REFERENCE IN INTERLANGUAGE: THE CASE OF THIS AND THAT. From linguistic annotation to corpus interoperability

Thomas Gaillat

UNIVERSITÉ SORBONNE PARIS CITÉ
UNIVERSITÉ PARIS.DIDEROT (Paris 7)

ÉCOLE DOCTORALE : *Sciences du langage - ED 132*

Unité de recherche: CLILLAC-ARP EA 3967

DOCTORAT EN CODIRECTION DE THÈSE

Linguistique anglaise

THOMAS GAILLAT

REFERENCE IN INTERLANGUAGE: THE CASE OF *THIS* AND *THAT*

From linguistic annotation to corpus interoperability

LA RÉFÉRENCE DANS L'INTERLANGUE : LE CAS DE *THIS* ET *THAT*

De l'annotation linguistique à l'interopérabilité des corpus

Thèse en vue de l'obtention du doctorat de linguistique anglaise
codirigée par Nicolas Ballier et Pascale Sébillot

Soutenue le 16 juin 2016

JURY

Nicolas Ballier, professeur à l'Université Paris Diderot (codirecteur)

Ana Díaz-Negrillo, maître de conférences à l'Université de Grenade (examinatrice)

Detmar Meurers, professeur à l'Université de Tübingen (rapporteur)

John Osborne, professeur à l'Université de Chambéry (rapporteur)

Pascale Sébillot, professeur à l'INSA Rennes (codirectrice)

Acknowledgements

So many people to thank! Where should I start? A statistical metaphor might help for this account.

As we reach the end of this PhD, we can gladly say that the null hypothesis (H0, i.e. “That thesis”) has been rejected based on a series of experiments. H1 is accepted in the following terms: “This thesis”. The work relied on a fairly sized sample of bright, kind and patient subjects whom I would like to thank in the following paragraphs. A cross-sectional analysis of their roles reveals a number of significant factors which can be reported.

H1 could not have been tested and validated without two extremely significant human factors, *i.e.* Pascale and Nicolas's involvement and guidance. Their involvement has been tremendous and their guidance with suggestions and proofreading fuelled my progression. I am indebted to them for all this (and not *that*!). Detmar Meurers's recommendations have also been very significant. He kindly took the time to explain key concepts and opened avenues that I never thought existed. I am so grateful to him for all this. I also would like to thank John Osborne and Ana Díaz-Negrillo for being part of the jury of the PhD. I also wish to thank Martine Schuwer for orienting me towards the rewarding field of research and for her constant support in this trajectory.

The design of the experiments required a number of tools which I needed to harness and this (or *it*?) leads me to name another set of key people who have also been supportive in this project. Camille Guinaudeau's sharp eye in reading my PERL

programs was very significant in terms of accuracy. Anca Roxana Simon's corrections in my R scripts unlocked difficult issues. Jonathan Kilgour's help in XML scripts made the corpora XML compliant. Vincent Claveau's and and Helmut Schmid's contributions cannot be factored out. Ana Díaz-Negrillo's sharing of the NOCE corpus was essential, and so was Pascale Gouteraux's involvement in the development of the Diderot-LONGDALE corpus. I am very grateful to both of them. Thanks also to Stefan Gries for his advice on modelling techniques. Recommendations from my colleagues of the CLILLAC-ARP team (Anne Talbot-Guyot, Alain Diana and Emmanuel Ferragne) were very much appreciated. Testing utterances on natives was possible thanks to Todd, Jane, Alex, Emily, Janet and Janet. My thanks go to every one of all these people. Many thanks to Vicki McNulty, a professional translator, for proofreading my English.

Family, friends and colleagues also played a central role. Vick's patience and support together with Eoin, Aisling and Séamus's encouragements were just crucial. My parents', Flo and Bert's discrete but nonetheless indispensable support cannot be forgotten. Greg and Perrine's support on bibliographical and MDF issues provided me with nice food for thought. François, Françoise†, Isabelle†, thank you for your kindness. Seb and Ala, I am not forgetting you! Finally, I am also grateful to my colleagues at the SCELVA and the University of Rennes 1 for the work arrangements that made this project feasible.

P-values for all the factors have been computed in the model and none of them is > 0.05, hence they are all significant. The R coefficient shows extremely high correlation between the factors and the outcome variable, *i.e.* “this thesis” can now be sent for validation.

To Vicki

Contents

Index of Figures

Index of Tables

Chapter 1 Introduction

Second Language Acquisition is a domain in which Learner Corpora have been playing an increasing role over the last two decades (Granger *et al.* 2015). These corpora provide insights into many aspects of learner language and help address research questions related to such fields as phonology, syntax and semantics as evidenced by the wide coverage of topics in the Handbook of the Learner Corpus Research. In fact, they show a large variety of language features used by learners, which makes them a powerful tool for exploring the intricacies of the learners' linguistic system called 'Interlanguage' (IL) (Selinker 1972). In IL, it is possible to study a range of linguistic and cognitive processes, one of them being how reference is established. Reference is a crucial component of human speech which can be analysed with two main concepts: deixis (reference to new information in extra- or intra-linguistic contexts) and anaphora (reference to already given information). This area has received a lot of interest in native languages including English (Cornish 1999; Kleiber 1992; Ariel 1994; Halliday & Hasan 1976). However, learner-language research on the subject has not been as intense and, yet, learners do experience difficulties such as overuse and misuse in wrong situational contexts (Petch-Tyson 2000). Many questions remain unanswered about the role of reference in IL. Developmental patterns on reference still need to be identified and little is known about specific reference-related learner features. One approach to the study of reference in IL involves the study of deixis and anaphora.

In deixis—understood as a procedure to locate a referent in space and time (Fraser & Joly 1979, 101)—many linguistic items such as *me, here* and *now* give speakers the possibility to refer to different types of entities depending on the context. In

anaphora, linguistic items such as personal or possessive pronouns help refer back to previously mentioned entities. In both concepts, the resolution of reference is usually carried out without any hindrance as each deictic or anaphoric form refers to a unique referent in a given situation or context. However, the case of *this* and *that* seems specific because both concepts partake of these two forms which are not simply lexical in nature. These two forms evolve in a multidimensional system which not only includes the syntagmatic and paradigmatic dimensions of speech but also the pragmatic dimension of discourse. Firstly, on the syntagmatic axis, they are multifunctional as they can have several functional realisations[1]. Depending on the context and their syntactic position, they can both be determiners, pro-forms or adverbials. Concerning the marker *that*, it can endorse two extra functions which are complementiser and relative pronoun. In short, for two forms there is a multiplicity of functional realisations depending on their position in the syntagmatic chain. Secondly, on the paradigmatic axis, they compete with each other but also with other forms depending on their function. Finally, *this* and *that*, especially as pro-forms, are markers of referentiality within the discourse dimension. They are used to point out specific referents, which mirrors a manner of conceptualising discourse. Pro-forms, as discourse markers, show a strong potential for referentiality due to their extendible capacity to either refer to single entities via nouns or more complex concepts via entire clauses, utterances or mundane situations.

In learner English, the study of *this* and *that* has not received much attention and yet, their complexity leaves room for learning difficulties. Much has been said about their central role in the construction of referential processes in native English (Halliday & Hasan 1976; Danon-Boileau 1984; Kleiber 1991; Cotte 1993; Ariel 1994; Biber *et al.* 1999; Cornish 1999; Stirling & Huddleston 2002; Strauss 2002). Yet, learners must experience the same needs, but maybe without the same

[1] In subsequent sections and chapters, we use the terms 'functional realisation' and 'functions' to describe the grammatical category of the forms, *i.e.* determiner or pro-form. We use the notion of 'position' to refer to the positional feature of the forms, *i.e.* subject or any other role. In the French enunciativist tradition, 'function' is used for what we refer to as 'position' (Moulin, Odin & Bouscaren 1996, 9).

outcomes. It is obvious that paradigmatic substitutions are possible between *this* and *that*—with an impact on meaning. But when the forms are filtered down according to their functions in learner English contexts, it appears that they also compete paradigmatically with other elements. As pro-forms, they compete with the pronoun *it*, hence forming a pro-form microsystem of reference. As determiners, they compete with the determiner *the*, hence forming a determiner microsystem of reference. In short, the paradigmatic axis seems to be problematic in learner speech. A few studies have offered a first insight into their use by learners (Petch-Tyson 2000; Lenko-Szymanska 2004; Zhang 2015; Young 1996) but the approaches always rely on one of the two following principles. Some only use data from one corpus, which appears as a limitation with regards to comparisons between learners of different L1s or with natives. Some approaches focus on the intra-deictic system represented by *this* v. *that*. In other terms, reported observations and analyses only focus on the two forms, which appears as another limitation, as they ignore any possible competition with other forms. This thesis is a research project on *this* and *that* with a special focus on comparing their use between several L1s and on taking into account other competitor forms (see Table 1). We focus on the exploration of the pro-form microsystem due to the strong potential for referentiality it reflects.

This & *that*	Pro-form microsystem	Determiner microsystem
Competitors	*It*	*The*

Table 1: Competitor forms of *this* and *that* depending on the microsystem of reference

1.1 Research question and hypotheses

Due to the multidimensional complexity of the two forms, we may wonder if their use in learner English is performed as smoothly as in native English. Learners may distort referential processes due to the versatile nature of these two forms. This leads us to our research question, which is linked to the way learners implement deictic and anaphoric procedures, especially within the pro-form microsystem. As *this* and *that* are two of their main components, investigating their use is a way to explore learners' construction of referential processes. Our research question could

be formulated as follows: To what extent do factors, such as the learners' L1 or the functions of the forms, come into play in learners' implementations of *this* and *that* in referential processes? Our assumption in this question is that, by observing and analysing how learners implement them, it is possible to improve our understanding of one aspect of learner's implementation of reference. In short, investigating the use of the forms in learner language is a way to analyse how deixis articulates with anaphora in IL.

Our research question leads directly into hypotheses which can be operationalised in the following manner:

i) Learners' patterns of use of the forms are L1 dependent due to influences from their L1.

ii) A learner-specific pro-form microsystem of reference exists, including specific linguistic features attached to *this* and *that* which are different from native English.

iii) Learner-error patterns are linked to specific linguistic features of the forms.

1.2 Aims

To test these hypotheses a number of steps need to be taken. First of all, we need to carry out a detailed study of the referential dimension of the linguistic system in native English. The aim is to provide a dynamic model of the many parameters of the system in order to have a detailed view of their interactions and of the linguistic features that characterise *this* and *that*. Elaborating a linguistic model of reference on native English allows us to interpret learner use of the forms and to identify potential interactions between several forms within what we call the learner microsystems of reference.

Secondly, testing these hypotheses requires an experimental design which relies on methods that are empirically grounded in corpora. As the purpose is to compare

speakers of different L1s, the aim is to contrast several corpora to reflect this diversity. By comparing native speakers with learners (Ellis 1994, 345) and learners of different L1s (Granger 1996; Gries & Deshors 2014), the purpose is to gain access to the different strata of IL (in Selinker's terms, strata are called 'units'). By combining this contrastive method with frequency counts of specific features and with statistical methods such as regression, the aim is to identify L1-specific patterns of uses.

Thirdly, several corpora need to be annotated in such a way that reflects the multidimensional complexity of the forms and which makes their data interoperable, *i.e.* the capacity to share and combine the data from different sources (Sérasset *et al.* 2009). The aim is to extend the annotation scheme provided by existing tools in order to have a fine-grained description which helps distinguish the forms according to their functional realisations, their syntactic positions and semantic features. Pro-forms can therefore be isolated and compared with the *it* pronoun, which gives a cross-corpus insight into the pro-form microsystem. Because we also need to be able to compare occurrences between corpora, the same annotation scheme needs to be implemented across all corpora. Consequently, the aim is to use machine learning technologies to automate the annotation process as much as possible in terms of accuracy.

1.3 Method and contributions

Our aims require the development of a formal framework which is carried out in five stages:

i. Building a linguistic model of reference
ii. Determining the learner microsystems of reference
iii. Choosing the corpora and the annotation scheme
iv. Extracting linguistic data
v. Analysing linguistic data

The following paragraphs give details on each of the stages.

i. Firstly, in terms of linguistics, we establish a model of reference to show how the two forms are integrated within the linguistic system. Our contribution is to show that focusing on the components of the referential framework and their articulations can reconcile the apparent dichotomy between deixis and anaphora (Halliday and Hasan 1976; Kleiber 1992; Cotte 1993; Cornish 1999). At textual level, the forms can be approached according to their function and their syntactic position. At contextual level, the endophoric/exophoric distinction supports different types of referential processes. At discourse level, the deictic/anaphoric distinction indicates whether the referent is new or not. By looking at the interactions of these components within referential processes, we provide a novel dynamic model of how referents are retrieved, depending on the degree of accessibility of their referring indexical expressions. In this view, *this* and *that* are indexical expressions whose functional realisations and semantic values influence the meaning of utterances.

ii. The second stage is a small-scale survey of learners' use of the forms on a limited number of learner utterances. We aim to understand the syntactic and contextual levels of interpretation of errors. We use a hybrid methodology based on error and form/function analysis and we propose a new typology of errors on *this* and *that* in light of the components of the referential framework. By focusing on the functional realisations of the forms as pro-forms or determiners, we establish new hypotheses on the way the forms are used. We uncover the two microsystems of reference mentioned previously that seem specific to learners. These two systems rely on specific linguistic features attached to the forms but their relevance remains to be assessed with regards to different L1s.

iii. The third stage in the development of the framework is a large-scale study of the forms to find statistical evidence of the pro-form microsystem. We adopt a multi-corpus approach and we determine the annotation scheme which

ensures that different levels of information on *this* and *that* can be retrieved. We use three corpora characterised by different L1s. The Penn Treebank (Marcus, Marcinkiewicz & Santorini 1993), a native written corpus, is used as a gold standard for its accuracy in tagging functional realisations of forms and for its single journalistic genre. A subset of the LONGDALE corpus (Meunier *et al.* 2008) includes a French L1 variety of spoken English and the NOCE corpus (Díaz-Negrillo 2007) reflects a Spanish L1 written variety of English.

Regarding the annotation scheme, there are two parts.

The first one is linked to the number of annotation levels to have, *e.g.* semantic, discourse or phonetic annotations (Leech, 2005). To address this task, we specifically devise a multi-layer annotation scheme for the interpretation of the two forms as well as *it*. We encode three levels of annotation: i) the functional realisations of the forms, *i.e.* determiner or pro-form ii) their syntactic position in the sentence, *i.e.* oblique or nominative and iii) their referential semantic value, *i.e.* endophoric or exophoric. The first two are achieved automatically, the third one carried out manually. Concerning a learner-error specific level of annotation, our approach consists in using the three-layer annotation scheme to describe learner language as neutrally as possible in terms of features. Following the principles set by Díaz-Negrillo et *al.* (2010), we show that by annotating properties that can actually be observed in corpora, we can contribute to automatic error detection.

The second part is linked to the annotation level dedicated to the pro-form and determiner realisations of the forms. This kind of information is expected to be found in the Part-of-Speech (PoS) annotation which provides information on the grammar categories of forms. A review of popular PoS annotation schemes on the market shows that there is a broad variety of

them. When examining the demonstratives as regards their annotation in different mainstream tagsets used for PoS-tagging, we see that PoS tags lack granularity as they do not allow basic grammar distinctions for the two forms. The labels used to tag *this* and *that* need to be refined in terms of functional realisation. Concerning *this*, the difference between the pro-form, determiner and adverbial functions needs to be marked. Concerning *that*, distinct tags are required for each of the following functions: determiner, pro-form, complementiser, relative and adverbial. In our work, we create a fine-grain tagset for the forms and we apply it automatically. We show that it is possible to PoS tag native and learner corpora with a modified PoS tagset which includes the distinctions in functional realisations of all occurrences of *this* and *that*. We also apply the same principle for the annotation of *it* in order to distinguish the pro-form and the non-referential functions attached to this form. To achieve this, we use TreeTagger (Schmid 1994), specifically retrained on a modified version of the Penn Treebank tagset (Marcus, Marcinkiewicz, and Santorini 1993).

iv. The fourth stage focuses on two types of extraction of data from the corpora. Firstly, we create a data structure with the NITE XML toolkit (Carletta *et al.* 2006) to search occurrences of the forms and automatically retrieve them with their close context thanks to multi-criterion queries that combine all three annotation layers. Secondly, we create methods to extract the occurrences and their linguistic features which are automatically placed in tables. These tables are abstractions of the corpora in the form of sequences also called instances. These tables can support data analysis with machine learning technologies.

v. In the last stage, we use the tables of linguistic data to conduct linguistic data analyses. Regression models are used to identify the linguistic features which are significant in the use of the forms depending on the L1s. Automatic classification based on the k-Nearest Neighbour algorithm

(Daelemans & Bosch 1992) is performed to identify patterns of use of the forms, including errors. New evidence of L1-specific patterns is uncovered in the acquisition process of *this* and *that*. Quantitative evidence of the learner-specific pro-form microsystem is given and L1-specific error types are identified.

1.4 Outline

This thesis is divided into three main parts. Firstly, the theoretical framework of reference related to *this* and *that* is covered in Chapters 2 and 3 in order to understand the linguistic motivations for learners' errors. Both chapters include state-of-the-art sections to distinctively explore the linguistic issue of reference and the methods used in the domain of Second Language Acquisition. Chapter 2 gives an overview of the state of the art about reference and the demonstratives in native English. Chapter 3 still focuses on reference but we look at the issue from the angle of learner English. Grounded in the functional distinctions of the forms, we elaborate a new typology of errors taking into account the two L1-specific microsystems of reference. Secondly, we focus on the methodology. In Chapter 4, we describe the multi-layer annotation scheme. In Chapter 5, we show how corpora are made interoperable. The final part of this thesis is about the analysis of the data and the discussion of the results. Chapter 6 shows how regression models help uncover significant features of the pro-form microsystem. Chapter 7 focuses on the automatic treatment of learner corpus data with two purposes: i) to factor out the relevant features of error patterns in the use of the forms and ii) to automate the detection of unexpected uses of *this* and *that*. The conclusion is presented in Chapter 8.

Chapter 2 *This* and *that* in the referential framework of native English

In this chapter, the aspects of the linguistic system in which *this* and *that* are at work are analysed. The focus of this thesis is on learner English (as instantiated in two learner corpora with different L1s) but, as native English is the target in learning strategies (Ellis and Barkhuizen 2005, 60), it is necessary to focus on the underlying principles that seem to guide the native choices of linguistic items in texts. This stage is required to understand better how learners make use of the two demonstratives, *i.e.* their errors. Ultimately, such a focus is necessary to model correct and incorrect uses of the forms with a high degree of detail. This chapter is dedicated to a state of the art of the referential framework of English. We progress from a general to a detailed approach by focusing on reference before examining the case of *this* and *that* in referring expressions.

In Section 2.1, we give an account of the referential framework which seems to underlie all referential processes. In Section 2.2, we cover the textual approach on reference and show the limitations of the antecedent and the situational/text paradigms to elucidate referential processes. In Section 2.3, a discourse-functional approach is described to show that the act of referring involves several levels of interpretation of speech such as text, context and discourse. In Section 2.4, we show how referential expressions are used within the discourse-functional approach and how they help integrate referents into discourse. Section 2.5 offers a focus on *this* and *that* according to several levels of interpretation. We show how natives use them in utterances to build their discourse.

2.1 The referential framework

Reference is a concept that has been well studied in the literature. Deixis and anaphora are presented as the two components of the referential framework (Halliday and Hasan 1976; Cornish 1999; Fraser and Joly 1979; Stirling and Huddleston 2002; Kleiber 1992 - among others). However, their distinction has proved to be somewhat cumbersome as the division follows along a pervasive line where some forms may be of both anaphoric and deictic values, not to mention the definitions themselves of the terms that are not stable among the community of linguists. Indeed, several visions of anaphora and deixis seem to emerge and correspond to two levels of analysis. This section covers the definitions of essential concepts before getting into the details of the theoretical accounts that reflect textualist and functionalist views of the matter.

Before seeing how reference is seen at each level, it is necessary to understand the underlying concepts in which each vision grounds itself. In fact, everything depends on the way the notion of 'text' is interpreted as there are many acceptions of the term, ranging from co-text to discourse and including context. Cornish (1999) proposes a vision that encompasses the different natures of the way speakers and addressees see this concept of text. Instead of opposing co-textual and cognitivist approaches, he insists on the manner in which the speech situation (or the co-text) is mentally represented and takes the form of a discourse model (Cornish 1999, 14). In so doing, he fastens his analysis on this distinction when dealing with the concept of text.

Text “denotes a typical instance of language *cum* other semiotic devices in use—*i.e.* occurring in some context and with the intention by the user of achieving some purpose or goal thereby” (Cornish 1999, 33). So, text in this representation is composed of language and other semiotic devices put together for a specific objective by the speaker. The author follows on by assimilating text to signs and signals as “the connected sequence of verbal signs *and* non-verbal, vocal as well as non-vocal (*i.e.* visual, auditory, etc.) signals [are] produced within the context of

some utterance act". For Cornish, text is different from discourse as the latter is defined as the "situated construction and interpretation of a message via a given text relative to some context, in terms of the speaker or writer's hypothesised intentions" (1999, 35). Here, the focus is on how the message is constructed and interpreted by the speaker.

Before elaborating on the notion of 'discourse', that of 'context' requires further details as it acts as a sort of function for the realisation of discourse constructed from the mere set of signal and signs. Cornish uses the adjective *relative* to express this go-between role that 'context' plays to link text to 'discourse'. In his 2010 article in *Functions of Language* (Cornish 2010), he provides a definition of context:

> "The domain of reference of a given text, the co-text, the discourse already constructed upstream, the genre of speech event in progress, the socio-cultural environment assumed by the text, and the specific utterance situation at hand" (2010, 209).

Context endorses more substance than *text*, *i.e.* the notion of 'text'. It embarks meaning in the form of memorised information related to the subject of the speaker's production. However, context must be understood as a 'mathematical' function that transforms text into discourse. It is "by invoking an appropriate context that the addressee or reader may create discourse on the basis of the connected sequence of textual clues that is text" (2010, 209). So, it is clearly thanks to context that discourse emerges from text.

In the cognitivist approach, 'discourse'[2] is the final product of a set of utterances whose acts are hierarchically organised in a situation. These acts may be of an indexical, propositional or illocutionary nature and are carried out with a communicative objective. Discourse is the product that results from these acts in a given situation of utterance. The discourse model that is produced is a representation that is stored in the long-term memory of the addressee and the speaker, and it can be retrieved at any later stage.

[2] To simplify our notation, we do not make typographical distinctions for autonymy. We write discourse with no quote or italics when we mean the concept of /discourse/.

The referential framework relies on the three textual elements described above (co-text, context and discourse) and the way linguists view these three elements has implications on how referential processes are regarded as far as referent identification is concerned. In the course of this chapter, Cornish's distinction between text and discourse is used as a prism to read and understand how referential procedures may be analysed. In the next section, we start with a semasiological point of view in which reference is viewed as the elicitation of the meaning of referring words as opposed to the identification of the concepts that underlie these words.

2.2 Reference: a textual approach

This section focuses on one type of approach that has been predominant in the field of anaphora and deixis. Resolving anaphors in texts has led researchers to analyse and classify anaphors according to two paradigms. Firstly, the match between an anaphor and its textual antecedent is sought. Secondly, this objective and the methods employed have led researchers to divide the domain of reference into two areas: text and situation. The following is an account of how these two paradigms articulate.

2.2.1 The antecedent paradigm

The co-textual approach is largely developed in experiments in the domain of computational linguistics. One of their applications may be automatic coreference resolution, which consists mainly in analysing anaphora via elements present in the chain of words. Resolving referential procedures is achieved within the text itself. In the case of anaphora, the focus is to identify the antecedent in the text. In other terms, the antecedent is the central part of the analysis. Under this conception, the antecedent and the anaphor form a pair that needs to be associated. Establishing this association revolves around resolving the link between the initially constructed entity and the non-semantic form that then appears within the same text. Here, the approach follows a syntagmatic line as anaphors are said “to refer back to a

previously mentioned item" (Mitkov 2002, cited in Cornish 2010, 213). For the sake of illustration, let us take Example 1, which was proposed by Kamp & Reyle (1993, ex 66 cited in Cornish 2010, 213):

1) Jones owns *Ulysses*. It fascinates him.

Here both anaphors, related to the pronouns *it* and *him,* find their resolutions on the syntagmatic axis with the NPs *Ulysses* and *Jones*. The difficulty in this case is to get the algorithm to match the pairs correctly as semantico-cultural aspects need to be taken into account to create the matches.

The previous example illustrates that in computational linguistics the tendency is to follow the syntagmatic axis by automatically searching for and classifying co-textual features. Lappin (2005) describes two types of approach to resolve anaphors: knowledge-based and inference-driven on one side and knowledge-poor on the other side. In the first case, complex systems "rely on rules of inference that encode semantic and real-world information in order to identify the most likely antecedent candidate of a pronoun in discourse" (Lappin 2005, 4). The second case corresponds to systems that "rely on features of the input which can be identified without reference to deep semantic information or detailed real world knowledge" (Lappin 2005, 5). In both cases, though, interpreting an anaphor is equal to finding its antecedent in the text. This trend focuses on morphosyntax and tends to minimise or underuse the semantic dimension in favour of the grammatical relationships between elements since "the semantic/discourse representations to which the inference rules apply are not reliably generated for large texts" (Lappin 2005, 5).

The co-textual approach can be illustrated by Soon *et al.*'s (Soon *et al.* 2001) experiment on the resolution of coreference of Noun Phrases. In this approach, they use NLP tools to extract possible pairs of what they call 'nested noun phrases', in which anaphors and NPs can both be found. Once the extraction is complete, potentially coreferential pairs are listed with their specific features such as the

distance between coreferring elements or their grammatical categories. The series of verified coreferential pairs are used in an artificial intelligence software tool, called a 'classifier', in its training phase. In this phase, the classifier learns the properties linked to actual coreferential pairs. In its performance phase, the classifier is given new anaphors and must automatically match them to candidate antecedents. In other terms, the classifier is used on new antecedent-anaphor pairs to classify them as positive or negative. The results show a F-measure[3] of 62.6% at best for the resolution of coreferential pairs, with errors related to the wrong assignment of semantic classes to words, *i.e.* nouns classified in wrong categories such as *person* instead of *object*. This illustrates the limit of real-world knowledge inserted in such systems. Another limit comes from the fact that such systems use antecedent-anaphor pairs and this eliminates occurrences of antecedentless anaphors. Consequently, as Cornish puts it, there are other factors than just the co-text to resolve anaphors (2010, 215).

2.2.2 The situational/text paradigm

The first level of analysis of anaphora and deixis corresponds to the co-textual approach. Here, reference is viewed according to a situational/textual dichotomy where deixis corresponds to situational reference and anaphora to text reference. Situational reference relies on deictic expressions that refer to variable referents that belong to the situation of reference. In chapter 17 of the Cambridge Grammar of the English Language on reference (Stirling and Huddleston 2002), Stirling and Huddleston provide a clear account of the varying nature of reference that characterises deictic expressions: “What is significant with deixis is that the shifting reference is systematically tied to features of the utterance-act itself” (Stirling and Huddleston 2002, 1451). They provide the following example:

2) “Could *you* pick *this* up and put it with *those* boxes, please.”

[3] Metric used to measure accuracy in the domain of information retrieval.

In Example 2, the demonstratives *this* and *those* correspond to locative deixis as it is necessary to know the locational parameters of the situation of utterance in order to clearly identify the referents, which is the case of the speaker and the addressee. Likewise, *you* corresponds to person deixis as it is necessary to have access to the parameters of the situation in order to identify who exactly the pronoun refers to. In all three instances of *this, those* and *you* Stirling and Huddleston show that the entities to which they refer depend on the "utterance-act" that is variable in nature.

Text reference is also clearly defined by Stirling and Huddleston: "Anaphora is the relation between an anaphor and an antecedent, where the interpretation of the anaphor is determined via the antecedent" (2002, 1453). They provide the following example:

3) *Max* claims *he* wasn't told about it.

In this case, *he* is the pronoun that refers to the entity to which the proper noun *Max* refers to and which is defined as the antecedent. The resolution of the anaphor *he* is validated when the antecedent has been matched to its antecedent within the text. So, the text is the place where anaphora takes place and the identification of the tie between the antecedent and the anaphor is, in this approach, the sole criterion for reference resolution.

Halliday and Hasan's concept of 'coreference' as part of their analysis of cohesion can be used to analyse anaphora. They define coreference as a tie of a particular kind between "a pair of cohesively related items" (1976, 3). Coreference is a central feature for text cohesion as this procedure allows two items to be tied together thanks to the fact that "they are identical in reference". In example 4 further down, with the first instance of *this,* cohesion is achieved thanks to a tie that links the demonstrative to the situational entity at the time of Churchill's speech: Great Britain. In the second instance, the tie links the demonstrative to the geographically located nation previously mentioned and referred to as 'Germany'. So, *this state* and *Germany* corefer to the same entity. Even though this analysis recognises the fact

that both elements refer to one entity, it eludes the interpretation of the referential expression as mundane reference. Because of this, it can be said that Halliday and Hasan's view of reference remains a textual one. As will be seen in Section 2.3, discourse analysis will fill the gap.

The notion of text may also be accompanied by the recognition of a context of situation. In their work on cohesion, Halliday and Hasan (1976) distinguish “the relations within language, patterns of meaning realised by grammar and vocabulary” from “the relations between the language and the relevant features of the speaker's and addressee's (or writer's and reader's) material, social and ideological environment” (Halliday and Hasan 1976, 20). By so doing, they recognise the context of situation as the extra dimension in which the text is embedded. They see this second aspect as a set of extra-linguistic factors that have “some bearing on the text itself” (1976, 20). However, on the basis that they are two different sets of phenomena, they choose to focus only on the linguistic factors that constitute the characteristics of texts in English. Indeed, as the purpose of their book is cohesion—reference is one of its components (1976, 5)—they classify it as one of the elements that partakes of the textual component of the linguistic system (1976, 27). Next to cohesion, they add the concept of 'information structure' that refers to “units of information on the basis of the distinction into given and new”, which would be located at the discourse level in Cornish's terms (Cornish 2011, 3) as will be shown in Section 2.3. For Halliday and Hasan (1976, 27) cohesion is a “potential for relating one element in the text to another, wherever they are and without any implication that everything in the text has some part in it”. Information structure, on the other hand, does not belong to cohesion as “there are no structural units defined by the cohesive relation” (1976, 27). So, in spite of the fact that Halliday and Hasan recognise the notion of information conveyed by the text in the form of context of situation, they decide to separate it from the notion of cohesion which includes referential procedures.

Consequently, this dichotomy positions their analysis of reference within the textual frame described previously. It can be said, though, that their approach is similar to the discourse functional approach (see Section 2.3), insofar as they endorse the notion of context by taking into account the roles of a speaker and a listener. But instead of seeing text and context of situation as two interacting aspects of language, they disconnect both concepts by drawing a dividing line in their analysis of the concept of reference.

In their terms, referential procedures branch out in two directions. Reference can be made to an entity located within the text, in which case it is said to be endophoric. When the entity corresponds to an object belonging to the situation of utterance, it is said to be exophoric. Figure 1 (Halliday and Hasan 1976, 33) shows the processes that are contrasted.

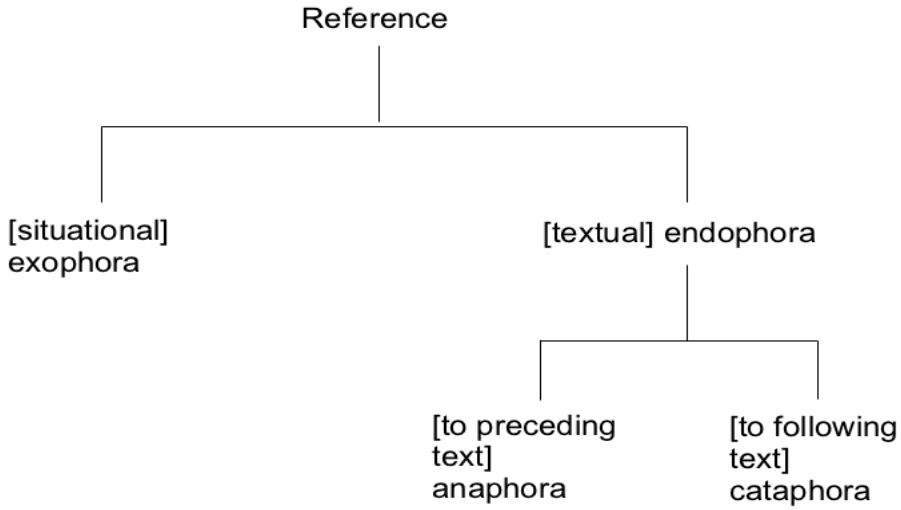

Figure 1: Halliday and Hasan's scheme to distinguish the class of referential processes

Following up on Halliday and Hasan, Fraser and Joly (1979, 106) give the following example:

4) “Ought we, instead of demonstrating the power of our Air Force by dropping leaflets over Germany, to have dropped bombs? (…) In *this* peaceful

> country, governed by public opinion, democracy and Parliament, we were not as thoroughly prepared at the outbreak as *this* dictator state whose whole thought was bent upon the preparation for war." (W. Churchill, War Speeches, 27/1/1940)

The first *this* is exophoric and refers to Great Britain, the country of the speaker. In other terms, this entity belongs to the situation of utterance and the speaker refers to an entity located outside his speech and physically present in the situation of utterance. The second *this* is endophoric and it is a direct reference to the previously mentioned entity Germany. In this case the entity is constructed within the text and its reference can only be resolved thanks to the previous construction. Fraser and Joly's analysis relies greatly on the way Halliday and Hasan (1976, 33) define the system of reference by designating exophora and endophora as the two branches.

Figure 2 shows how the authors developed the initial diagram in order to account for the conceptual link between endophora and exophora. They indeed raise the question of the order of appearance between endophora and exophora. Just like Halliday and Hasan (1976, 62), they argue that deixis (or reference interpreted in the Greek sense of *deiknunai*, that is 'the act of showing') begins with exophora and "endophora seems to be a particular case of the former insofar as the 'linguistic' context is included in the 'situation', in the broad sense of the term" (Fraser and Joly 1979, 107 my translation). They summarise this inclusion by saying that contextual ostentation is the result of the transformation[4] of situational ostentation. That is the reason why exophora includes the *before* label and why the *after* label is attached to endophora.

[4] In French they use the terms "une transformée de l'ostension situationnelle".

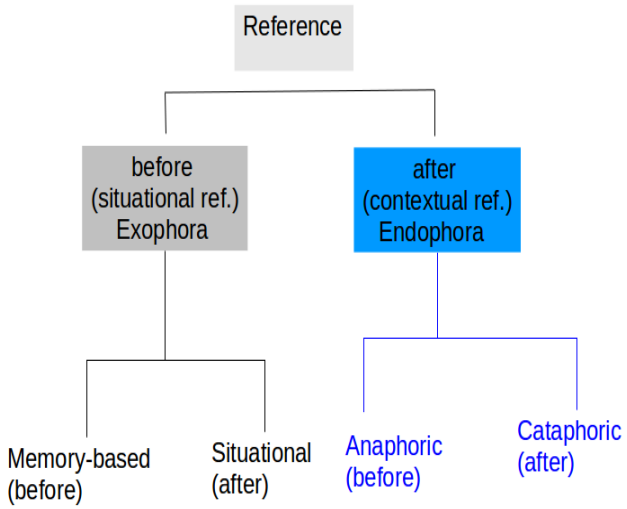

Figure 2: Fraser and Joly's scheme to distinguish the types of referential processes by taking memory into account.

As can be seen in Figure 2, they also add an extra level of distinction within each branch and take into account the notion of memory. Exophora can be memory-based or not. In other terms, memory-based exophoric procedures refer to entities that do not exist any more in the situation of utterance. Their absence is known to both the speaker and the addressee and, as such, the referential process is still valid. Likewise, endophora splits into two parts: memory-based and forward-looking reference. Memory-based procedures may refer to entities previously mentioned in the text and are called 'anaphoric'. Forward-looking procedures may refer to entities not yet mentioned in the text and are called 'cataphoric'.

The fact that situation and text are the two components of reference appears as a main difference with the cognitive approach based on a difference between what is considered as 'new' and 'given' information. Cotte's approach (1993, 47) may be seen as encompassing both views. The distinction between deixis and anaphora depends on the addressee's representation of the temporal and situational location of the referent. For discourse-functionalists, the reference to a salient entity or the focus on a new entity seems to be the dividing line. In fact, depending on the point of view of the speaker in relation to hi(s)her discourse and when taking into

account the interlocutory act, the speaker may sustain the fact that the object of the discourse is new to the addressee even though it is only a simple anaphoric reference as far as his knowledge of the communication situation is concerned. For Cotte, exophoric and endophoric references and focus are processes that all exist in the act of pointing that is at the root of the use of deictics. Section 2.3 shows how the discourse-functional approach uses the speaker and addressee's relationship to develop a view that does not reject endophora and exophora but, rather, that inserts these procedures in a broader set of phenomena. This approach adopts an onomasiological point of view in which referential processes are conceptualised in a functional and cognitive framework.

2.3 Reference: a discourse-functional approach

So far, the views that have been covered encompass text and context with variations but they do not articulate them in the way some linguists do by linking text to the discourse formed in the mind (see Cornish 1999; Ariel 1994; Kleiber 1992). In these authors' views, referential expressions target referents that are mental entities. Ariel believes—in reaction against the distinction between linguistic antecedents (for anaphora) and world object-antecedents (for reference)—that "it is more plausible to hypothesise a connection between linguistic expressions and mental entities" (1994, 27). These mental entities are the elements that constitute discourse, and the question of how discourse is being constructed finds an answer in Cornish's view of referential processes. For him, these processes use text and context to produce discourse, which implies the acknowledgement of the speaker-addressee's relationship. He insists on the fact that the "establishment of referent is a joint achievement, undertaken by the interlocutors collaboratively, and is not simply the responsibility of the speaker alone" (Cornish 1999, 20). This approach adopts a cognitive angle for the understanding of deixis and anaphora by linking the surface structure, composed of linguistic and situational signs and signals, to the way the mind processes reference and by ensuring that both interlocutors are on par with their attention focus. Consequently, the mind plays a

central role since it is the location of the unfolding of discourse that relies on knowledge-based and memory processes.

The following sections focus on explaining the essence of the discourse functional approach to reference. In Section 2.3.1, the role of information and the way speaker and addressee interact are addressed. Section 2.3.2 focuses on the notion of referent and the way it is made accessible in discourse. Section 2.3.3 details the process of retrieving referents from various types of contexts and Section 2.3.4 is an attempt to present a model of the reference system in which referential processes are dynamically constructed.

2.3.1 The role of information and the speaker-addressee relationship

As was mentioned in the previous section, reference is made of the two components that are deixis and anaphora. In the textual approach, their distinction is made according to the situation/text paradigm. In the cognitive approach, both components are also distinguished from one another but differently from the situation/text paradigm. The latter appears as erroneous: “It is wrong to suggest […] that anaphora yields a sense or a referent via an intratextual linkage between expressions within some co-text, and that deixis (and thus 'exophora') assumes a relation between a segment of co-text and a referent located in the utterance situation” (Cornish 1999, 117). In fact, deixis and anaphora are seen as discourse functions used to “ensure that the speech participants are on the 'same wavelength' with respect to their focus of attention at any point in the discourse” (Cornish 1999, 24). Focus of attention in the interlocutors' mind is the point of articulation of the referential processes that take place in the course of conversation. Anaphoric and deictic functions are used to coordinate both interlocutors' mental activity taking place in their mind during the construction of meaning and its by-product that is discourse.

The question is to identify the underlying paradigm that acts in the distinction between both components of reference. The aforementioned authors, who have a

pragmatic approach to text, centre their approach on the notion of information. As a comparison, Halliday and Hasan (1976, 27) recognise it under the term 'information structure' but they do not integrate it within the issue of reference. Nevertheless, information is processed in the brain in order to reach certain purposes such as communication. So, instead of seeing information as a side development of cohesion, it may be relevant to see it as its core objective, which makes it part of cohesion. Information needs to be structured and organised and the way to organise it is to classify it according to two fundamental cognitive states that are 'known' or 'unknown'. In the course of conversation, the information shared by the speaker and the addressee is constantly classified and updated and its mental representation is either new or already known. Ariel (1994, 27) uses the term 'givenness' to specify expressions that provide various degrees of accessibility for information (we will cover the issue of accessibility in Section 2.3.2). Under this view, reference is a procedure used to organise information. Kleiber also follows this line. He does not see deixis and anaphora from the point of view of the speaker sending a message in relation to where (s)he is (*e.g.* situational or text reference). Instead, Kleiber sees the mechanism at an information level where either the information carried by the referent is new to the addressee or it is a repetition of an already salient referent (Kleiber 1992, 618). The process of deixis in this case is to point out new information whilst the process of anaphora retrieves already-extracted information in the previous co-text. Cornish sees the components of reference in the same way, that is operating at the level of memory organisation (1999, 117). Information, stored in the memory, is classified and later retrieved via referential processes. To do so, the classification and retrieval processes rely on the 'new' or 'given' paradigm. As previously quoted, Cornish insists that the same focus of attention must be enjoyed by the addressee and the speaker in the conversation. He uses the expression 'same wavelength' and because of their need to fine-tune their perception of “what the predication currently being constructed (by the speaker) or interpreted (by the addressee) is about”, the two participants use deictic and anaphoric expressions to adapt the focus of attention and to make sure that they both refer to the same entities. Cornish provides two definitions of deixis

and anaphora. Their distinction is grounded on the principle of what is given or new.

> “Deixis on this view is the use of a member or members of a set of devices, whether linguistic or paralinguistic in character, whose object is to ensure the refocusing of the interlocutor's attention on a particular discourse entity, a refocusing which is rooted in the current utterance context; while anaphora is the use of a member (or members) of a complementary set of purely linguistic devices whose role is to ensure that the interlocutor maintains the focus of attention already established at the point where the anaphor occurs.” (Cornish 1999, 25–26).

The point is that information is structured and restructured permanently in discourse and the speaker needs either to introduce new entities or to refer to previously constructed discourse entities. Deictic and anaphoric procedures are grounded on the principle of new or given information. Hence, the fact that some expressions in speech may or may not be coreferential is not a decisive element in the interpretation of these entities. What is at stake is to correctly identify the discourse referents that are targeted in what Cornish calls the 'discourse model representation'. Interpreting an anaphor, for instance, cannot be done by simply establishing the tie with its textual antecedent since this antecedent may well not exist in the previous co-text. In fact, situational and text reference are identical in their 'pointing' process. The same mental process is at work in order to let the addressee identify the referent intended by the speaker (Cornish 1999, 22). This view follows on from the idea set by Halliday and Hasan in which they posit that exophora and endophora partake of the same process because linguistic context is part of the more global situation (Halliday and Hasan 1979, 107). If we adopt this view, the notion of antecedent attached to text becomes obsolete because it does not reflect the case of the deictic procedure in which a referent is also pointed at (more details on the way the deictic procedure applies are given in Sections 2.3.3 and 2.3.4). Ariel makes the same claim as she insists that the distinction between situational ('referring' in her own terms) and coreferring (equivalent to textual) is artificial. For her, “the factors involved in the search for mental antecedents do not distinguish between reference and anaphora as such” (Ariel 1994, 27). Kleiber also submits the criterion of textual or situational location to the more prevalent and influential criterion of known or unknown information. He indicates that the

"anaphora/deixis opposition changes its content as it turns into a distinction at memory level: what is anaphoric is an expression that refers to an already known entity or an already existing entity in discourse. It can also be an entity already in the attention focus of the addressee; what is deictic is an expression that introduces a new referent in the focus." (Kleiber 1992, 55).

So, what is paramount to reference as a whole is that both interlocutors have the same focus of attention on the same entity at the same time. Deixis and anaphora are the discursive means that allow this. In this approach, the term *antecedent* used in the textual approach becomes unclear or even inappropriate. Does it refer to the word form present in the text, *i.e.* the 'signifier'? Does it refer to the entity represented in the mind, *i.e.* the 'signified'? Or does it refer to the real-world 'referent'? In fact, it may apply to all three answers, which shows the need for a shift in terminology by using the term *referent* instead.

2.3.2 The referent and its accessibility

At this point in the characterisation of the approach, it is relevant to detail how the notions of referent and antecedent interact. In fact, the discourse functional approach takes into consideration the notion of referent rather than that of antecedent. Indeed, the articulation point of the resolution process of an anaphor in a text, or of a pointing signal in a situation, depends on the link with the mental referent. The discourse functional approach introduces a layer of referential interpretation in which entities are represented mentally (Ariel 1994, 27) and accrue properties as the discourse unfolds (Cornish 1999, 46). A referent is accruable insofar as it acquires and increments various characteristics when the discourse is being developed by the interlocutors. This is what Cornish refers to as *antecedents*. In the course of discourse progression the referent initially linked to real-world entities takes on a more context-related set of features that makes it unique in the mind of the interlocutors. As Figure 3 shows, the referent accrues properties at discourse level and, in the case of anaphora, such entities are named

antecedent by Cornish (1999, 44). At text or situation level, Cornish uses the term 'antecedent-trigger' to refer to the utterance tokens, the gestures or the percepts that are present in the signs or signals produced by the speaker and that trigger the first appearance of a referent in discourse. It then is submitted to evolutions and the presence of an anaphor later in speech provides the need for retrieval of that discourse referent. Therefore, depending on the level of interpretation of reference there are two types of items: antecedent-trigger (in text and situation) and antecedent which Cornish also synonymously names as discourse referent. The trigger places the referent in the memory of the interlocutors and the antecedent represents the referent at discourse level, including all the properties that have been added to the initial concept by the speakers.

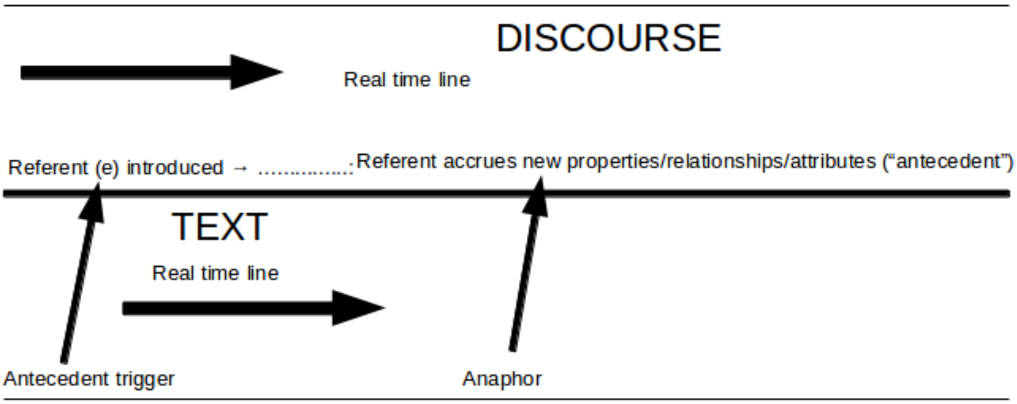

Figure 3: Cornish's system of referent accessibility (Cornish 2011)

Once established that the referent is the starting point of referential processes, it is possible to analyse the manner by which this discourse referent is referred to as the discourse unfolds and as speech in the form of sign or signals is being produced. When discourse develops, the interlocutors point out different entities of different types and construct their referents (see Section 2.4.1 for a discussion on the nature of these entities). Later in their conversation they need to refer to these entities and, in order to do so, the question of the accessibility of the referent is raised. Kleiber, who also favours the discourse approach, clearly states the importance of the referent:

"The cognitive approach draws from the way the interlocutor recognises the referent

> or, in more cognitive terms, the accessibility of the referent." (Kleiber 1992, 617, my translation).

In this paper presented at the Sorbonne in 1990, he draws the same distinction between anaphora and deixis as that chosen by Cornish and Ariel and explained above. The way the referent is seen and understood by the two interlocutors acts as the central force of reference. Resolving deictic and anaphoric processes tantamounts to recognising the referent as being already known or new. This depends on how accessible it is in the memory of the interlocutors. Kleiber shows that the saliency of the referent in the unfolding discourse is the driving force behind the choice of one of the two referential procedures. In other terms, the referent is already salient in the discourse or it is new due to its introduction at the moment of utterance. Cornish talks about the notion of attention focus. The object of deixis is to ensure the refocusing of the attention on a particular discourse entity, either because it is totally new or because a new aspect of an already existing referent in the discourse is to be pointed out by the speaker. Conversely, the object of anaphora is to maintain the attention focus on a referent already established in a discourse. The speaker presupposes that the addressee holds the referent at a high level of focus too. Accessibility is the guiding principle that interlocutors follow when referents need to be retrieved. Either these are considered as being in-focus and anaphoric procedures are put in place to retrieve them or they are considered as not being in-focus. Hence, the need to apply deictic procedures to bring them to the foreground and allow clear understanding by both interlocutors. The principle of saliency becomes all the more obvious when referential processes are not easily resolved due to the lack of saliency of the referent in the discourse. The following disconcerting sentences open the first chapter of *Tokyo cancelled*, a novel written by Rana Dasgupta in 2005:

> 5) "THERE WAS CHAOS. Will someone please explain why *we* are *here*? What are *we* going to eat? Who has thought of that? Who is in charge *here*? Let me speak to him!" (Dasgupta 2005).

A novel can be regarded as a conversation between a speaker, the writer, and an addressee, the reader. During this conversation, it is essential for the writer to introduce the parameters of the utterances that (s)he intends to convey to the reader. Example 5 shows that the novel begins almost immediately with direct speech, thus using the *in media res* principle sometimes used in storytelling. Consequently, the incipit, which is supposed to help set the place and time of the story, begins with dialogues that are not clearly understandable for the reader. No introduction of characters and place is made by Dasgupta, leaving the reader without any of the parameters needed to interpret the various anaphors. This example is relevant because it illustrates that referential processes cannot be easily resolved if referents are not salient in discourse. At this point in the story, the anaphors *we* and *here* are two prototypes of blurred reference as the referents supposed to be known by the reader (passengers of a flight and airport) have not been introduced previously by the narrator. The referents have not been set in focus as expected. The reader is thus left uncertain due to referential gaps. Saliency is crucial for clear reference and, obviously, meaning effects can be achieved on the addressee when the writer deliberately chooses to circumvent the conventions.

The aforementioned example is interesting in two ways. First, it shows that the principle of accessibility is at the heart of reference and second, it illustrates a case of total blurriness of the referent. However, there are other examples in the novel that show that accessibility is not a binary condition. Example 6 shows a case of intermediate accessibility with the use of *we* by the character introduced as *man.* The pronoun's referent is not clear as no previous mention to a group has been made. Its elucidation relies on the identification of the group that the *man* represents. The context helps understand that the *man* probably works for an airline company and the reader understands the use of *we* as a mark of speaking on behalf of this company.

6) A queue formed, of sorts, at the one open desk where a *man* tried to hold off the snaking, spitting vitriol long enough to find a solution. *We* understand

> Madam it's very late yes the little one looks quite unhappy please bear with us.

In fact, referents have degrees of saliency and Ariel (1994, 27) emits the hypotheses that these degrees are coded within language itself in the form of linguistic expressions. The speaker chooses the relevant expression according to the anticipated degree of saliency of the referent in the addressee's mind. Specific markers are used for specific degrees (see Section 2.4.2 for details on indexical expressions, also called 'referential expressions', and their level of accessibility). This raises the question of the factors that constitute these degrees and Ariel provides an answer:

> "A variety of factors contribute to the degree of accessibility with which entities are entertained in one's memory. [...] I have isolated two main factors [...]: the prominence of the antecedent and the nature of the relation between the antecedent and the Givenness marker, now dubbed Accessibility marker." (Ariel 1994, 28).

Psycholinguistic findings lead Ariel to elaborate her hypotheses on the way the referent (the 'mental antecedent' in her own terms) and the indexical referring expression are articulated. Firstly, she indicates that referents are more or less prominent in discourse depending on the nature of the entity referred to. When referents are humans, or entities encoded as topics or subjects, their level of accessibility is high, compared with entities that do not have such properties. In the case of several competing referents, the level of accessibility is decreased. Secondly, accessibility also depends on the textual distance that exists between what Cornish calls the antecedent trigger and its anaphor. Short distance is a factor of high availability for the discourse referent. This is understandable if we consider that the construction of the referent, via its antecedent, in the discourse is so recent that it is highly available in memory. Consequently, it becomes the first candidate for retrieval. If the reference is made endophorically, then the accessibility of the referent comes from its previous mention in speech. If the reference is made exophorically, then the accessibility is grounded in the situational perception of the speaker (Kleiber 1994, 55).

The accessibility of the referent varies in degrees and Cornish elaborates on this by describing a principle of topic-worthiness used to 'rank' discourse referents in the memory. This notion of 'rank order' depends on the topic-hood of the referent, *i.e.* its topic worthiness and the existing rank order of the referents is constantly updated as the discourse unfolds (Cornish 1999, 159). In other terms, a highly ranked referent is highly accessible. It must be mentioned that the ranking method follows principles established in Centering Theory (CT) (Grosz, Weinstein, and Joshi 1995) which is not going to be detailed in this thesis. However, it will suffice to say that CT applies rules at local and global level, *i.e.* sentence unit or text unit, and proposes a method to identify topical elements of discourse. Just like Ariel, Cornish tries to establish criteria that act on the ranking, and thus the accessibility, of the referent. At text level, he identifies indicators that play a role on the way a referent may be highly accessible. The subject position and the fact that a referent emanates from a syntactically autonomous element, rather than an embedded one, are textual traces showing that referents can be ranked highly in terms of memory representation. Similarly, at discourse level, the fact that a referent is related to “a broader topic of the discourse unit” makes it highly accessible. As there are less accessible referents Cornish also proposes the existence of an accessibility scale (1999, 162). This scale ranges from implicit to low. When discourse is constructed by the speaker, direct reference may indeed be made to referents, be it for a new entity or to retrieve a known entity. In this case, accessibility is high as the referent is sufficiently salient and so the speaker's indexical form hits its targeted referent without any hindrance.

However, there are cases when reference is made indirectly and this has a blurring effect on the identification of the referent by the addressee. Indirect anaphora as described by Cornish (2005, 202) is a referential process in which the identification of the correct referent is done more laboriously by the addressee. He defines the concept, first described by Erkü and Gundel, as follows: “indirect anaphora is any use of the anaphoric procedure which does not consist in straightforwardly retrieving the referent of a prior linguistic mention from within the co-text […] or

of a subsequent one, in the case of cataphora; nor of a referent which is visible and salient within the situation of utterance". This definition highlights the importance of the retrieval process which needs to be straightforward in the case of direct anaphora. Conversely, in indirect cases, access to the discourse referent requires more processing time due to co-textual elements that do not clearly set the discourse parameters. Kleiber (2001) also covers this indirectness aspect by stating criteria that define what he calls "associative anaphora". Firstly, the anaphoric expression introduces a new referent in the form of a new NP. Secondly, a different referent must have been mentioned in the discourse. Thirdly, both referents enjoy a metonymic relationship and the use of the first one naturally implies and constructs the existence of the second one. Hence, the second referent is known when introduced in discourse.

Working memory space is limited as Cornish puts it and this has an impact on the way memory deals with incoming information. Storing information and keeping it with a high level of activation is impossible especially if we consider that the discourse is ongoing and new referents are added to it. Information needs to be structured hierarchically in order to avoid overwhelming information in the mind (Chanquoy, Tricot, and Sweller 2007). A procedure of raising the focus on incoming discourse entities and dropping it for obsolete discourse entities is put in place by the brain. This cognitive process relies on principles linked to the topic-hood of an entity whether it is local or global in terms of discourse: "the local topic-entity is the highest rank of these, but by default the global discourse topic entity takes precedence even over it" (Cornish 1999, 209). An entity which is the topic of discourse remains highly accessible and only the global topic of the discourse may supersede a local discourse topic. Based on experimental results, Cornish indicates that other referents are organised around the discourse topic and that those which are connected to it spatially enjoy a high level of accessibility as they are foregrounded. Those that "are spatially dissociated from the entity in the foreground […] assume a background status within the [discourse] model two

clauses after they are initially introduced, and where no intervening reference to them occurs" (Cornish 1999, 209).

The accessibility of the referent is the articulating point of the discourse-cognitive approach as it determines whether the referential procedure is deictic or anaphoric. In other terms, the speaker's choice of the referent depends on hi(s)her perception of the addressee's understanding and awareness of the entity to refer to. The more salient a referent, the less it requires to be put in focus and thus the use of the anaphoric procedure. The less salient a referent, the more the speaker will need to put it in focus and thus the need for a deictic procedure arises. In section 2.4.2, the focus is placed on how the saliency of referents correlates with accessibility markers according to various degrees. Prior to this, the focus is raised on the way the referent is retrieved.

2.3.3 The process of referent retrieval

Referent retrieval is a process that does not rely primarily on the distinction between situation and text. Cornish (1999, 116), Ariel (1994, 27) and Kleiber (1992, 619) reject the vision in which endophora matches anaphora and exophora matches deixis, that is to say that deixis is about reference outside the text and that anaphora means reference within the text. The discourse-cognitive approach is grounded in the construction of a discourse model in which text and situation "are but domains in terms of which the relevant discourse representation is constructed" (Cornish 1999, 117). Endophora and exophora are not denied in this approach but instead they are seen to play a facilitating role in the move from speech, in the form of text, to discourse. Endophora and exophora can be seen as means for the retrieval of referents from the context of utterance. For Cornish (2010, 219) "there exists different 'fields' or domains of reference on which both deictic and anaphoric procedures may operate". He distinguishes four domains—explained in the next paragraphs—in which referential processes are applied and which are part of the utterance context. As well as the two aforementioned domains, deictic and

anaphoric procedures may also stem from long-term memory and previously constructed or anticipated discourse.

There are several retrieval processes depending on the domains of reference. Canonical deixis and exophora (which can be part of anaphora in Cornish's own account) operate from the utterance situation. Textual deixis and endophora (also part of anaphora) operate in the domain of the co-text. Discourse deixis or anaphora are retrieval processes applied to discourse already created or anticipated. Finally, anadeixis (to be defined further down page 53) and anaphora are applied on the domain of long-term memory shared between interlocutors. The following paragraphs cover each of the retrieval processes with examples provided by Cornish.

'Pure deixis' can be seen as the prototype of deixis as it bears two properties: mundane reference and new information. Example 7 shows that *that* pro-form is used to point at an element of the utterance situation, *i.e.* the real world, while making it the new topic of the unfolding discourse of the speaker.

7) Hey, look at *that*! [The speaker gestures toward a strange bird perched on a nearby tree] (Cornish 2011, 4).

In the utterance situation, anaphoric exophora is at work to maintain the focus on an already salient discourse referent. Cornish's example is particularly explicit:

8) “[A and B turn a corner on the pavement, and suddenly find themselves face to face with a rather large dog] A to B: Do you think *it*'s friendly?” (Cornish 1999, 112).

Here, the pronoun *it* is used to retrieve the referent via the utterance situation. The entity dog is seen by both interlocutors and A uses the pronoun *it* because (s)he knows that B also has the dog in mind. This means that the discourse referent “dog” is salient in both interlocutors' memories. So the retrieval process follows the anaphoric procedure since the pronoun is used to maintain the referent in focus. As surprising as it may seem, exophora, in this case, is part of anaphora. *Via* exophora

the link between speech and discourse is established, even though there is no textual antecedent. 'Antecedentless anaphora', as it is also called, appears to be contradictory if it is not coreferential: “If there is no 'co-textual' antecedent for the anaphor, there can be no intra-textual relation—hence no anaphora” (Cornish 1999, 116). His claim is that even in this case of exophoric reference, there is an antecedent in the form of discourse representation. The context of utterance partakes of the construction of the referent in the interlocutors' mind before it is retrieved with the anaphor. Consequently, exophora acts just like endophora in its retrieval process of an already salient entity. This is a similar view to the one developed by Halliday and Hasan and Fraser and Joly. It is reported in Section 2.2.2, in which endophora is seen as the conceptual equivalent of exophora. For Cornish, exophora may belong in some cases to anaphora and, to be more accurate, it could be referred to with the coinage “exophoric anaphora”.

'Textual deixis' belongs to the domain of the co-text. It simply corresponds to the utterances in which the speaker consciously refers to previously mentioned speech. The words that have been written or pronounced are actually retrieved as words in their morphosyntactic forms. Their semantic value is put in the background and a new discourse focus is established. Cornish (2010, 219) provides the following example:

9) “A: Our rhododendrons are in blossom right now. B: Oh really? How do you spell *that*, by the way?”

In this occurrence of *that*, the initial addressee refers to the spelling of the word form *rhododendrons* rather than its semantic value. The presence of the expression *by the way* is a textual clue that shows the break in continuity from the current discourse on the blossoming of the flowers in order to shift the semantic focus to the actual spelling of the word, thus making the reference autonymic and new. The answer examplified in 10 shows a case of textual anaphora.

10) “B: I know *it*'s got three “d”'s” (Cornish 2010, 219)

In this example, the autonymic reference carried out with *it* refers to a known textual element of the discourse.

'Discourse deixis' belongs to the domain of discourse already created or anticipated. The reference is made upon the discourse which has been or is going to be constructed (cataphora). The speaker refers to an utterance for which hi(s)her discourse representation is already constructed but not yet known by the addressee. It is different from textual deixis as far as the referent is concerned. Textual deixis refers to a precise antecedent whose identity is clearly established in the memory. Discourse deixis is a process to refer to what can be called the "co-discourse", that is "the surrounding discourse which has just been constructed (or which is at the point of being constructed), which is operated upon by the addressee to appropriate the intended referent" (Cornish 2010, 220). See the following example:

11) "A: Listen to *this*: a man went into a butcher's shop one day wanting to buy a whole sheep, and..." (Cornish 2010, 219)

This example gives evidence that *this* refers to the upcoming utterance and not the man in the shop only. It is a reference to the complex discourse representation of a set of entities that constitute a short story rather than a single non-accruable referent. It is the case of the upstream retrieval procedure used to introduce a new discourse entity. The imperative form with the verb *listen* gives evidence of the refocusing of discourse on what is to be said and its discourse representation.

'Discourse anaphora' falls within the same category as discourse deixis insofar as the reference to a discourse representation is concerned. The distinction comes from the fact that it is carried out on a previously created representation in discourse and not a new one. Example 12 is an illustration of this process.

12) "...Would you believe *it*?" (Cornish 2010, 219)

In this case the pronoun is used to refer to a discourse representation which is the result of several entities having been passed into discourse. It does not correspond

to an exact entity in the text, but, rather, to the discourse constructed up to that point.

The fourth domain in which retrieval processes operate is the shared long-term memory and Cornish coins the term 'anadeixis' for the specific process that involves both deictic and anaphoric procedures. In this kind of process, the speaker provides a dual perception of the discourse representation (s)he refers to. The reference is made upon an already existing representation present in the long-term memory of the addressee and it is concurrently made to shift the focus of discourse towards this representation. The long-term aspect necessarily implies the existence of the representation model in the interlocutors' memory. However, the fact that it is long established requires a retrieval process that treats it as new to make it the topic.

13) “A: Do you remember *that holiday we had two years ago in the Bahamas?”* (Cornish 2010, 219)

The segment in Example 13 is processed as already known and yet new. It is, by nature, already existing as both interlocutors experienced the trip together. However, it is not salient in the unfolding discourse as the event's representation is stored in long-term memory and needs to be brought back to the surface. So the retrieval process implies a dual process of memory identification and discourse refocusing.

The long-term memory can also support anaphora as in Example 14, which is the answer for the question asked in 13.

14) “B: I do indeed! *It* was really awful, wasn'it?” (Cornish 2010, 219)

In this example, the pronoun refers to the long-term memory of an event that has just been raised again in the conversation. It is not a reference to an entity, nor to an on-going discourse construction, but to a set of known memory elements that constitute an experience from a time long past.

To recapitulate the various retrieval processes of the discourse referents, the following table is proposed.

Retrieval processes: deixis v. anaphora	**Domains of reference**
Canonical deixis: exophora	Utterance situation
Exophoric anaphora	Utterance situation
Textual deixis	The co-text
Endophoric anaphora	The co-text
Discourse deixis	The co-discourse
Endophoric anaphora	The co-discourse
Anadeixis	Shared long-term memory
Endophoric anaphora	Shared long-term memory

Table 2: Retrieval processes in relation to the four domains of reference

Table 2 recapitulates the four fields identified by (Cornish 2010, 219) on which retrieval processes are based. As argued previously, deixis and anaphora are to be distinguished in terms of mental representation of the referent. The textual and situational aspects are not used to distinguish both procedures but, instead, partake of a set of domains of reference that together compose a multi-dimensional coordinate space for the retrieval processes.

The act of retrieving a referent from these domains is carried out according to the focus to be given to the referent. Should it be new, a deictic procedure is put in place and, should the focus already be high, an anaphoric procedure follows. It is essential to mention that, in this system, anaphora and deixis can both apply to situational and textual domains as well as to those of discourse and long-term memory. The retrieval of the referent is done according to its focus and thus its degree of accessibility. The four domains in which the processes operate constitute the context and, as such, they dynamically transform text into discourse. Depending on their source domain and the saliency of the referents, the retrieval processes connect the text to the discourse. The fact that the processes are characterised by an intensity of saliency and a direction of reference makes them comparable to vectors defined as mathematical objects representing a quantity and a direction. It could be argued that a specific text unit is linked to a discourse referent via a dynamic force whose direction and intensity rely on the contextual space made of co-text, situation, co-discourse and long-term memory coordinates.

In the next section, we provide a model for the interpretation of referential procedures.

2.3.4 The reference system: a dynamic model

This section is an attempt to draw a 'flat' representation of the reference system with arrows that represent accessibility. Figure 4 tries to produce a comprehensive overview of how a referential process is produced from the moment of utterance to its realisation in terms of cognitive referent. It is inspired by the often quoted diagram drawn by Halliday and Hasan (1976, 33) and the discourse cognitivist approach and more particularly Figure 1 in (Cornish 2011, 3). This approach does not eliminate the situational/text paradigm, but rather, it embeds it within a more global system made of different components: text, context and discourse.

There are three components in the system. Discourse corresponds to the cognitive level of speech. Text corresponds to the realisation of speech in terms of vocal or textual signs and symbols. Context includes the various domains in which reference occurs and the types of referential processes that are at work. The two main horizontal arrows show the progression of text (*via* speech) and the construction of discourse that parallels it. The dividing line corresponds to the shift in reference from new to known information in the sense that, first, an entity is new in the discourse and, second, it is retrieved as a known element.

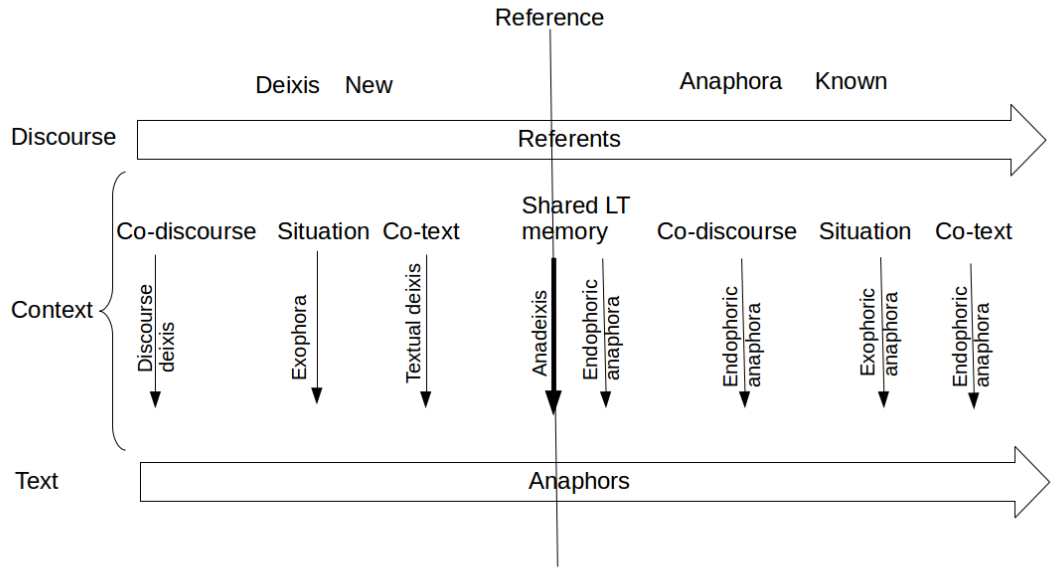

Figure 4: Referential system: processes at work from discourse to text

The idea is to represent referential processes dynamically with vertical arrows. In the diagram, they correspond to the speaker's point of view as their direction indicates a move from discourse to text. The speaker's construction of an utterance commences at discourse level since the speech act is naturally initiated cognitively before resulting in a string of signs and symbols. After the cognitive identification of the type of referent (a new or known entity), the speaker chooses the coordinate that guides the retrieval process. In parallel, the saliency of the referent determines its level of accessibility. As saliency is gradable, several degrees can be applied to the retrieval process. Saliency could be symbolised with the length of the arrow. Should accessibility be low, the retrieval process would be blurred and it could be represented as a short arrow.

What the figure shows is how a referential process is dynamically constructed and how it conveys the referent to its anaphor. When the speaker cognitively initiates an utterance the entity appears at discourse level first and it is new. When the utterance is pronounced, the entity is actualised in speech—hence the downward arrow—with an anaphor and *via* a specific context such as exophora. If the speaker

cognitively initiates a known entity, (s)he actualises it with an anaphor *via* another context such as endophoric anaphora. When an utterance contains a reference made to a shared long-term memory, not already present in the ongoing discourse, the entity is actualised via anadeixis (the border line between anaphora and deixis).

From the addressee's position, what matters is the necessity to access the referent pointed at in the speaker's speech as accurately as possible. Referent retrieval is the main task of the addressee and so the referential processes are inverted. There is a move from text to discourse. The addressee's endeavour is to re-construct the referential process within the coordinate space and based on the anaphors given in the speech. The arrows could therefore be presented in the opposite direction of those presented in Figure 4. This dual view follows Cotte's (1993, 46) idea of the direction of reference between speaker and addressee that determines the type of procedure. However, it does not link the situation/text and the new/known paradigms to a specific direction. Instead, it enlarges the situation/text dimension and articulates it within the discourse paradigm.

Stirling and Huddleston (2002, 1454) also notice the close relation between deixis and anaphora. In their account, the nature of the relation relies on the situation/text paradigm. For them, some markers in a given utterance may refer simultaneously to the situation's parameters and to a previously mentioned entity. Even though our approach is guided by a different paradigm, their view shows that some referential processes might be of a dual nature. This duality may be found in the notion of anadeixis which is the frontier between anaphora and deixis (see Section 2.3.3 for a definition of the notion). In establishing a pervasive border between both referential procedures, Cornish (2010, 221) puts forward the notion of a scale of anaphoricity and deicticity in which he places indexical expressions. What matters is the fact that reference is seen as oscillating between the deictic and anaphoric poles of the scale. The role of indexical expressions is described in Section 2.4 but the point is that reference is variable in nature and that it is not

clear cut. It is this scale, and the blurriness it permits, that is used on purpose by writers such as Dasgupta (2005) (see Section 2.3.2) to mislead their readers. It is also this scale that learners do not master completely and that leads them to errors and this issue is covered in Chapter 3. The description of the reference system is an attempt to show how discourse is converted into text. In other words, discourse turns into speech in the form of text and, as such, text represents an important field to explore in order to gauge the gradability of the aforementioned scale. The text is used to explore the predicational context that governs indexicality. Section 2.4 focuses on the end-products of reference that are referential expressions in their predicative environments.

2.4 Referential expressions and predicational context

In this section, we remain at an onomasiological level and we show how referential expressions are used within the discourse-functional approach. The predicational context that surrounds the form plays an essential role in the process of referent integration into discourse. In his publications (see Kleiber 1991, 1992), Kleiber defends a cognitive approach in which anaphora resolution relies on memorial processes. His view, shared by Cornish and Ariel, comes as a response to purely linguistic views of anaphora resolution as described in Section 2.1 but it is important to show that it is not constructed in opposition with the textual approach. Kleiber underlines the need to take heed of the textual dimension in terms of linguistic criteria:

> "There is nevertheless a danger in moving too quickly towards an approach that would be almost totally pragmatic. It would be that of forgetting the linguistic, syntactic and semantic criteria, so as to retain only the cognitive facts as decisive. It would thus reinforce the commonly accepted poor idea that identifies context as the main factor. This theory does not explain much when it defends the idea that interpretation is grounded in text content and general or specific extra-linguistic knowledge, which it presupposes." (Kleiber 1994, 35, my translation).

What Kleiber wants to say is that the cognitive approach must avoid the bias of exclusively taking into account constraints like 'conceptual consistency', that is to say, the fact of verifying the coherence of reference with knowledge. The problem

is that in this case, referential interpretations are always correct and this is because such theories only take correctly formed examples. His claim is that when observing incorrectly formed examples the purely cognitive approaches find their limits. He gives the following example of indirect anaphora:

15) * “We arrived in a village. This/that church was located on a hill” (Kleiber 1994, 37, my translation).

In this erroneous example, indirect anaphora cannot work due to the use of demonstratives both in French (*cette* in the French version) and English. In fact, the correct form should be article *the* or its cliticised *l'* version in French. Nevertheless, the cognitive interpretation allows the reader to identify the referent. The problem of acceptability of the utterance in 15 shows that some specific linguistic, syntactic or semantic elements are still required to have a clear interpretation of the reference, thus the need to deal with the anaphoric expressions themselves. In this case, it is clear for Kleiber that it is also crucial to analyse how indexical expressions function (1994, 39).

Cornish also regards the insertion of indexical expressions as essential in the interpretation of reference:

> “Rather than the 'antecedent', under the traditional view of anaphora, then, it is the anaphor itself in conjunction with its immediate predicative and utterance context which determines its interpretation and reference. It is the immediate context of the indexical expression whose interpretation is at issue and not the antecedent expression (the antecedent-trigger in my usage) which carries the greatest weight as far as this is concerned.” (Cornish 1999, 68).

Within this conception, the role of the antecedent expression is minimised and considered with some degree of irrelevance. However, the indexical expression is given full consideration when interpreting referential processes. The indexical expression is made of the anaphor itself and its predicative and utterance context. A set of elements constitutes the predicational context and, as such, it is the subject of analysis. In this section, the predicational environment of indexical expressions receives the main focus and the objective is to identify the linguistic constraints that

are applied to referential processes. The various components of the textual environment are reviewed including the roles of *this* and *that*.

2.4.1 The predicational environment and its constraints

The indexical segment in Cornish's terms is the place where the interpretation of an anaphor can be achieved. The text features that surround the anaphor play a role in the resolution of the referential procedures, *i.e.* the identification of the discourse referent. The first component of the indexical segment is the predicator that, according to Cornish (1999, 69), governs the anaphor. The predicator—usually a verb—is "a semantico-grammatical unit rather than a purely grammatical or syntactic one". It produces an array of semantico-grammatical constraints that point towards a particular referent in the discourse. In so doing, it acts as a sort of filter used by the addressee to exclude all potential referents but the one intended by the speaker (Cornish 2010, 230). So, one important stage is to understand what the process of matching the constraints to a discourse referent entails.

Prior to this, it is necessary to take a closer look at discourse entities because referents represent entities whose nature impacts the grammar used to determiners or refer to them. Discourse entities have been the subject of studies reported in (Cornish 1999, 47–51). Their essence is that their classification can be made according to five categories of 'referent-order entities' (Lyons 1968, 347) that are numbered from zero to four. The typology order reflects the level of abstractness that entities may represent. Cornish provides the five examples that are quoted underneath in order to illustrate the categories (Cornish 1999, 48):

16) That tree-like shape that you can see on the horizon is actually *a building*.

17) *The Empire State Building* is the tallest in New York.

18) *The building of the Channel Tunnel* took seven years to complete.

19) *John believes that building a tunnel 30km long under the sea* is impossible.

20) Buy a box of “Maz” and you will receive a free copy of War and Peace. Note: *this offer* is completely genuine.

Example 16 illustrates zero-order entities. They are entities that are not entities and that actually correspond to the designation of predicates themselves. Here the predicate *a building* is used as a property for the subject expression. Contrary to 16, Example 17 refers to an actual existing building which is uniquely identified. In that context, the italicised expression denotes what Lyons calls a discrete object with “a physical existence” (Lyons 1968, 424). This makes it a first-order entity. In 18, the segment in italics refers to an event or process. *Building* is an activity, which puts it in the second category of order-referents. In example 19, the segment, a third-order entity, corresponds, not to the process of building, but to an actual thought attributed to John. As such, it is a proposition that can be asserted or denied. Finally, *this offer* in 20 is a fourth-order entity that constitutes an illocutionary act by which the speaker performs the act of making the free copy of the magazine an offer.

The question of what is entailed by the process of matching the predicator's constraints and the referent can now be answered. The construction of discourse by interlocutors implies the assignment of specific order-entity categories to anaphors. To do so, the predicational context of the utterance is set so that the type of entity is specified for the anaphor's referent. For Cornish (1999, 70), the predicator governs the anaphor with its semantico-grammatical properties and allows the assignment of a “referent-order” potential assumed by the anaphor. In this process, grammatical features are used to further filter in or out possible anaphor interpretations. Cornish points aspect, tense, and various types of modality for their role in the narrowing down process of interpretation (Cornish 1999, 84). These grammatical elements act on the indexical segment in which the anaphor occurs. For illustration's sake, the following example found in (Cornish 1999, 86) shows how the aspect on the verb can have an influence on the interpretation of an indexical expression introduced by *that*:

21) John doesn't usually like red wine, but he certainly seems to be enjoying *that Italian one*.

Here, the *be+ing* aspect associated to the verb *enjoy* renders the utterance specific and synchronised with the moment of utterance. This contrasts with the first utterance where /wine/ is referred to in its generic form. So, the fact of adding the *be+ing* aspect onto the verb leads to interpret *that Italian wine* as a discrete entity in the form of a specific bottle of wine at a particular moment in time. Were the verb *enjoy* used with the present simple auxiliary *does* as in "…but he does seem to enjoy that Italian one"*,* the interpretation would be different insofar as the entity referred to would be generic. *Italian one* would refer to a sub-class of wine, not to a single occurrence of his drinking a wine he enjoys. This example and its manipulation gives evidence of the impact of tense and aspect on the possible interpretation of an indexical segment. However, it is only a possible interpretation at this point in the construction of discourse, since it requires validation with the discourse referent whose actualisation is conditioned by the existence of the mundane referent. In the example, the aspect -ING entails the presence of the referent in the situation. The speaker places the focus of the addressee on the new fact (in the ongoing discourse) represented by *Italian one (wine)*. This is a case of deictic reference but, should the alternative utterance be considered with the present simple and the auxiliary *does*, then John might not necessarily belong to the situation of utterance. The speaker might be making a reference to a type of wine rather than a specific bottle on a table with John next to it. In that case, the reference would be anadeictic in nature in the sense that the two interlocutors refer to a type of wine known to both of them and that the speaker presents the *Italian one* quality as new for the addressee.

The process described in the previous paragraphs shows how the predicational context sets constraints on the indexical expression so that it is oriented towards a possible interpretation of its referent. The purpose of this operation is to bring forward a potential interpretation of the indexical expression to finally integrate the utterance into the unfolding discourse (Cornish 1999, 98). The integration into

discourse is a crucial point that also relies on the choice of the indexical expression according to the utterance context and thus the type of deictic or anaphoric procedure.

As explained in Section 2.3.2, discourse referents are given a level of accessibility via various domains. When the discourse unfolds, the predicational context prepares the ground for a possible interpretation and the indexical expression, depending on its accessibility degree, brings forward a candidate for reference. In order to obtain coherence, the referent candidate and the potential for interpretation created by the predication must match. Therefore, the choice of the right indexical expression is crucial.

2.4.2 Indexical expression and discourse integration

Indexical expressions are the subject of several surveys among which the views of Ariel and her linguistic and cognitive approach shall be reviewed here. The interest of her theory is that it grounds itself in the speaker-addressee relation and the need to make a discourse referent retrievable (Ariel 1994, 3). She refers to evidence that "different anaphoric expressions trigger different processing procedures" and so she proposes "that referring expressions (anaphoric expressions among them) signal specific (relative) degrees of accessibility of mental representations" (Ariel 1994, 4). She clearly specifies the existence of a link between referring expressions and the level of saliency of their referent. As there are different degrees of accessibility of a referent (see Section 2.3.2), Ariel claims that indexical expressions are a function of this accessibility. In other words, specific indexical expressions yield a specific degree of accessibility. Consequently, it can be said that they determine the retrieval procedure of the referent. She presents her Accessibility Marking Scale as follows (Ariel 1994, 30):

> zero < reflexives < agreement markers < cliticised pronouns < unstressed pronouns < stressed pronouns < stressed pronouns + gesture < proximal demonstrative (+ NP) < distal demonstrative (+ NP) < proximal demonstrative (+ NP) + modifier < distal demonstrative (+ NP) + modifier < first name < last name < short definite description < long definite description < full name + modifier.

Most of the degrees expressed in the scale are self-explanatory but some of them call for comments. Firstly, the cliticised pronouns placed in the lower end of the scale do not exist in English but exist in French, *e.g.* the pronoun *le* that becomes *l'* in a sequence such as "Il l'a vu". Another example may be given with *la* where the cliticised pronoun cannot be detached from the verb in its post-verbal position as in: "Prends la lui." In Spanish, cliticised pronouns also appear such as in "Dáme*lo*" where *lo* is attached to the verb form. Secondly, the proximal or distal demonstrative can be preceded by a modifier. Ariel provides an example in (Ariel 1988, 84): "this/that hat was bought last year". Here, the demonstratives, considered in their distal/proximal dimension (see Section 2.5.2 for a closer look at the demonstratives and their values), render the referent more salient if they are followed by an NP and a modifier such as a clause which provides further information on the entity talked about.

Cornish follows Ariel and adds two elements to the scale. Firstly, he frames the scale between the two poles that are deixis and anaphora (see Figure 5). Secondly, he uses generic labels to refer to groups of indexical expressions categorised as pronouns, adverbs and NPs. His labels show the distinction between proximal (P)- and distal (D)-related groups as well as the successive nature of the two properties —P always comes before D within each adverb and NP category. This scale reflects the degree of variability that a form may have depending on contexts (Cornish 2010, 221).

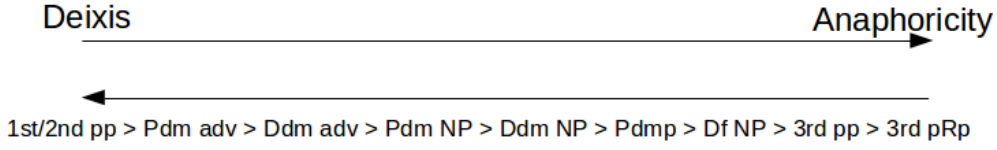

Figure 5: Cornish's scale of deicticity/anaphoricity and grammar categories of forms

In this classification, *Pdm NP* corresponds to *this NP,* that is to say the Proximal demonstrative (Pdm) followed by a Noun Phrase (Cornish 2011, 5). Conversely, Ddm refers to the distal (D) demonstrative *that* which can be positioned left of an adverb, a Noun Phrase or nothing and plays the role of pronoun (dmp). The scale shows that first and second person personal pronouns correspond to 'pure deixis' when third person personal pronouns equal to anaphoric procedures. The positions in between reflect the variability that morphological forms such as *this* and *that* may have. A demonstrative will be favoured over a definite NP (Df NP) or a third person personal pronoun construction (3rd pp) when reference tends to be deictic and, conversely, the last two—including third person reflexive pronouns (3rd pRp)—will tend to be used in anaphoric procedures.

Indexical expressions are connected to specific referring processes and it is the speaker, in the act of utterance, who chooses the form according to the type of reference and the level of accessibility (s)he wishes to give to the referent. Ariel suggests that:

> "Natural languages code degrees of Accessibility in memory in their Givenness expressions. Thus, when the speaker wishes to refer to some Given entity, she must choose a linguistic expression which signals that degree of Accessibility with which she judges the addressee to entertain it in his memory." (Ariel 1994, 27).

Here, Ariel focuses on anaphoric expressions and she points out the indexical as the text component used to link up to the discourse referent. The indexical expression is used as a recipient of the level of accessibility and it allows the interlocutors to retrieve the referent (as seen in Section 2.3.3). So, the indexical expression is the result of a predicational process in which entity properties are narrowed down to a type of order-entity. In conjunction, the type of referential procedure needs to bring the relevant referent in focus and it does so with the use of a specific indexical expression. It entails that the indexical is the place where the compatibility between the referential procedure and the sense created by the predicational context, up to the point of the utterance of the indexical expression, is at stake. This compatibility ensures the integration of the utterance into the discourse. When compatibility is

reached, there is coherence between the utterance and the previous utterance already present in the discourse.

The reference system functions at discourse level with referents and at text level with indexical expressions which result from the predicational and the referential processes. This represents the referential framework in which the demonstratives function.

2.5 The case of *this* and *that* in the reference system

As suggested in Ariel and Cornish's scales, the role of the demonstratives is central in the reference system. This section is a semasiological approach of the two forms and it shows their various properties at syntactic and discourse levels. It offers a focus on *this* and *that* according to several levels of interpretation. We show how natives use them in utterances to build their discourse.

2.5.1 The syntactic role of *this* and *that*

This thesis focuses on the referential processes involved in the use of *this* and *that*. So only their use as demonstratives is analysed here, which excludes the uses of *that* as relative pronouns and complementisers. Nevertheless, in the course of some experiments whose setups are presented in Chapters 6 and 7, these functions will be taken into account. The two morphemes have received many labels in the literature depending on their semantic and syntactic values; from simple determiners to adjectives, from suppletive pronouns to completive pronouns depending on their position in the sentence and their function. The following is an attempt to clarify what the various labels encompass.

Quirk *et al.* (1985, 217) make the distinction between the determiner and nominal functions. Biber *et al.* (1999, 274, 349) use the term *demonstrative* but they consistently specify the notion with the terms *determiner* or *pronoun*. Stirling and Huddleston (2002, 1504) use the same terminology as far as their function is

concerned but, contrary to Biber *et al.*, they prefer the terms *dependent* or *independent* in relation to the role of NP which they might endorse. Demonstratives can be found in two syntactic positions. In front of a noun phrase, they will be determiners, *i.e.* dependent on a head, and when acting as the head of a noun phrase they will be pro-forms, *i.e.* independent. However, these two terms lack accuracy when the position of the forms in relation to the substantives they denote is taken into consideration. On the contrary, Biber *et al.*'s terminology allows for a clear distinction between the head and pre-head positions.

Some French linguists have also delved into the matter. Fraser and Joly (1979, 102) consider *this* and *that* as suppletive and completive pronouns due to their distinctive function regarding the need or the absence of a substantive with which the process of location in time or place is operated. In a more recent survey, Lapaire and Rotgé (1991, 50-51) also use the functional distinction proposed by Biber *et al.* but they use *pro-form* instead of *pronoun* because when *this* or *that* are used as heads they may refer semantically to more than just a Noun Phrase. Indeed, referents previously denoted by entire clauses in speech may be referred to by *this* or *that,* and "not all pro-forms are pronouns" (Stirling and Huddleston 2002, 1462). Lapaire and Rotgé (1998, 57) also discuss the sense of the term *demonstrative* and criticise the focus it places on the denoting function of the forms. They point out the fact that such an approach tends to elude the importance played by the speaker in the act of reference. Therefore, they favour the term *deictic*, not for its etymological meaning, but rather for the significant role of the speaker in referential processes this vision gives (Lapaire and Rotgé 1991, 59). In this thesis, the term 'demonstrative' will be favoured in order to remain consistent with the distinction between deixis and anaphora covered in Section 2.3. Under the view presented in that section, the fact that demonstratives can be of anaphoric or deictic value would make the 'deictic' denomination more confusing than that of 'demonstrative'. This choice also reflects the importance of the semantic value of the forms as opposed to their syntactic value. This is the reason why we also retain the terms *determiner* and *pro-form* to specify their syntactic functions.

What the aforementioned labels tend to describe is the role played by the forms in their close syntagmatic vicinity. This may give the idea that only local constraints apply in the construction of indexical expressions with demonstratives. However, the indexical expressions also interact with other syntagmatic constituents of the sentence. Table 3 summarises the syntagmatic configurations for demonstratives:

Syntactic function	Syntax
Determiner	TH-* + NP
Pro-form	TH- + VP
Pro-form	VP + TH-

* TH- encodes both *this* and *that*

Table 3: Syntagmatic configurations for demonstratives

Table 3 shows how the forms can be found at sentence level. It is relevant to note that, as a pro-form, *this* or *that* may be found in various places in the sentence. They can be part of a noun phrase acting as a subject for a verb phrase and they can be the object of a verb phrase.

The observation of demonstratives also gains significance when their distribution is observed at context level, *i.e.* the co-text, the domain of reference, the discourse already constructed up to the point of their occurrence and other elements partaking of the utterance parameters. A significant element for the description of their position is also added by Cotte (1993, 58) when linking the concept of referent to the one of co-presence. For him, *this* only appears when anaphora occurs shortly after the first use of the referent. He describes two constraints. The anaphor and its referent must be in two distinct utterances of the speaker but they must both be present without anything in between. In addition, he notices that after the construction that makes the referent appear, the next process of pointing to this referent is often carried out with a demonstrative before being done again with another form:

> "The salient referent, retrieved from the contiguous utterance, is often introduced by a demonstrative, before a different form is used to refer to it again" (Cotte 1997, 157 my translation).

This means that the following order of appearance applies:

Utterance 1: NP → Utterance 2: Reference via TH- → Utterance n: Reference via other device.

Cotte (1993, 57) provides an example:

> 22) "The solution in the second pattern is to keep the constituent INTENSIFIER + ADJECTIVE intact and to interpose the indefinite article before the noun. *This* does not prevent two successive main accents when the intensifier is monosyllabic and the adjective is monosyllabic or is stressed on its first syllable, but *it* does prevent a succession of three, …"

This series of utterances follows the order given above. Cotte interestingly manipulates the utterances by swapping the personal pronoun and the demonstrative. His conclusion is that, where the personal pronoun is possible in both places, it appears that the demonstrative is not possible in the last utterance and can only be placed in the immediate utterance that follows (Cotte 1993, 58). Stirling and Huddleston (2002, 1507) also mention this process, which they call anaphoric chains. In an example containing two utterances, they show the same order as the one described by Cotte:

> 23) "He discovered that she had slept with several other boys before him. *That* shocked him a good deal, and they had a quarrel about *it*."

Here in their view, *that* is anaphoric to the preceding clause but antecedent to the following *it*. Even though the analysis is grounded in coreferentiality between the antecedent and the anaphor, it still highlights the syntagmactic property which spans several utterances.

So, at the level of local, sentential or contextual syntax, it appears that the use of the demonstratives follows constraints that are not only grammatical such as the agreement constraint. As seen in Section 2.4, other constraints, whose evidence is to be found in referential and predicational processes, influence the distribution of demonstratives in natural language. The choice of the forms follows a series of procedures that are connected to the referential framework and that also depend

on the nature of the forms themselves. As expressed in Section 2.4.2, the indexical expressions are the recipients of referential procedures conveyed by the demonstratives in their own way and this is what creates sense in occurrences.

2.5.2 Building sense with *this* and *that*

The traditional sense attributed to the demonstratives is based on their proximal/distal distinction. Quirk *et al.* speak about near and distant references (Quirk *et al.* 1985, 217). For Biber *et al.*, the demonstratives in their determiner and pro-form functions also specify whether the referent is near or distant in relation to the speaker (Biber *et al.* 1999, 272, 347). They ground their analysis in the situational/textual paradigm (see Section 2.2.2) as "the reference of noun phrases with demonstrative determiners may be established on the basis of either the situation or the preceding or the following text" (Biber *et al.* 1999, 272). In their chapter on deixis and anaphora, Stirling and Huddleston approach the demonstratives with the same distinctive view of distance v. proximity (Stirling and Huddleston 2002, 1504). They also recognise their referring value in the way demonstratives are used either anaphorically or deictically. In doing so, they assimilate anaphora to endophora and deixis to exophora.

However, Biber *et al.*'s distributional analysis on the pronoun form shows evidence of the limiting character of this view as they observe that a certain number of instances of pronouns show "the lack of a consistency with the pattern of proximate v. distant form". For them, this indicates that "proximity is insufficient to account for the distribution of the demonstrative pronouns" (Biber *et al.* 1999, 349). Fraser and Joly (1979, 125) also point out the fact that some uses cannot be explained with the distal/proximal view. They provide several examples in which distance is not the criterion to refer to an entity in the discourse. In the following example, the speaker holds a pair of glasses which belongs to the addressee:

24) "I bet you don't need *those* glasses … If I were you I'd just throw *those* glasses away." (Fraser and Joly 1979, 125)

According to them, the contrast with *these* cannot be explained with the notion of distance since the speaker is holding the glasses. For them, the notion of speaker's sphere helps interpret the use of *those*.

The resulting interpretation of a demonstrative form under the traditional approach is thus carried out in two stages. Firstly, the referent is identified within the text or within the situation of utterance and, secondly, it is specified as being close to or distant from the speaker. This approach, as limited as it might be in light of Section 2.4, highlights the fact that there is a distinction between the referential process that leads to the demonstrative and the internal value of the demonstrative itself.

As far as reference is concerned, the point has been made that the situational/textual distinction does not suffice (see Section 2.3). The demonstratives are in fact selected as part of referential processes that involve the need to either point out new information about a discourse referent or retrieve a discourse referent already in focus in the unfolding discourse. They encode the assumed level of accessibility of their discourse referents (Cornish 1999, 52). This accessibility could be compared to force in the sense assigned in physics. This accessibility applies to the referent thanks to the demonstrative and it links the demonstratives to the referent by invoking various types of retrieval processes as shown in Figure 4 page 56.

The internal value of the demonstratives also plays a major role in their being selected by speakers. Demonstratives are deictic in the sense that their referentiality fluctuates according to the place and time of the utterance (Lapaire and Rotgé 1991, 57). Fraser and Joly add that deixis—in the sense of reference in their work —is a process for spatio-temporal localisation in which the speaker is the centre (Fraser and Joly 1979, 105). The speaker uses the demonstratives not only to signify that there is an observation, but also to re-actualise it punctually in the *here* and *now* of the utterance. This Guillaume-inspired statement points out the implicit process that is at work. *This* and *that* act as “actualisers” by not only pointing to the

referent but also by positioning the referent in the *here* and *now* of the speaker's speech. For them, *this* and *that* regulate the extensity of the noun and, in doing so, the notion expressed by the noun is actualised (Fraser and Joly 1979, 156).

The demonstratives have the deictic ability to shift in meaning and the speaker is considered at the centre of the situation of utterance (Fraser and Joly 1979, 110, Lapaire and Rotgé 1991, 59). This brings the discussion to the system that defines the person and Fraser and Joly use the notions of the "moi" (*self*) and the "hors-moi" (*non-self*) to link it with the *hic et nunc* and to form a triangular configuration of three interacting representations: *self* or *non-self*, *here* or *there* and *now* or *not now* (Fraser and Joly 1979, 112). The representation of time and place always depends on the speaker. The use of demonstratives by the speaker is carried out within the configuration so that *this* corresponds to the speaker's self and *that* corresponds to the speaker's non-self. The full referential process could be summarised as follows: a speaker addresses an interlocutor to refer deictically or anaphorically in relation to a particular domain of reference. Once the referent is made clear in the discourse, the speaker uses the form as a marker to show how the referent is to be viewed in relation to hi(s)her personal sphere.

With their notion of speaker's sphere Fraser and Joly (1980, 26) provide a diagram (see Figure 6) of the microsystem that, according to them, accounts for the use of *this* and *that*.

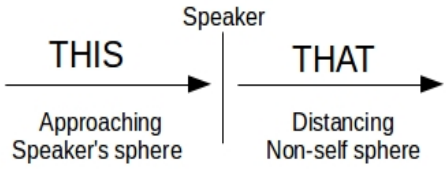

Figure 6: Fraser and Joly's representation of the microsystem organising *this* and *that*

To admit this view shows that once the discourse referents are clearly established in the interlocutors' minds, they are shifted by the speaker from one sphere to the other according to hi(s)her interest in them. This way of marking referents allows the speaker to give specific meaning effects to the referents. Fraser and Joly's study covers a wide variety of meanings that may result from the use of demonstratives in various contexts (1980, 35-45). Table 4 summarises them.

this	Proximal	Interest	Focusing	Foreground	Present	Speaker's responsibility in relation to referent
that	Distal	Rejection	Conclusive	Background	Past	Addressee's responsibility in relation to referent

Table 4: Possible meaning effects linked to the use of *this* and *that*

This table calls for a certain number of comments regarding the distinctions between meaning effects. It could be argued that all of these effects find their origin in the same set of cognitive operations as Lapaire (1987, 544) does in his work on articles. For Lapaire, Cotte, Adamczewski and other enunciative linguists, the TH- part of the demonstratives is the mark of a complex operational and relational anteriority by which the referent is retrieved. This corresponds to a first stage (Lapaire 1987, 526). The second stage corresponds to the -IS or -AT pseudo-morphemes—they are not strictly morphemes as they cannot be combined with a large variety of bases. Their role is to determine whether the cognitive operations performed on the referent, so far, are to be validated or not. Bouscaren provides a definition of the notion of validity by stating that a predicative relation is considered valid when anchored within some situation of utterance (Bouscaren 1991, 13). In this analysis, -IS would indicate that the cognitive operations that have brought the discourse to where it is are not closed and denote a refusal to finish with these operations. Conversely, -AT seals what -IS refuses to consider as closed. It is the mode of closure. No more cognitive operations are to be carried out on the referent (Lapaire and Rotgé 1991, 64). These two modes find their realisation in the form of the meaning effects described in Table 4 since whatever is rejected, concluded, placed in the addressee's sphere or put in the background means that no more cognitive operations on the referent are to ensue.

With his distinction between two types of reference, Danon-Boileau's analysis also results in the same interpretation of meaning effects (Danon-Boileau 1992). He posits two concepts that are grounded in the speaker-addressee relation. On the one hand, *consensual deixis* is a mode of reference where speaker and addressee are not distinguished in the act of seeing the referent. The entity belongs to the discourse common to both interlocutors. He assigns this type of deixis to *that*. On the other hand, *disconnection deixis* (discordance) is a mode of reference in which the speaker distinguishes himself/herself from the addressee. It corresponds to *this* and allows the speaker to keep control of the referent and its content (Danon-Boileau 1992, 420). With this distinction in mind, the meaning effects described in Table 4 can also be explained. For instance, in the case of focus with *this*, the speaker controls the referent and wants to force hi(s)her personal view of the referent without taking heed of the addressee's. In the case of conclusive *that*, the speaker presumes that the addressee bears some responsibility on the way the referent has been constructed up to its mention in the discourse. It reflects some kind of agreement on the final shape of the entity at discourse level.

So, what has been shown so far is how sense is given to the expressions in which *this* and *that* operate. There is a first stage in which a referent is activated in the discourse for both interlocutors and the second stage corresponds to the way the referent is going to be presented. This is the stage that ultimately affects the sense of indexical expressions as it is when meaning effects are created and proximity v. distance is only one of them.

As this chapter draws to a close, it is worth recalling that the challenge has been to provide a referential framework in which to understand how reference between a discourse referent and its corresponding text unit is established. In so doing, our analysis has alternated between semasiological and onomasiological interpretations of the issue. The objective has been to integrate several levels of interpretation of the forms depending on whether they are regarded in relation to concepts or to their meanings. We have shown that reference occurs at a cognitive level with the

construction of discourse in the act of communication. The textual clues that are produced by the speaker, *i.e. this* and *that*, are the result of cognitive operations that include the application of referential processes in a first stage, and the assignment of properties to the referent in a second stage. Figure 7 is an attempt to summarise the process that leads to the selection of *this* or *that*.

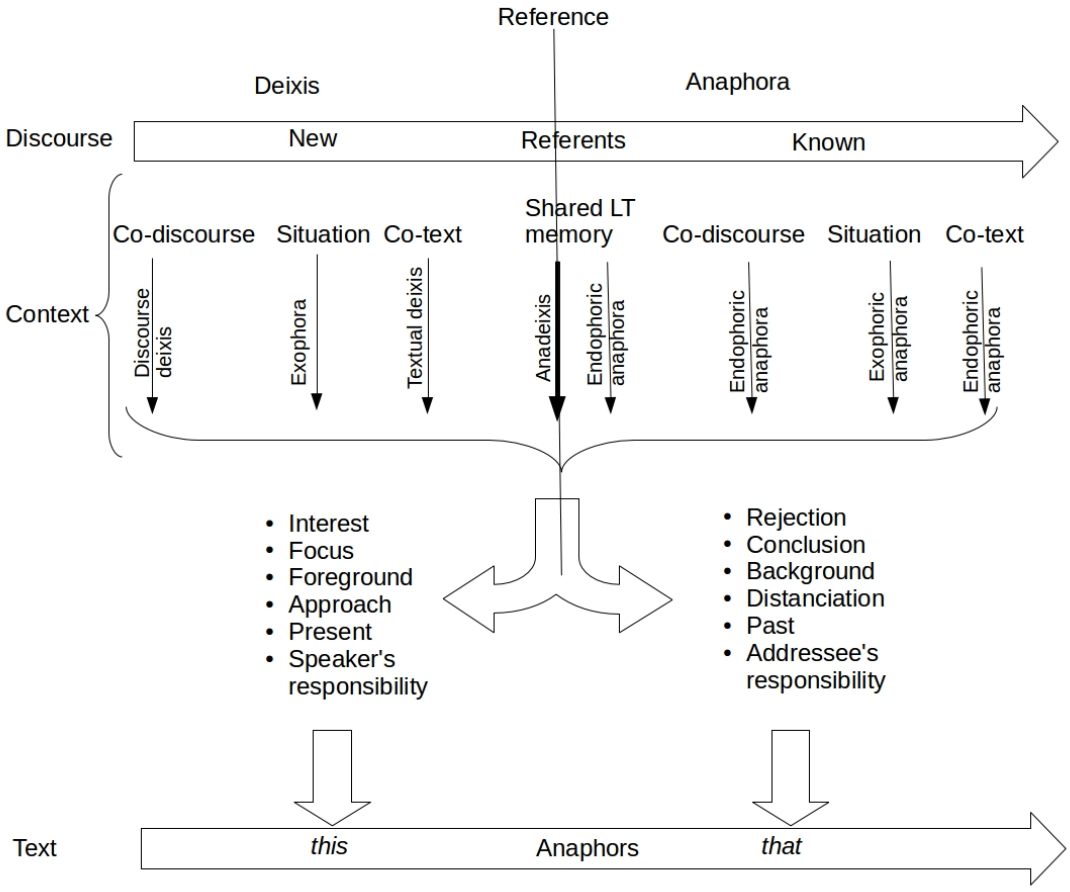

Figure 7: The selection process of *this* and *that*

To illustrate the diagram and recapitulate the various stages of reference with a demonstrative, let us finish with example 4 given on page 35. We repeat it underneath (numbered 25) for the sake of convenience.

25) "Ought we, instead of demonstrating the power of our Air Force by dropping leaflets over Germany, to have dropped bombs? (…) In *this* peaceful country, governed by public opinion, democracy and Parliament, we were not as thoroughly prepared at the outbreak as *this* dictator State

> whose whole thought was bent upon the preparation for war." (W. Churchill, War Speeches, 27/1/1940)

When Churchill made his speech in England, he first wanted to refer to his country. The first stage of the reference process consisted in retrieving the referent *England,* not so salient in the discourse at this point, since the mention of war implied several countries. The use of a deictic exophoric procedure permitted not only to support the retrieval process with the situational knowledge of his addressees but also to show a new aspect of the NP *country*, *i.e.* peaceful. The purpose of this first stage was to select the right indexical expression in order to make the referent England accessible. If, instead, the following incorrect definite construction "... In the peaceful country" had been used the entity *England* would not have been accessible. The choice of an indexical expression is dependent on how the speaker thinks the referent can be retrieved by the addressee (Kleiber 1992, 617). The anadeictic procedure is used with the positive adjective *peaceful* to place the focus on the notion of England, which is obviously present in the long-term memory of the addressees. Churchill's purpose is to make the referent accessible and the notion of England is not salient enough to just use the article *the* due to the ambiguity with the exact country in question. The choice of the demonstrative provides a higher level of accessibility and results in making the entity *peaceful England* accessible. The second stage of the referential process consists in selecting *this* or *that* with the compatible sense of the entity previously made accessible. The choice of *this* reflects the fact that Churchill places the referent within his sphere by focusing on the way his country is run democratically and peacefully.

The second reference made with the demonstrative *this* in Churchill's speech extract corresponds to another type of referential process. The first stage is to make the referent *state* accessible and its previous mention, with the proper noun *Germany,* has already made it salient in the discourse. However, the presence of another referent of the same type (*country*) in the discourse leads the speaker to use a demonstrative to maintain a high level of accessibility via endophoric anaphora. The use of *the* would not allow the stress to be made as he intended. The

choice of *this* rather than *that* comes in the second stage. *That* would have been acceptable by placing *Germany* outside Churchill's sphere. However, the war leader of the time did not want to deny the problem that Germany represented at the time and the fact that England was at war with the Nazis. Consequently, his choice of using *this* with a stress is to show his determination in facing the enemy.

2.6 Summary

The focus of this chapter has been on native English as a first step prior to better understanding incorrect uses of *this* and *that*. A state of the art on the issue of reference and the use of demonstratives has helped prepare a model for the interpretation of correct uses of the forms. We have clarified two aspects. The first one is understanding how referential procedures are applied. The second one is the identification of referential expressions such as *this* and *that* and their role in utterances. Regarding the first aspect, it appears that three levels of interpretation need to be taken into account to understand referential procedures. At text level, syntagmatic information can be retrieved to help assess the predicational context. At context level, exophora, endophora, the co-discourse or the long term memory of the speakers are the elements that help a referent integrate into discourse. At discourse level, referents are classified in relation to the new or known nature of the information they convey. They are organised hierarchically, accrue properties and remain in the long term memory of the speakers. Regarding the second aspect, referential expressions such as *this* or *that* can be said to support referential procedures. These expressions carry a degree of saliency that helps choose the type of reference and the level of accessibility of the referent.

Overall, we have surveyed the state of the art in the domain of reference and contributed to the discussion on reference by proposing a model of the referential framework that reflects the complexity of speech in terms of levels of interpretation. We have shown that different theories on reference can fit in a single framework in which referential processes are represented as arrows to show

the dynamism of referent accessibility. Such a model enables the researcher to understand referential processes in native English. It is also a solid base for the analysis of learners' implementation of referential processes in English. To interpret learner specificities in utterances, the text, context and discourse levels can be used. In the next chapter, we use this model to analyse and diagnose learner uses of *this* and *that*.

Chapter 3 Learner English and *this* and *that* in reference

In this chapter, the focus is placed on *this* and *that* in learner English. We proceed twofold. First, we survey the state of the art of the methodologies applied to learner English and, second, we use methodological tools used in SLA theories to analyse *this* and *that* in learner English. We start with a theoretical viewpoint before focusing on a practical application. The approach may seem odd but the purpose is to show how existing SLA methods can be used in a practical qualitative study of *this* and *that* to determine the linguistic criteria used in a subsequent large-scale quantitative study.

We contextualise our analysis within the emerging community of learner corpus research—in 2011, the founding conference celebrated the 20th anniversary of learner corpora in Louvain-la-Neuve. The twofold process aims to review methodologies before carrying out a qualitative analysis to identify hypotheses concerning the way learners use these forms. To unmask the characteristics of learner use of *this* and *that,* we analyse specific occurrences in light of the model developed on native English. The purpose is to provide a typology of learner errors regarding *this* and *that* and to identify the features which are specific to erroneous configurations. In order to prepare the elaboration of a typology, a methodology adapted to the study of learner English needs to be identified. Therefore, a survey of the state of the art of Second Language Acquisition research methodologies needs to be carried out showing how learner corpora inform SLA research questions (Lüdeling and Hirschmann 2015). This work leads us to the development of a hybrid methodology based on error and form/function analysis which helps

establish the typology and isolate the core linguistic issue of this thesis which is the existence of a learner-specific microsystem in the use of the forms as pro-forms.

Section 3.1 focuses on the systematic variability of interlanguage and specific analytical methods. It is dedicated to a state of the art in the domain of learner English. In Section 3.2, we look at the way learners use the demonstratives. The uses of *this* and *that* are studied to deduce patterns of systematic variability linked with specific contextual factors. Based on a set of occurrences, we create a typology which relies on criteria—identified in the native model presented in Chapter 2—such as the context and the functional realisations of the forms. The existence of a learner-specific microsystem linked to the pro-form functional realisation of the forms is exploited. In Section 3.3, we use the typology of errors to identify a research question on the microsystem and to determine the elements which need to be taken into account for a large scale analysis of the pro-form microsystem.

3.1 Description of learner language

The purpose of this section is to understand what makes learner language different from native language and, thus, a particular object of study. Interlanguage, the notion coined by Selinker (1972), appears to be the relevant theoretical framework to understand these differences and the various analytical methodologies which have stemmed from it. This section is composed of two parts. In Section 3.1.1, we present the notion of interlanguage and its variability as a system. In Section 3.1.2, methodologies for the analysis of learner language are reviewed.

3.1.1 Interlanguage

Interlanguage is a system which is specific to learner language. Selinker provides the definition:

> "[...] This set of utterances for *most* learners of a second language is not identical to the hypothesized corresponding set of utterances which would have been produced by a native speaker of the TL (target language) had he attempted to express the same meaning as the learner. Since we can observe that these two sets of utterances are not identical, then in the making of constructs relevant to a theory of second-language learning, one would be completely justified in hypothesizing, perhaps even *compelled* to

> hypothesize, the existence of a separate language system based on the observable output which results from a learner's attempted production of a TL norm. This linguistic system we will call '*interlanguage*' (IL)." (Selinker 1972, 214)

Selinker raises the issue of differences in the way a learner formulates specific ideas compared with the way a native speaker would formulate the same ideas. The differences are the evidence that pushes Selinker to conclude on the now broadly-accepted existence of a learner-specific linguistic system. It follows that such a system can be described thanks to the observable output produced by learners, *i.e.* the utterances, their syntax and their lexicon.

When learners produce utterances which are different from natives, their utterances may include errors. Errors can be seen as deviations from a norm. When it comes to observing learner language, the analyst is faced with a variety of deviations from the target language whose characteristic is that of being the language used by natives[5]. On the surface, the deviations may take the form of errors that can be located on specific segments. At the time of utterance, learners make lexical, syntactic and phonetic choices that appear as inconsistent with the target language. In that respect, they correspond to what Corder names a "deviation from norms of the target language" (Corder, 1967 cited in Ellis 1994, 51). Hence, errors are detected at the surface structure of the output. The notion of 'error' must not be mistaken with that of 'mistake' as the latter "happens when the learner fails to perform competence" when errors "are the result of a lack of knowledge" (Ellis 1994, 51). All in all, errors are specific realisations chosen by learners as solutions for specific language needs. They are part of a variety of signs at the surface structure that are traces of the underlying learner-specific system. This concept has laid the ground for studies such as Dulay and Burt's on the syntactic errors of children in their learning of a second language (1974).

[5] Although, the term native is subject to discussions (see Brutt-Griffler and Samimy 2001 for a discussion on identity construction), the debate will not be opened here and native language will be considered as that spoken by the people born and living in a country whose official language is the target language of the learner.

Vogel introduces the idea of the inherent instability of interlanguage in an individual as the source for errors:

> "The instability of interlanguage is shown via a set of specific features such as a high proportion of errors and disturbance phenomena. They result in many hesitations, in a simplification of structures, in a reduction in semantic complexity, in the use of compensation or avoidance strategies." (Vogel 1995, 69, my translation[6])

The instability of interlanguage implies that learner devise strategies to seek and find answers within their linguistic knowledge. He adds that in general these strategies are put in place when learners' linguistic tools become limited and cannot satisfy their needs in the situation (Vogel 1995, 70). Consequently, it seems that instability is the source of variations within learners' utterances. Just like native speakers, learners also vary in their use of language forms. One learner may use a particular speech form correctly in one context and not in another one. However, this only reflects one type of variation. In fact, in the course of their productions, learners' choices may vary along with the changes that occur in time, situation, and individuals (Ellis 1994, 134). For Vogel, interlanguage evolves over time and is transformed gradually to reach the norm set by the target language (Vogel 1995, 66). The factors that explain variations are found across the domains of psycholinguistics and sociolinguistics:

> "Sociolinguistic approaches treat social factors as primary, although as we will see, they may also refer to psycholinguistic mechanisms to explain how situational factors result in variability. Psycholinguistic approaches are concerned with identifying the internal mechanisms responsible for variable performance and pay little attention to the social factors that motivate them. Sociolinguistic and psycholinguistic approaches, therefore, are complementary and a full account of variability in learner language requires both." (Ellis 1994, 120)

There are several levels of variability. For Ellis (1994, 134), the first level is related to time and includes the distinction between synchronic and diachronic variation. The second level relies on the notion of learner and the distinction between intra-learner and inter-learner variations. Ellis (1994, 472) gives an inter-learner account

[6] Original version: "L'instabilité de l'interlangue se manifeste dans une série de traits spécifiques tels que la proportion élevée d'erreurs et de phénomènes de perturbation, qui se concrétisent par de nombreuses hésitations, la simplification des structures, la réduction de la complexité sémantique, l'utilisation de stratégies de compensation ou d'évitement." (Vogel 1995, 69)

of three surveys based on social factors considered as influencing individual learner differences in language learning. Conversely, intra-learner variation describes the use of specific language features observed in the speech of any individual at any given time. The third level is systematic variability. It is based on language features and whether their unstable character may be systematic or not, in which case it is called free variation. Systematic variability depends on several factors: "Systematic variability arises as a result of external factors to do with the linguistic context, the situation context, and the psycholinguistic context." (Ellis 1994, 135). Each of these contexts influences the forms found in learner language and the specific language functions they were performed for, hence the idea of form-function variation. To sum up, there are three levels of variability for interlanguage: time, learner and systematicity.

Systematic inter-subject variability appears to be the most relevant level for the identification of psycholinguistic factors in the use of *this* and *that* by learners. Ellis places the study of forms in relation to their functions at this level, due to the influential role of context for the choice of forms employed to perform specific functions (Ellis 1994, 135). Since learners may choose different form realisations for the same need, a study based on form-function correlations may help understand the contexts in which form-function variations occur. For instance, (Young 1996, 172) studies articles used by Czeck and Slovak learners of English of different proficiency levels. The paper concludes that there is an important effect of form-function relations in developing interlanguage. Learners map L1 meaning onto L2 forms and there is evidence of systematicity in indefinite articles as they are mapped to count and number categories.

To shed light on why form-function variation occurs, it may be relevant to delve into the notion of sign and its arbitrary property first pointed out by Saussure and later taken up by Frei (2011 [1929]). In an analysis of letters from French prisoners to their families during the First World War, Frei covered this notion to provide an explanation for the reason why errors were committed in the writings. For the

linguist, errors had a function in their being made as they reflect how language works. In doing so, he laid the foundations for what is today called Functional Linguistics (FL). He is the founding father of the 'grammar of errors' and can be regarded as a usage-based linguist more than a functional linguist à la Halliday. For Frei, language is governed by a series of basic needs that act as guiding principles in the act of utterance. Two of these needs, *i.e.* economy, and invariability play a role in the question of variability. Frei shows that the need for economy is the reason why language is composed of mobile signs that can convey a large number of meanings. In addition, these signs are as invariable as possible in order to avoid complexity and memory overload. In short, the reason why the sign is invariable and mobile is to avoid memory effort (Frei 2011 [1929], 171). Bearing in mind this functional approach of language and the mobile property of signs, it is possible to elaborate an explanation for the systematic variability found in learner uses of forms with specific functions. The sign is mobile and this mobility aims at simplifying the description of the world. One sign is used in many contexts and situations in a target language. The learner is aware of this mobility as the same principle applies to the L1. However, s/he is not aware of all the specific contexts and syntactic configurations that trigger the use of the sign, hence variations in selecting forms in contexts. Ellis specifies the types of contexts that influence the sign, or the form in his terminology, in its functional use:

> "Form-function variation is also systematic and is contextually induced. All three types of context—linguistic, situational, and psycholinguistic—may influence which forms learners employ to perform specific language functions." (Ellis 1994, 135)

In the act of production, the discourse that unfolds generates a multiplicity of contexts calling for functions and thus possible forms. If these forms or signs are not recognised as such, then the learner will mismatch form and context. Therefore, the learner has to learn the contexts and situations in which to apply the sign. In the case of an unknown configuration or situation, the learner might resort to other signs than those that are expected to fit the configuration better. Consequently, due to its mobility, the sign is used with variability. As contexts and configurations evolve and become complex, the sign and its mobility are used to

simplify the description of the complexity. So configurations change but the use of the same sign allows an economy in terms of working memory (see Chanquoy, Tricot, and Sweller 2007 for details on the theory of cognitive workload and its applications). Frei explains how one word form is found in many places.

> "The need for economy forces language to replace the multiplicity of particular signs by mobile signs that can translate a great number of distinct meanings in their own turn. This generalisation does not only include generic signs (*i.e.* tree, house, vehicle, parcel, etc.). It also includes signs for categories that determine the concepts of thing, being, quality, action, *etc.*" (Frei 2011 [1929], 165, my translation)

Conversely, the speaker needs to establish a link between the new configuration and the sign. If this is not achieved then the sign is not used. If there is a confusion of configuration the sign is used wrongly leading to an error. If the configuration is correctly identified the sign is correctly used. Because discourse is complex, the learner finds it hard to match all discourse configurations with signs, hence the correct and incorrect uses for the same sign.

So interlanguage is the hypothesis that supposedly shapes the difference between learner and target languages. Interlanguage is an unstable system that shows evidence of variability in learner output. Exploring why learners vary their realisations to implement the same functional needs is a way to better understand the notion of systematic variability that affects interlanguage. To do so, a specific methodology needs to be applied. The next section is a review of different methodologies used in the analysis of learner language.

3.1.2 Methodologies to analyse learner language

In this section, several methodologies for the study of Second Language Acquisition (SLA) will be explored. The purpose is to identify their principles since they may be retrieved and followed in experiments described in subsequent chapters.

3.1.2.1 Previous approaches

A number of approaches have been developed in the last decades. In the following section, we give an account of their key concepts.

3.1.2.1.1 Contrastive Analysis

Contrastive Analysis (CA hereafter) is an outdated method for analysis that had its heyday in the 1960s, at a time when behaviourism was mainstream in the study of human sciences. It was believed that if difficulties were to arise among learners of an L2, they would be due to the differences that existed between the learner's L1 and the L2 (van Els *et al.* 1984, 38). This meant that language learning was viewed as a process of transfer or interferences between the L1 and the L2. For cases in which both languages were similar or identical in their structures, positive transfers were to be expected. For cases in which both languages differed, interferences would be expected and negative transfers would be observed. In other terms, learners would apply L1 structures in L2 output.

Hence, researchers set out to identify similarities, to group differences of various kinds and to classify them according to degrees of difficulty. The purpose was to extract lists of errors that learners were bound to commit when learning their L2 and to use them for teaching purposes. CA was intended to predict errors but the problem is that it did not fulfil its promise as evidence was given that there were cases in which learners were not influenced by their L1 when making mistakes. Instead, they were the results of creative processes. A weaker form of the CA approach was later used to explain errors instead of predicting them (Díaz-Negrillo 2007, 28). It also accepted that transfer-related errors corresponded only to a partial view of interlanguage.

Nevertheless, the approach served its purpose in the longer run as it yielded research efforts towards Error Analysis (EA hereafter) and Computer-aided Error Analysis (CEA) (Dagneaux, Denness, and Granger 1998) while having established

the principle of language transfer in L2 acquisition. It also oriented SLA research towards an interlanguage comparative approach—Contrastive Interlanguage Analysis (CIA)(see Section 3.1.2.2.2 for more details)—that is now considered as an essential comparative principle in the domain. For Granger “the interest of L1–L2 comparisons is even more obvious as they help teachers identify the lexical, grammatical and discourse features that differentiate learners’ production from the targeted norm and may therefore be usefully integrated into the teaching programme” (Granger 2008, 341).

3.1.2.1.2 Error Analysis

In the history of SLA, errors have been one of the elements preeminently used by learner language analysts to characterise and explain the progression path followed by subjects. Early on, Corder described a procedure to identify errors in which utterances were tested with questions on the interpretation that could be made of utterances (Corder 71a cited in Ellis 1994, 53-56). However, this type of procedure was not without caveats. It relied on a certain number of assumptions that raised questions of clarity and reliability as learners were used as informants in the interpretation process of the errors (Ellis 1994, 54).

The process of error identification was also tackled via a descriptive approach of the elements to be found on the surface of learners' utterances. By categorising errors, researchers expected to lay the ground for further explanation of errors. Dulay, Burt and Krashen (1982, 150) are cited by Ellis for their insistence on the “need for descriptive taxonomies of errors that focus only on observable, surface features of errors as a basis for subsequent explanation” (Ellis 1994, 54). In line with this principle of error taxonomy, they established a simple system of error descriptions supposed to encompass all possible surface structure discrepancies with a target language. Their taxonomy included information expected to indicate “the cognitive processes that underlie the learner's reconstruction of an L2” (Ellis 1994, 56). Omissions, additions, misinformations, misorderings were some of the

categories suggested to classify the various types of errors (Dulay, Burt and Krashen, 1982 cited in Ellis 1994, 56). For these researchers, their taxonomy provided information on the strategy operated by learners on the surface structure. However, Ellis challenged this view in so far as it presupposes that learners simply operate on the surface structure and that they do not create or invent their own structures, which is a characteristic of interlanguage as defined above. For Ellis, a taxonomy grounded on surface structure description must represent the mental processes that lead to the creation of errors. In addition, Ellis points out other criticisms for Error Analysis. The focus on errors is a primary bias (Ellis 1994, 68) since the fact of focusing on errors only shows a partial view of what learners actually produce. It is evident that learners' productions include errors as well as other aspects such as overuse and underuse of specific patterns or forms (Granger 2008, 342). This shows the need to rely on a broader framework in which errors are considered as one of the components of learners' productions. Interlanguage, in these terms, appears as a suitable framework as the notion implies the fact that all utterances—correct and incorrect—are considered when studying learner language. Interlanguage supports a better understanding of the rules that the learner applies in the act of utterance. Conversely, EA does not focus on all the types of differences between target language and learner language. For instance, learners may use avoidance strategies which allow them to express their ideas without committing any error. This type of behaviour is not detected when carrying out EA-based surveys. In this respect, the contribution of learner corpora brings a more balanced insight into interlanguage as they provide a more global view of the phenomenon (see Granger, Gilquin, and Meunier 2015) and support the possible analysis of errors (Díaz-Negrillo and Fernandez-Domingez 2006).

Another point of criticism raised by Ellis is the fact that EA studies are carried out over a population at one point in time and thus give “a static view of L2 acquisition” (Ellis 1994, 68). This appears to be a hindrance for the analysis of the gradual process of acquisition since the errors cannot be related to a particular stage in the learning process. Corpus-based longitudinal studies, however, appear

as a solution to overcome this limitation as particular instances of learner language are connected to stages in time. Longitudinal or quasi-longitudinal studies (Granger 2004, 131) allow for comparisons between stages and, thus, the discovery of common features.

3.1.2.1.3 Frequency analysis

Another approach stems from the way errors can be characterised. Errors can be interpreted in terms of specific misused word forms but also in terms of overused and underused patterns. This second aspect helps researchers deduce behaviours that do not comply with native use for instance. In this respect, frequency analysis is a useful methodology to measure whether learners adopt idiosyncratic uses of existing forms among natives. As Granger (1994, 27) puts it:

> "Learner language is not merely distinguished by its errors, however […] The foreign-soundingness of advanced learners' writing is often due to over- and underuse of certain words and phrases. Such phenomena are not captured by error analysis but are clearly brought out when the data is submitted to text retrieval software [...]".

What she meant at the time was that research on learner corpora could benefit from computerised corpora and, thus, could allow new measurements such as the frequency of use of words. Inspired by CA in its language comparative approach, frequency analysis was going to become a major line of work in the domain allowing for the profiling of language users in relation to the paradigm of nativeness.

> "One of the things that the computer can most easily produce is frequency lists. The raw frequencies of words and phrases used by learners when compared with native speaker frequencies, often reveal very interesting differences" (Granger 1994, 27).

This line of research pushed comparisons between learners of different language backgrounds to new horizons since high volumes of data could then be processed. Results in the form of normalised frequency allow for comparisons according to categories such as the learners' L1s, text genres, discourse types and other factors that bear some influence on language production.

The methodology is grounded on the principle of observing how often specific language patterns are used. It starts with the identification of specific forms in a corpus, followed by a word count. The word count undergoes a normalisation procedure which involves dividing the raw frequency count of a form of specific categories by the total number of words in a given corpus. The ratio is then multiplied by normalised values such as 100,000 or a million. With normalised results, it is possible to compare values within and between corpora. For instance, Biber *et al.*'s results on the frequency counts of demonstratives in conversation are presented per million words (Biber *et al.* 1999, 349). The authors show the different uses of *this* and *that* in native conversation depending on their determiner or pro-form functions in the corpus. In addition, these results are comparable with other results obtained from other corpora and normalised in the same way. However, the question of normalisation must be cautiously considered regarding the use of corpora of small sizes or learner language. Counts in small size corpora might lead to over-generalisations and variability in learner language may lead to uncounted patterns (Cortes 2007, 105–106).

When analysing developmental patterns for instance, the first step is to identify the different devices used to perform a specific language point under observation. Next, the classification of the patterns in relation to their occurrence in time leads to the elicitation of a learning sequence for the particular point. It can be applied on longitudinal data and thus provides information on variations in time of learner output. Ellis was successful in showing evidence of the use of variants for particular language variables. It was shown, for instance, that an existential sentence was realised by three variants: no verb (*There church*), contracted verb (*There's church*) and full verb (*There is a church*). These variants include errors and the approach based on frequency established a link between the uses of specific variants by learners and linguistic contexts and specific learner profiles (Ellis and Barkhuizen 2005, 94). In other words, snapshots of interlanguage can be taken for each microsystem, hence the possibility to examine learners' use of linguistic forms in their own right, and “not in terms of whether they correspond to target language

forms" (Ellis and Barkhuizen 2005, 94). As it must be applied longitudinally, this approach requires sound and solid longitudinal data which are not so easy to collect. Ellis points out an issue related to the decision of matching a variant to a particular stage in a learning sequence. The question is to know when to determine the threshold from which a particular variant can be considered acquired and thus matched to a given stage. There is variability in the definition of stages, which casts a shadow of artificiality on the act "of defining a sequence of acquisition in terms of a set of stages" (Ellis and Barkhuizen 2005, 98). Nevertheless, the sole notion of frequency computation *i.e.* the 'underuse'/'overuse' paradigm has established a strong principle in the study of learner data.

3.1.2.1.4 Functional analysis

If language is considered primarily as a tool to satisfy communication needs, its study requires the identification of language forms with semantic, semantico-grammatical pragmatic and discourse functions (Huebner 1985, cited in Ellis and Barkhuizen 2005, 113). By establishing mappings between functions and forms, it is then possible to provide a layout of the language system where grammatical forms are used to convey meaning.

Ellis documents two types of functional analysis. The first one, called form-function analysis, starts from the form itself and the researcher is compelled to search the data to find all the various functions that are performed. Frequency counts can then be done and functions can be matched to particular profiles of learners. This process can be applied cross-sectionally and longitudinally and provides answers on how learners of different profiles use the forms, which in turn casts light on the variable aspects of learners' interlanguage. The second type of functional analysis, called function-form analysis, starts from the function. Its purpose is to explore the communicative need of the language by examining how a particular function is performed in terms of number of forms. It relies on the idea that "the need to perform a particular function motivates a learner to attend to a particular form"

(Ellis and Barkhuizen 2005, 126). After the selection of one function, the researcher then identifies all the forms that are used to perform it and frequency counts provide information on what forms are dominant in the fulfilment of the function. The set of frequencies gives a classification of forms that can be used as evidence for the acquisition stages and their links with learners' profiles. Tarone and Parrish (1988) report on an experiment on article usage by Japanese and Arabic subjects. The approach relies on frequency counts of the article + NP relation considered as a function. Results show that depending on the type of NP, learners' accuracy varies with the type of task in which the learners used the function. Overall, the functional approach grounds itself in the measurement of what Ellis names as form-function mappings. This principle provides a strong basis for the quantitative analysis of learner language taken in context.

3.1.2.2 Current approach – the Learner Corpus Research framework

All the methods described in the previous sections followed one another over the last decades but it does not necessarily mean that they subsided. Instead, some of their principles have subsisted and have been gradually blended to produce methodological principles applicable to learner corpora. So, as Díaz-Negrillo puts it: “it is also true that current views are founded on the realization of the limitations of previous approaches” (2007, 26). It could be added that current views are also founded on some of the principles set in previous approaches. Today's learner corpus research tends to benefit from these decades as it builds on this experience. LCR applies these methodological principles and increasingly relies on statistical methods to process large amounts of data and to uncover trends in language use (Gries 2013 [2009], 4).

As explained in previous sections, analytical methods used to study learner language showed limitations. For instance, EA suffered from shortcomings in methodology and scope, as it narrowed down learner differences with natives to the sole idea of errors. CA, for its part, based errors solely on the notion of cross-

linguistic interference and, in actual fact, the theory's claim to predict errors was proved untrue. The weaker version of CA, based on *a posteriori* explanation of errors, was disqualified on the grounds that it defeated the purpose of the idea of prediction.

The criticisms that were raised may have found answers with the advent of Learner Corpus Research (LCR hereafter) and Computer-aided Error Analysis as one of its corollaries (see, among others, Dagneaux, Denness, and Granger 1998; Granger 2002). With the growing power of computers, the field of corpus linguistics developed an approach grounded in the compilation and the analysis of native language corpora. It was not long until learner corpora were also subjected to computer processing. The objective in this case was not only to work on errors, but to include the approach in the broader framework of interlanguage (Granger 1994). In this respect, other indicators such as 'overuse' and 'underuse' have made it possible to compare different patterns of learner language with native language. LCR has also shown to be a fruitful framework in the exploration of interlanguage by linking learner corpus exploration results with the language levels of the Common European Framework of Reference (CEFR). Hawkins and Buttery (2010) give evidence of criterial features of learner English. They show that various linguistic patterns may be correlated to specific levels of the CEFR. LCR also offers a suitable framework to deal with the problem of variability in learners' interlanguage. It allows the definition of longitudinal corpora that are likely to be helpful for the investigation of the evolution of learners' productions over time. Patterns, functions, errors and other aspects can be traced back to their initial appearances and their evolutions can be monitored. All in all, LCR opens the door for a comprehensive exploration of interlanguage in which both negative and positive aspects are taken into account. LCR relies on two main components: Error annotation and Contrastive Interlanguage Analysis.

3.1.2.2.1 Error annotation

It was argued that the compilation of large corpora of learners' language would provide the basis for the analysis of SLA. LCR relies on many components such as errors but not only. Barlow supports this view and clearly links the design of learner corpora to the procedural elements that compose EA.

> "One impetus to compile learner corpora follows from the Error Analysis tradition of identifying, describing, and explaining errors, and many of the issues related to error analysis and linguistic analyses based on categorization of errors […]." (Barlow 2005, 335)

The categorisation of errors found a concrete realisation in the act of recording learner language with methodological guidelines for the compilation of a corpus (Wynne 2005), which included the requirement of annotation. (Further details will be given on this aspect in Chapter 4). It is within this particular aspect of corpus compilation that errors have received a new interest. This time, though, their inclusion within methodologically-compiled corpora helps overcome the limitations mentioned above. They are the subject of consistent precise guidelines and are part of broader contexts that include other less visible features of learner language such as avoidance strategies, variations or frequencies of use of idioms or patterns. For Granger, the advent of the computerised learner corpus was the opportunity to bring back EA, but this time on the basis of large corpora: "the researcher will now have access to large databases, which can be submitted to automatic or semi-automatic analysis (Granger 1994, 26)."

In this context of computer-based analysis, the issue of error annotation is one of the first to be discussed in the domain. The definition of several types of annotation has accompanied the development of learner corpora but the question of error identification has been central ever since the early days of LCR. As opposed to EA, which relied mostly on a manually collected and a limited number of occurrences, the computer-based realisation of the task has entailed the collection of large samples of data that guarantee that errors are actually considered within the framework of interlanguage. However, the identification process of errors has not

been without discussions since there are a number of issues that appear when collecting information on errors (see Chapter 4 for the distinctions between annotation schemes of learner corpora). One of these issues is the way errors are tagged in a given corpus. For instance, (Lüdeling *et al.* 2005; Lüdeling and Hirschmann 2015) support the view that a multi-level annotation system solves tagging problems that can occur in a flat annotation system—a system in which error annotation is appended to the raw texts of a corpus.

The International Corpus of Learner English (ICLE hereafter) project (Granger 1993) was the first of its kind to include error annotation. In that respect, the Louvain-based team of researchers at the Centre for English Corpus Linguistics of UCL developed an annotation system of errors in order to label all the learner errors found in the corpus (see Section 4.1.1.2 for a detailed account of the annotation scheme). The purpose of such an endeavour was to be able to automatically retrieve errors belonging to the same category and, thus, "draw up comprehensive catalogues of errors" (Granger 1994, 27). The error-annotation system, however, relied on the manual annotation of errors, which irremediably raises the issue of error identification again, as well as the granularity of the analysis (Díaz-Negrillo 2007). Once the set of tags has been described theoretically, its application on a learner corpus is not without any difficulties. Error interpretations may vary between annotators, leading to low inter-annotator agreement. So this phenomenon has been under scrutiny and discussions in recent years have been drawn towards the necessity to distinguish errors according to theory-neutral labels.

Lüdeling and Hirschmann (2015, Chapter 7 Section 2.4) clearly point out the problem and show that such an endeavour is not straightforward. By contrasting ungrammatical examples, they show that, in spite of grammar rules, the interpretation of errors leaves room for subjectivity. They recall that the nativeness paradigm helps determine the construction which is actually targeted by learners, *i.e.* the target hypothesis. They show that there may be several target hypotheses

for one error and, sometimes, errors cannot even yield any interpretation as no target hypothesis is identifiable. Their point is that errors lead to different interpretations and so they advocate for the need to state the target hypotheses used to interpret an error. They also note that errors could be assigned to one or several categories, hence their support for a multi-level annotation system of learner language.

Díaz-Negrillo *et al.* (2010) have advocated for an annotation system that would rely on Part-of-Speech (PoS hereafter) tagging. PoS annotation, in their view, should systematically encode the distributional, morphological, and lexical aspects which are specific to learner language. Meanwhile, they also recognise the limits of the single PoS annotation used for native corpora: "Native PoS annotation for learner language is problematic for several classes of cases in which the evidence from distribution, lexis, and morphology systematically does not converge on a single PoS classification." (Díaz-Negrillo *et al.* 2010, 12). Consequently, their view is that of a multi-layer PoS annotation scheme that would encode each distributional, morphological, and lexical aspect on a different layer. This type of encoding is not error annotation but it could be said that it supports error detection.

All in all, error annotation remains an evolving line of research in which the question of error tagsets is central (see Section 4.1.1.2), hence our contribution in the field in our PhD.

3.1.2.2.2 Contrastive Interlanguage Analysis

Learner corpus research also stemmed from Contrastive Analysis and frequency analysis insofar as the idea of processing a mass of data to count occurrences of forms and to compare results between learners of different L1s and also with natives has been an important line of work. It has given results in terms of under-use and overuse of linguistic forms. Contrastive Interlanguage Analysis (Granger 1996; Granger 1998, 12) lays the foundations for this kind of analysis that sees

native language as a comparison standard for the analysis of learner corpora. As it operates on corpora, it seems appropriate to include it under learner corpus analysis. For Barlow, studies based on this paradigm provide “evidence for the nature of interlanguage, focusing on the non-native aspects of learners' speech or writing” (Barlow 2005, 343). Word counts compose the basis of studies in this line of research. However, some studies have gone further in the refinement of the data as annotation is also taken into account in the frequency measures. Annotation tags linked to words prevent semantic ambiguity inherent to some word forms. CIA can also support form-function analysis insofar as form counts can be made for specific functions related to a form. As a result, information on language use in the form of patterns emerges thanks to comparisons made diachronically and synchronically between speakers. Ellis defines patterns as “linkages at the conceptual level” (Ellis and Barkhuizen 2005, 269), that is to say constructs composed of the codings of word forms whose repetitions yield similarities. These constructs can be identified and grouped according to various conceptual categories such as time, user profiles, functions and so on. Analysis based on statistically validated data can yield results in several areas by highlighting factors that are significant in the use of forms by learners. Barlow (2005, 343) gives a number of factors that help explain characteristics of interlanguage and some can be interpreted as evidence of the transitional nature of learner language, namely L1 transfers and paths of interlanguage development.

L1 transfers, also called crosslinguistic influence, are evidence of specific learning processes. Odlin (1982, 27 cited in Ellis 1994, 301) provides a definition of the concept: “Transfer is the influence resulting from the similarities and differences between the target language and any other language that has been previously (and perhaps imperfectly) acquired”. By comparing output from at least two groups of learners or non-native speakers (NNS hereafter) of different L1s and output from native speakers (NS hereafter), it is possible to detect differences in the use of particular forms or grammatical patterns according to the NS's or the NNS's L1s. Close observation might reveal that a particular form is used normally by NNS of a

L1 and not normally by NNS of another L1. The observed difference in use may suggest that, for the second group of NNS, there are interferences with their L1. These interferences are repeated via specific patterns. Granger places this approach under CIA:

> "CIA also involves NNS/NNS comparisons. By comparing different learner populations, researchers improve their knowledge of interlanguage. In particular, comparisons of learner data from different mother tongue backgrounds help researchers to differentiate between features which are shared by several learner populations and are therefore more likely to be developmental and those which are peculiar to one national group and therefore possibly L1-dependent" (Granger, Hung, and Petch-Tyson 2002, 11).

In this view, the occurrence of features can be traced to transfer issues in case these only appear within one group of learners of a specific L1. Transfers can be of two different kinds. Positive, they show evidence that the learners' L1 patterns apply correctly to the L2 and, thus, facilitate acquisition. Negative, their occurrence proves that some L1 patterns interfere with the acquisition of L2 patterns since they are different. Overall, transfers give evidence of the influence of the learners' L1 on the L2 and help understand their importance within interlanguage.

The use of CIA with functional analysis (see Section 3.1.2.1.4) and frequency analysis (see Section 3.1.2.1.3) also helps study transitional processes in SLA with the purpose of making developmental patterns emerge. The mutual comparison of frequencies in relation to the source of the data and the learner profiles which are attached, may yield results that show stages of acquisition. The patterns correspond to systematic paths of interlanguage development. For Ellis, the definition of developmental patterns is grounded in the distinction between two types of question.

The first type refers to the order of acquisition of language features in relation to one another. In other terms, "do learners acquire some target-language (TL) features before others?" (Ellis 1994, 73). This question is based on the assumption that interlanguage is an evolutive system composed of an increasing number of features. The order in which these are introduced in the system is thus of interest

for the researcher. This kind of evidence can be the result of a function-form approach (see Section 3.1.2.1.4) in the study of learner language. Classifying the linguistic forms used to achieve particular linguistic functions gives evidence of preferred forms at different stages of the learning process. It makes it possible to establish a roadmap of syntactic, semantic, phonological and morphological features in order of acquisition. One current line of research is reported in (Hawkins and Buttery 2010). The paper suggests four types of criterial features whose values help determine proficiency CEFR levels of learners. They identify positive and negative language properties linked to error or usage frequency levels. Frequency counts are used measure the distance between learner and native use and help classify language properties from A1 to C2 levels. Earlier in the seventies, studies carried out by Dulay and Burt (1975) and Krashen (1977) (cited by Ellis 1994, 94) demonstrated the existence of an order of acquisition of morphemes. For instance, it was shown that the *-ing* form, plural and copula were the group of morphemes acquired first before auxiliaries and articles. The classification was accomplished by calculating the accuracy with which the morphemes were actually used in relation to the contexts in which they were expected to occur. The assumption was that the more accurate a morpheme use, the earlier it was considered to have been acquired by the learner. The accuracy order was thus interpreted as an acquisition order. Such studies showed that interlanguage development followed stages and it was possible to describe them by focusing on specific language needs and on the grammatical items that are linked to them.

The second type of question is related to developmental patterns and the notion of sequence. In this case, the approach is still based on frequency and functional methodologies but the question that is asked relates to a specific linguistic item and its path of acquisition. In other words, the question is: “How do learners acquire a particular TL linguistic feature?” (Ellis 1994, 73). A series of studies shows that learners do not follow a straightforward path when acquiring particular language features. For instance, the question of the assimilation of pronoun use appeared in several studies and results showed evidence of a gradual acquisition depending on

several factors such as the subject or object position of the pronoun (Gundel and Tarone 1983 cited in Ellis 1994, 97) or its semantic classes such as person and number (Felix and Hahn 1985 cited in Ellis 1994, 96; Lightbown and Spada 1990 cited in Ellis 1994, 97).

CIA and the frequency approach in learner corpus research have been complemented with statistical methods applied to corpus linguistics. Frequency results provide distributional information that needs to be described and tested to verify its validity (Gries 2010, 5; Gard 2008). Furthermore, it also establishes a framework for predictive analysis of the forms that are observed in a particular sample. Statistical methods are also used in learning algorithms used in Natural Language Processing tools to process corpora (Manning and Schütze 1999). Automatic PoS taggers are one example of these and their use relies on algorithms that predict tags on the basis of statistics elaborated on training data. A number of studies have been conducted on the issue of tagging learner data (Nøklestad and Søfteland 2007; de Haan 2000; van Rooy and Schafer 2003). PoS taggers are usually implemented in the coding process of the data rather than for the analysis of the data. Their reliability on learner data has proved to be weaker than on native corpora but they allow for faster annotation processes. Manual correction is usually implemented to ensure stronger support for linguistic analysis. The work presented in subsequent chapters will document how such algorithms operate to assist with the processing of learner data.

Overall, the fact that several approaches are combined in the process of analysis of learner corpora provides the basis for research on the reasons why differences between natives and learners are observed, helping the researcher to understand better how the learners construct their utterances. The combination of methodologies provides information on dominant or non-dominant forms used by learners which, put together, may underline developmental patterns. Cross-linguistic influences can also be deduced from the analysis of the data. The work in this thesis is grounded in LCR supported by Natural Language techniques. These

techniques are used to process learner corpora at different stages, which includes data annotation, data formatting and ultimately data analysis. Subsequent chapters document how a form-function and frequency analysis across several corpora, with the use of NLP techniques, helps compare speakers of different L1s in order to analyse how *this* and *that* operate at interlanguage level.

3.2 A qualitative study of *this* and *that* in a learner language environment

In this section, after a quick overview of previous studies on *this* and *that* in learner English, we conduct a qualitative survey of learner data in order to provide a typology of errors, which leads to our research question on pro-forms. In the last sub-section, we detail the necessary criteria for modelling the pro-form microsystem.

3.2.1 Previous studies

Chapter 2 of the thesis covers how *this* and *that* are used in native English and more generally the issue of deixis and anaphora; the focus is now placed on the issue of learner English, including unexpected uses. As Lenko-Szymanska (2004, 90) puts it "students have problems in using demonstratives". Her study focuses on Polish learners of English and shows consistent differences in the patterns of use of the two forms. Her results show two areas of difficulty in relation to the frequency of occurrence and, using Biber *et al.*'s terminology, between proximal *(this* and *these)* and distal *(that* and *those)* demonstratives. Patterns of overuse seem to emerge when analysing the use of demonstratives as a whole. Compared with native forms *that* appears to be overused by learners.

Such studies corroborate teachers' intuitions that referential processes constructed with the help of demonstratives are not always correct. Some empirical research has been carried out on the two forms and their use by non-natives. Petch-Tyson (2000, 52) conducted a comparative analysis of the forms between natives and non-natives of various L1s. The study shows that variations exist in the uses of

demonstratives depending on the learners' L1. Petch-Tyson's study gives indicators concerning overuse and underuse of the forms depending on the L1s. Significant results give a baseline of use per L1 and per functional realisation. However, her study does not provide quantitative information concerning the factors which are involved in the use of the forms, nor does it provide confidence intervals for the frequencies. In addition to Lenko-Szymanska's aforementioned study, Liang (2009) did a frequency analysis of the distribution of the forms in three English written corpora of Chinese L1. The results give indications of most frequently used forms without any significance tests. (Cornish 2010) provided an account of third-year French students' judgement on specific occurrences of *that* in native English. (Bordet 2011) analysed the way French PhD students make use of *this*, as a discourse marker, in their thesis abstracts.

All the aforementioned studies have shown that *this* and *that* are used differently by learners depending on factors such as their function as determiners and pronouns or their referential (non-) situational values. In these studies, evidence of overuse and underuse links specific syntactic and semantic distinctions to the speakers' L1. Nevertheless, their approach is syntagmatic as they only focus on *this* and/or *that* in their utterance contexts and regardless of other forms with which they may interact. The next section focuses on a paradigmatic approach of the issue.

3.2.2 Error typology

In this section, we carry out a qualitative analysis of learner use of *this* and *that* along the paradigmatic axis. The objective of this novel approach—the forms are not looked at syntagmatically in order to justify their use—is to extract learner uses of the two forms and to observe them according to different criteria such as erroneous use, semantic context, functional category and error type. The underlying purpose is to provide a typology of errors by linking possible substitutions at functional and contextual levels with incorrect or unexpected uses.

Based on a subset of the Diderot-LONGDALE corpus[7] of 48 manually transcribed recordings of French learners of English, and totalling more than 300 minutes of speech, we manually analyse occurrences of *this* and *that* recorded in a free speech context where speakers explained one of their favourite experiences and answered questions on their daily life from a native speaker. Our goal is to address the following question: What kinds of errors do learners make when referring to an entity?

In order to answer this question qualitatively, 406 learner utterances are manually collected from the corpus and placed in a spreadsheet for sorting. The objective is to classify them according to several criteria. Each utterance that includes an occurrence of *this* or *that* is placed in a specific column. We also add utterances that include the *it* pro-form when they appear as erroneous in contexts which would support a demonstrative. Other informative columns are added: the determiner or pro-form function of each occurrence, the endophoric or exophoric characteristic of the context, the erroneous uses of the forms.

As far as erroneous uses are concerned, the annotation process must be evaluated with regard to natives. In order to assess our annotation predictions of erroneous uses, it is necessary to identify what the correct uses are in the same contexts as the learners'. For this purpose eleven learner erroneous utterances were proposed to a panel of six natives (see Annex A for the utterances). These natives were asked to fill in gaps corresponding to the positions of *this, that* and *it*. The results are shown in Table 5.

[7] http://www.clillac-arp.univ-paris-diderot.fr/projets/longdale (last accessed on March 31st, 2016).

	Rater A	Rater B	Rater C	Rater D	Rater E	Rater F	Learner
1	this	that	it	it	it	it	that
2	other	it	it	it	it	it	that
3	it	that	it	them	it	it	this
4	it	that	it	it	it	it	that
5	it	that	other	it	it	other	that
6	that	that	it	that	it	it	it
7	that	that	it	that	this	that	it
8	the	this	the	the	the	the	this
9	that	that	that	that	this	that	it
10	the	the	the	the	the	the	this
11	a	a	a	the	a	a	this

Table 5: Selection of forms by natives and non-natives for 11 specific utterances

Results show that, as well as selecting *this, that* or *it*, natives also choose the determiner *the*. In all cases though (except case 6), the vast majority of natives favours a different form from the learners' choices shown in the last column of the table. Nevertheless, agreement between natives is not as perfect as it could be as their choices show a certain degree of variability. As recommended by (Carletta 1996), computing the kappa coefficient for multiple raters reflects this reality:

```
> kappam.fleiss(df)
 Fleiss' Kappa for m Raters
 Subjects = 11
   Raters = 6
    Kappa = 0.42
        z = 10.3
   p-value = 0
```

The measure shows that agreement between the six natives is fair according to Landis and Koch's scale (1977) reported in (Artstein and Poesio 2008). This judgement means that there is no substantial agreement between speakers. Carletta's paper reports on a minimum 0.8 value to allow tentative conclusions (Carletta 1996). At any rate, these findings highlight the existence of a competition between the forms at determiner and pro-form levels. The results show a certain degree of variability in the selection of a form among natives. There are cases in which all natives disapprove the learner's choice and cases in which some natives agree with the learner. In spite of this variability, the results also show that learners' choices are not favoured by a majority of natives. Instead of qualifying learner's uses as errors it might thus be more appropriate to refer to them as *unexpected* v. *expected* uses

The aforementioned data collection protocol and the results lay the ground for a comparative approach in which specifically filtered-out utterances are isolated. When observing occurrences including morphemes used in an endophoric context, two main patterns emerge. Errors can be detected within the deictic system itself, as substitutions between one and the other can be found. The second pattern that stands out indicates that errors do not just occur within the deictic system but find their origin in the interactions that exist, depending on the function of the demonstrative, between the deictic system and the determination system, or between the deictic system and the pronominal procedure. These interactions between various grammatical forms denote the existence of two microsystems. We successively present the two patterns by focusing, first, on substitutions in Section 3.2.2.1 and, second, on interactions in Section 3.2.2.2.

3.2.2.1 Substitutions

In this section, we show that two types of substitutions exist. They may occur between endophoric and exophoric contexts and they may appear within endophoric contexts.

3.2.2.1.1 Endophoric/exophoric substitutions

Referring to the diagram on the selection process of the two forms (see Chapter 2, page 75), some occurrences suggest that errors can be traced back to the level of context and more specifically the type of referential process involving endophoric anaphora and exophora. Some errors show a confusion in the referential procedure where one form is replaced due to a confusion between these two levels. The error in this case consists in applying a form with an exophoric deictic sense in an endophoric anaphora context. This results in a blurred referential process, as the referent becomes unclear due to the absence of a clear entity in the situation of communication.

26) .. we can't go out (eh) for a long time so (er) . (eh) .. we used to (erm) . to p= to pass our . to pass time (eh) . in some shops or (er) .. or church (eh) . like *this* (erm) … (DID0150-S001)

Here, the demonstrative *this* is used with its collocate *like* in order to perform a situational reference process with something new seemingly pointed at. However, in actual fact, the addressee's expectation is an endophoric anaphora retrieval process of the previously mentioned entities such as *church* and *shops*. The error probably finds its root in the lack of command of the learner who finds it difficult to finish his sentence and thus uses a collocation that gives the impression of summarising what he actually means. From a research point of view, it is hard to collect such data, as there is no specific chain of words that indicates an exophoric or endophoric level.

Another typical occurrence of confusion between exophora and endophora among learners appears when temporal reference is at work. Fraser and Joly (1979, 142) showed a difference between the temporal uses of *that* and *this*, *that* referring to endophora and *this* referring to exophora. In English, *this* is seen as coinciding with the moment of utterance. The form is used to refer exophorically to a present-future period in time whilst *that* is used to refer to past events. Fraser and Joly point out the problem:

> "Whereas *this summer, this evening* only have a prospective value because they refer to the notion of present-future, the French instances of *cet été*, *cette nuit* may well refer to the past and the present." (Fraser et Joly 1979, 142, my translation)

They show that the use of temporal *this* and *that* is linked to the contextual difference between endophora and exophora, *i.e.* the situational parameters required for interpretation of *this*, while *that* refers to a previously mentioned period of time. Some examples of student utterances show evidence of confusion between exophora and endophora as learners apply present-past retrieval processes with present-future retrieval processes.

27) <Speaker B> For example for instance Dumbledore the . Hogwarts' director (em) died at the end of the the the the movie </B> <Speaker A> oh </A><Speaker B> and Harry Potter at *this* moment was (er) has . (er) .. (em) .. Dumbledore has throw thrown (em) .. (DID0116-S001)

28) we we waited for people to . to arrive because I got a . I got earlier cause I haven't class *this* day . (DID0162-S001)

In Examples 27 and 28, the two learners referred to a past period of their time by using the wrong determiner form of the demonstrative. As expressed by Fraser and Joly, *this* can only be used in referring to an upcoming event seen as a situational element, which is not the case here. So learners, use *this* with an exophoric retrieval procedure in an endophoric context (the dates have already been given), which erroneously locates the event in a time period yet to happen. Example 27 specifically reveals a different use of deixis between French and English speakers. In French *ce,* and its cliticised version *c'*, are used in the following prototypical expressions: i) *en ce moment* and *ii) à ce moment (-là)*. In English, these two expressions yield three different determiners: *the*, *that*, *this* as in i) *at the moment* and ii) *at that moment or at this moment (point in time)*. The ternary subsystem of the three time-referral markers may not be identical in scope for the category involved. The ternary representation of deixis may be different as summed up in Table 6. We can see the importance of prepositions (and therefore of our feature *oblique*) and of -ED in the analysis of these markers. The French language seems to have a ternary use of the same word. *Ce* is used to mark an exophoric reference to the temporal sphere at large (first interpretation of i)) or to the *hic et nunc* (second interpretation of i)). The third use is endophoric. Conversely, English speakers make a distinction between contexts (endophoric or exophoric) to choose a different determiner. All this seems to point out a sub-system for temporal reference: preposition + determiner + *moment*. French speakers apprehend it with one determiner in mind whilst English speakers avail of several options.

	Immediate *hic et nunc* reference	Current events looser present reference	Past or future reference
French	C'est le moment	En ce moment	À ce moment-là
English	This moment	At the moment	At that moment

Table 6: A comparative subsystem of temporal reference

The two previous examples must be contrasted with Examples 29 and 30, in which learners make an acceptable[8] use of *this*. In this case, recordings took place in October, which was still the same year as the summer they refer to. Consequently, *this* has an exophoric value as it is necessary to know the year to understand the reference. This contradicts somehow Fraser and Joly's 'prospective value' of *this*, and shows that the use of deictics is not clear cut.

29) so *this* summer I worked at (eh) Disneyland Paris (DID0148-S001)

30) oh . I don't know I haven't been to many places and (em) *this* summer I went to Bordeaux (DID0117-S001)

One more comment should be made about the possibility to use *last* in examples 29 and 30. It shows that the utterance's location in time has an impact on determining a period of time. It reveals a micro-paradigm[9] in which *last* and *this* compete to determine days, months, and seasons. Depending on specific exophoric characteristics, speakers will favour one form or the other. Table 7 summarises the micro-paradigm.

	prospective *past* reference	Current events looser present reference (October)	Past reference (after January)
French	Cet été	Cet été	L'été dernier
English	This summer	Last summer this summer?	Last summer this past summer (written data)

Table 7: Difference in temporal scope for deictics within a time referral micro-paradigm

[8] Natives have endorsed this possibility (as well as last summer).

[9] Because this micro-paradigm seems very restricted, we plan to investigate it a later stage, our tagging strategy being more general in the determiner v. pro-form distinction. See Laurie Buscail's PhD (Buscail 2013) for a more systematic investigation of deictics in French and English based on spoken PFC and PAC corpora.

3.2.2.1.2 Internal endophoric anaphora substitutions

What is meant here is the substitution between the two demonstratives within endophoric anaphora (see Chapter 2 Section 2.3). The classification of a form as an error may, in this case, be subject to much debate due to the fact that sometimes both forms seem interchangeable in a given context even though they are not most of the time, especially in anaphoric procedures (Fraser and Joly, 1980, 28). So in a lot of cases, it appears that choices denote variations in meaning effects. However, some choices remain incompatible with their context as suggested by tests on natives with the same contexts.

31) I go to I went to (em) (em) I don't remember (em) oh I don't know I don't remember the name (er) it's (em) near to north Vancouver there's (em) another ca= another bridge a suspension bridge but very short but there is a: forest and you can you can (em) walk on the forest . it's great <begin laughter> there is a big tree very big <end laughter> I don't see a tree like *this* very highs (DID0114-S004)

32) I took medias I got the media class that I didn't have in France . and it was pretty cool *this* is the only class that was really changing . a lot (DID0167-S001)

33) you can you can you can't lose yourself in London because there's a: a lot of underground and and buses *this* is what I like (DIDO118-S001)

Examples 31, 32, 33 were chosen due to the lack of ambiguity they offer. In cases 31 and 32, the expected meaning effect is to refer objectively to an entity by distancing it from the speaker's sphere. *That* would have been more appropriate especially if we consider that the past tense should be used. In both cases, the retrieval process is carried out correctly as the speaker clearly activates an endophoric anaphora context—also called co-discourse anaphora by Cornish (1999). The learner error occurs in the second phase of the referential construction when the selection of the right indexical must be made. Ariel (1994, 4) clearly specifies the requirement of the existence of a link between referring expressions and the level of saliency of their referent. In both cases, the entities *three* and *class* are already activated in the discourse of both interlocutors and there is no need to

refocus on them. The use of *this* does exactly that. It is counter-productive to try to bring to the foreground a discourse entity that is already fully salient. *That* with a focus maintaining value would apply better in these two contexts. In 33, the expected meaning is to conclude on the list of reasons why she loves London. Here, the learner has correctly identified the endophoric anaphora retrieval process as she acknowledges the already existing status of the entity in the discourse. Nevertheless, the selection of the right form fails in the second stage of the referential procedure as the identification of the selection of the form does not match the conclusive value of the statement. The use of *this* shows the characteristic of focus activation of the entity in the discourse while in fact the learner wants to de-activate it. Her pragmatic purpose is to conclude, in which case, *that* would be more appropriate.

3.2.2.2 Interactions: two microsystems

In this section, we show that depending on the determiner or pro-form function of the forms, learners tend to make substitutions with other forms, hence revealing the existence of paradigmatic interactions. By favouring a novel paradigmatic approach which correlates incorrect utterances with the two functions of the forms, we uncover the existence of two microsystems based on each function.

3.2.2.2.1 Endophoric use: substitution *it* v. *that/this*

Within the endophoric procedure, the pro-form use of the demonstratives appears to interact closely with the pronoun *it*. For some learners the choice of the personal pronoun versus the demonstrative pronouns appears to be a difficulty that leads to errors. This phenomenon is to be expected as all the forms have the same syntactic function.

The occurrences that follow include errors which show that learners understand the referential procedure but do not understand the conditions in which they can use the demonstratives. The following table lists a series of occurrences classified

according to the function of the form. By native use, we refer to the fact that the occurrences were also submitted to natives in order to see what choices they made in the same contexts. A classification is proposed according to native choices (first column). This column indicates the form the natives chose for the utterance quoted in the "learner use" column.

Native use	Function - Position	Learner use
It	Pro-form Subject	I liked (er) . the Independence Hall ***that*** was really interesting . and (er) . I went to: the: U Penn U Penn University and I really liked it . (DID0160-S001) we we see a (em) a romance for (em) the guy's eyes because most of the time ***that's*** the girl who is telling the story about was bad and and blablabla (DID0121-S001)
It	Pro-form Object	<A (native speaker)> would you consider pizza an Italian food </A> <B (learner)>(em) yes but it's not it's not really f= it's typic but it's not (em) we can eat that everyday everywhere now and . but (em) my grandma does ***this*** by herself (DID0115-S001) first of all what I loved the most of the world is ice hockey . just but I I loved ***that*** (er) I loved ***that*** before going there (DID0158-S001)
This/that	Determiner Subject	(13) so at the end she's an old lady she writes a book . and actually the book the scene that we saw when she apologize is (er) . she wrote that story on her book so ***it*** was her way to (eh) try to (em) to apologize to them through the book but (er) </B> (DID0164-S001)
This/that	Determiner Object	French French is very proud actually and they say yeah we're very open minded we can yeah but that that's not true we can see ***it*** with all the problems in this at this moment (DID0118-S001) (er) you know you can talk with people in the streets (er) it's (er) . it's really nice I'm not used to ***it*** in France so (DID0157-S001)

Table 8: Interactions between the demonstratives and *it*

As can be seen, errors occur in all four cases of use. Since this is a qualitative approach, statistics about each case have not been compiled (see Chapter 6 for a statistical analysis of learner use of the forms). Here, we characterise tendencies by means of authentic examples which have been sampled from the Diderot-LONGDALE corpus. This method does not assume any statistical accuracy, which remains a task to be completed because such information would allow us to have a better view on the major sources of error. In order to achieve such a task, it seems relevant to annotate the determiners differently from the pro-forms as both types of

information would help identify the functional realisation of each form and allow accurate calculations. This means that the distinction must rely on linguistically-rich annotation data found in the present qualitative analysis.

So what are the elements that help to distinguish unexpected from expected uses between the demonstratives on the one hand and the pronoun *it* on the other hand? At discourse level, Stirling and Huddleston (2002, 1507) note the possibility of finding a form within an anaphoric chain "with *that* anaphoric to the preceding clause but antecedent to the following *it*". They provide us with an example: "He discovered that she had slept with several other boyfriends before him. *That* shocked him a good deal, and they had a quarrel about *it*". Pierre Cotte describes the anaphoric process of *this* as necessarily being mentioned shortly after the construction of the existence of the referent (Cotte 1993, 58). In other terms, Cotte proposes that the following succession of items is to be expected: [Reference via Noun Phrase] → (followed by) [Reference via *this/that*] → [Reference via *it*]. This claim is yet to be tested in corpus based environments, with a specific focus on the fixity of the forms.

At text level, in (Cotte 1993, 57-58), demonstratives are said to point to the referent when *it* only repeats the referent. This assumption could be linked with the idea supported by Cornish (2010, 221) with the discourse functional approach in which the saliency of the referent depends on the anaphor that is used to refer to it. Normally, the more salient a referent is, the less it needs to be highlighted and so the choice of the anaphor varies. In Cornish's scale, the demonstratives in their pro-form function are situated next to another pro-form that is the third person personal pronoun. One hypothesis is that this proximity causes difficulties for learners. They might have problems in measuring the degree of accessibility carried by the form and, thus, inadequately use one form instead of another. *It* is a form that is stronger in degree of anaphoricity than the demonstratives, so it relies on a more salient entity than *this* or *that.* In some cases, the information carried by the form clearly does not require any more focus and the pronoun *it* is then sufficient

for the reference. However, learners show evidence of problems in this area. In the DID0160-S001 and DID0158-S001 examples of Table 8, the entities “Independence Hall” and “ice hockey” have just been mentioned in the co-text and so their degree of saliency at discourse level is high. In each case, the entity is clearly established in the memory of the interlocutors. Consequently, the use of *it* should be sufficient and the use of *that* appears as unexpected. In the DID0115-S001 example, the new information item is 'grandma' so 'pizza' should not receive renewed focus. The fact that a demonstrative is chosen instead of the pronoun introduces a logical contradiction, hence the unexpected use. Conversely, there are cases in which the learner only repeats the referent with *it* where, instead, a demonstrative should be used to bring a specific meaning expected to be found in the context as in DID0164-S001 and DID0118-S001 of the table.

Still at text level, the DID0121-S001 example can be looked at in light of Biber *et al.*'s definition of anticipatory object where “a clause has been extraposed” (Biber *et al.* 1999, 332). The learner has opted for an extra-positional construction that relies on a non-referential *it*. However, the learner envisages the form and its position as being referential and, thus, uses a referential form with *that*. This does not explain why the error is made but it shows that the extra-positional construction is conducive to errors.

3.2.2.2.2 Degrees of specificity: substitution *the/this/that?*

A second microsystem[10] that seems to apply to the use of demonstratives by learners appears when observing them in the determiner position. The following grid lists a certain number of occurrences where the determiner *the* is not chosen by learners even though tests on natives (see protocol of the test presented in Table 5 page 104) show its selection. Conversely, learners may choose *the* while natives favour *this* or *that*.

[10] Presented in (Gaillat 2013a).

Native expectation	Learner use
The	French French is very proud actually and they say yeah we're very open minded we can yeah but that that's not true we can see it with all the problems in this at ***this*** moment (DID0118-S001) <A> okay how old were you when you visited for the first time </A> <B> I was twelve years old and I went there with my mum because my family lives there my grandm my grandmother and my my uncles and my cousins (laughter) </B> <A> (mhm) And do you have any particular memories from that first trip </A> <B> oh yes because I didn't know how to speak ***this*** language so when I went there I didn't even know how how to say . hello to my grandmother (DID0112-S001)
This/that	I remember there was this little comic book store just just . like one block away from the . the: the Empire State Building so every time I like had an hour I would go and buy my coffee at Starbucks <begin laughter> <end laughter> anyhow (er) . Greenwich Village was very nice (er) I remember spending a lot of time at (er) ***the*** . this bookstore called Barnes and Nobles . and . those were . like it was like huge in the middle of Greenwich Village . so I spend a lot of time over there . (er) .(DID0155-S001)

Table 9: Interactions between the demonstratives and article *the*

As previously performed with *this/that* and *it*, it is relevant to analyse the occurrences in which the variations between native and learner English occur so as to pinpoint elements that may be the causes of such variations.

First, it is important to note that demonstratives are "closely related in meaning to the definite article" (Biber *et al.* 1999, 272). It can therefore be argued that it is within this proximity that learners tend to make the unexpected selection when constructing the degree of specificity of the referent. It appears important to envisage the determination process as a gradual referential construction in which various stages can be identified. As Biber *et al.* put it, "in addition to marking an entity as known, they—the demonstratives—specify the number of the referent and whether the referent is near or distant in relation to the speaker" (1999, 272). This statement is interesting insofar as it shows the existence of two stages. One related to marking the entity and the other one related to giving the construction a meaning of proximity or distance. Even though we have argued that such a meaning does not suffice to explain the use of the determiners, Biber *et al.*'s assumption shows the gradual approach in the construction. According to the authors, the degree of specificity is also performed via the phonological role given

to the morpheme: “In addition, the demonstrative determiners are stressed, whereas the definite article is almost always unstressed” (1999, 272).

Guillemin-Flescher (1993, 181-208) covers the issue of deixis and more specifically the various features that characterise the determiners *this, that* and *the*. They include an element of distinction with the *th-* part of the morphemes. In other terms, the first stage in using a determiner is to distinguish the referent stating its existence. The distinction part, common to all three determiners, is then followed by two different patterns. The definite article helps identify the occurrence of the referent with its notional concept; a process similar to Biber *et al.*'s marked entity. The demonstratives allow for the localisation and differentiation of the referent within a given class. This distinction may be paralleled with Cornish's anaphoricity scale (see Section 2.4.2 in Chapter 2) because the demonstratives, in their determiner function, are characterised as less anaphoric in degree than *the*. It ensues that the use of the definite article determines a more salient referent in the discourse than the demonstratives. The underlying assumption of the distinction function of the demonstratives is that the referent is less salient in the discourse and, thus, requires a stronger degree of accessibility which is completed with the demonstrative.

Our sample data suggest that the distinction between the use of *the* and the demonstratives is not always made correctly by learners. One first feature seems to appear with temporal referential terms such as *moment* as substitutions are found. The DID0118-S001 example in Table 9 illustrates this point in which the learner makes use of the determiner *this.* Tests[11] on natives show a clear preference for *the*. A closer look at the context shows that it is a situational reference as the moment referred to matches the moment of utterance. This moment is shared by both interlocutors and is clearly known. Consequently, the referential process corresponds to exophoric anaphora and it might be the reason for the confusion since the learner may have internalised the use of the deictic form to refer exophorically. Nevertheless, *moment* being a known parameter of the common

[11] When we use this expression, we refer to the test presented in Annex A.

situational knowledge, natives prefer *the* as it is sufficient for referent accessibility (see Section 3.2.2.1.1 for comments on the existence of a micro-paradigm with *moment*). This happens to the extent that "at the moment" has become a collocation (the COCA[12] spoken subset shows a frequency of 31.96 instances of the collocation per million words). For a discussion on the use of prepositions and determiners with *moment*, see Section 3.2.2.1.1 page 108.

Another feature of misuse appears in the DID0112-S001 example of the table which is characterised by an anaphoric referential procedure via an endophoric context. In this case, the learner has not yet established the entity *language* in the discourse. The referent is not salient, even though it might be argued that it could be a case of associative anaphora in the sense of Kleiber (2001), *i.e.* the entity *country* would trigger the mental construction of the entity *language*. However, the association is not strong enough since the choice of *this* implicitly evidences a process of distinction with another *language* entity of the same class. However, none other has been mentioned in the previous context. Consequently, the use of *the* appears more appropriate to natives as it would allow this association and, thus, the endophoric process to be carried out fully. In contrast, *this* calls for an absent second entity, which results in a blurred referential process. To summarise, there is a tendency to over-determine referents that only require marking with *the*. In English, the *hic et nunc* is not marked the same way (see Section 3.2.2.1.1 for comments on the existence of a ternary deictic system in French).

The DID0155-S001 example of Table 9 shows evidence of the opposite trend. Here, the learner uses *the* instead of *this* but immediately corrects herself. We show this example of self correction as no other such type of error, corrected or not, was found in our sample data. In spite of its rare occurrence, its presence denotes the potential for confusion. In this example, *the* is initially used to refer to the entity "book store" which has not yet been constructed at the level of discourse. Hence, a broken referential process, which is rapidly repaired by the learner. To summarise,

[12] The COCA corpus is accessible at http://corpus2.byu.edu/coca/ (Last access March 31, 2016)

some learners do not make the difference between a reference procedure which is only limited to marking an entity as known, and a procedure which requires the need to single out a referent in relation to another one, both of them linked to the same class entity, *i.e. this* or *that.*

3.2.3 Defining a research question

Like in some previous research, our approach aims to distinguish the functional uses of the forms in various native and non-native speaker contexts. However, it is also fashioned in a manner that compares the forms on a paradigmatic axis. Learner language is not only characterised by dominant uses of the forms but also by unexpected uses of forms which interact with each other. We have shown that learners' choices of referential forms evidence variability in terms of syntactic position and thus semantic value. Depending on their position, they interact in two different microsystems where learners tend to have confusional selection processes. In light of the reference model set in Section 2.5.2, we have characterised the two microsystems according to functional, positional and semantic criteria: one based on the pro-form and the other one based on the determiner function of interacting forms. These two microsystems open two respective lines of research.

We have decided to narrow down our research question to that of the pro-form microsystem. We intend to confirm (or not) the existence of the pro-form microsystem and to assess its significance. We also need to identify the linguistic features whose variations impact the expected or unexpected outcome of learner utterances. Our research tackles the question of learner errors and their identification in large corpora. To follow this line of research, we need to use the typology presented in 3.2.2 to identify criteria that help model the microsystem.

3.3 Modelling the pro-form microsystem for a large-scale investigation of *this* and *that* in learner language

In this section, we model the learner-specific pro-form microsystem of *it*, *this* and *that*. In doing so, we identify the criteria which act as variables in the system. These criteria stem from the linguistic features used in the typology and are to be introduced within the annotation schemes of several corpora. In this manner, the microsystem becomes observable and automatic analysis of learner output is possible. To describe the model, semantic and syntactic features are described successively in the next paragraphs.

3.3.1 Semantic features

The typology has led to the identification of two types of unexpected uses of the forms by learners, one of which corresponds to the semantic level. The substitutions that were described (see Section 3.2.2.1) involve the semantic notions of endophora and exophora intertwined with the semantic notions of deixis and anaphora as presented in Section 2.3. It was established that some learners experience difficulties in identifying the referential processes imposed by the contexts and this leads them to choose forms that appear unexpected, if not incorrect. Therefore, it would be relevant to investigate in what way the endophoric and exophoric features combine with the notions of given and new information.

Following our previous analysis, there are several hypotheses to be tested. The confusions all result in unexpected choices at different stages of the referential processes. They may firstly mismatch the givenness-or-not nature of the referent with the kind of cognitive process to be implemented (see Section 2.4 for details on referent accessibility). Consequently, they may choose the wrong coordinate or domain for the retrieval process, *i.e.* which context for their utterance. This raises the question of whether learners can identify or not the type of reference they need to use. To analyse occurrences of *this* and *that*, it would appear relevant to collect semantic information related to a context feature that can take endophoric and

exophoric values. The second kind of semantic information to collect would be related to discourse. It could take the form of a discourse feature that could be assigned two values: deixis or anaphora.

The second stage of the referential process that may be a hindrance for learners is that of choosing the right form. The purpose is to match its potential accessibility with the actual degree of saliency of the referent in the unfolding discourse. It is also to match the form with the type of vision that the learner has of the entity s/he is constructing, *i.e.* whether the entity is considered in the speaker's sphere or not (see Section 2.3.2). This means that semantic information related to saliency and speaker's stance on the referent should be taken into account.

Another, more peripheral, element for the semantic study of *this* and *that* is related to the pronoun *it*. As shown in the previous section, there is a micro-paradigm related to the same syntagmatic position that *it, this, that* can take. In order to investigate this microsystem, it is first necessary to isolate referential forms of the pronoun *it* only. Indeed, *it* can be found in extra-positional, impersonal, and cleft constructions and also in constructions related to time, distance and weather. In all these case, uses of *it* are not anaphoric and do not refer directly to salient entities as explained by (Huddleston and Pullum 2002, 1481). As a consequence, the analysis of the micro-paradigm requires the exclusion of non-referential *it* pronouns. It means that some form of coding should be implemented to provide the distinction between referential and non-referential pronouns.

3.3.2 Positional and functional features

The second type of unexpected uses identified in the typology of Section 3.2.2 is related to the interactions of the two forms with it and the. Paradigmatically, and depending on their position as subject or object, *this* and *that*, in learner English, seem to compete with the pronoun *it* and with the article *the*. This constitutes two learner-specific micro-paradigms that need to be investigated in order to understand the reasons that led to particular choices. Two questions arise in the

study of these two systems. Firstly, is the function taken up by the form that of a determiner or a pro-form? And, secondly, within the pro-form-based microsystem, does the syntagmatic configuration correspond to an oblique or nominative case?

Let us first focus on the first question. One assumption of this work is that grammar is guided by the expressional needs of the speaker. In other terms, forms are used to realise a number of different meanings and "a specific meaning can be performed using a variety of linguistic forms" (Ellis and Barkhuizen 2005, 111). In the case of *this* and *that* competing with other forms, the functional criterion comes as a primary distinctive feature as it entails the distinction between the two systems. At discourse level, the first thing that learners have to do is to either name an entity in the act of utterance or they just have to refer to it if it is already salient. Depending on the learner's discourse construction, *this* triggers the use of either the determiner function or the pro-form function. The problem is that either of these functions can be performed by more than one form, hence the difficulty to select the correct one in a given context. It is this intrinsic form variation—added to the natural native variation—for each of the two functions, which is at the heart of each microsystem. Any study on the matter would have to integrate this functional distinction by labelling the forms according to their functions. This labelling system would then allow for the disambiguation of all the forms present in a corpus and thus support a distinctive quantitative survey of the forms.

The second question corresponds to the pro-form microsystem. The positional criterion plays a role because the position of the form in the syntagmatic chain appears as a feature to sort out learner occurrences. Oblique and nominative cases seem to play a distinctive role in referential processes. Some learners show evidence of linguistic variability in the use of the forms in the object position when others show variability in the activation of the subject position. There are reasons to believe that learners do not handle the oblique case the same way as the nominative case. A comparison between examples 34 and 35 shows that learners choose different forms in nominative and oblique cases.

34) <A> would you consider pizza an Italian food </A><B> (em) yes but it's not it's not really f= it's typic but it's not (em) we can eat that everyday everywhere now and . but (em) my grandma does *this* by herself (DID0115-S001)

35) we we see a (em) a romance for (em) the guy's eyes because most of the time *that*'s the girl who is telling the story about was bad and and blablabla (DID0121-S001)

This needs to be investigated in corpora with a form of coding that embeds the positional feature and its two subject and object values.

3.4 Summary

This chapter has a dual purpose. We have presented a state of the art in analytical methods of learner English and this lays the ground for the development of a typology of learner errors in the use of demonstratives. This endeavour requires the use of a hybrid analytical method grounded in two approaches: Error Analysis and form-function analysis. By manually sorting out a limited number of occurrences according to their determiner or pro-form functions and their endophoric or exophoric context, we manage to establish a typology in which substitutions between the two forms are not the only type of error. We uncover two paradigmatic microsystems in which the demonstratives are found to compete with determiner *the* and pro-form *it*. The question of the pro-form microsystem is chosen for further research.

The typology of errors helps model the pro-form microsystem with specific linguistic features that seem to have some weight in the use of the forms. To confirm the existence of the pro-form microsystem and the relevance of the features, evidence needs to be found in comparisons between several corpora of different L1s. A large-scale learner-corpus study, based on a form-function and CIA approach, would help understand how the forms are used and acquired by learners. In order to investigate the data, it is firstly necessary to add annotation information to the text sources. And so the linguistic features highlighted by the work on the

typology appear to be relevant annotation information to add to the corpora to be explored. The following features and their values represent the substrate (the main variables) of the qualitative analysis carried out on reference and learner errors:

- Context: endophoric/exophoric
- Discourse: anaphoric/deictic
- Function: pro-form/determiners
- Position: nominative/oblique

The ultimate stage before analysis is to operationalise these features as variables so that they can be taken into account within statistical models. The features we have just highlighted represent syntactic and semantic properties and the difficult challenge is to convert them into variables. In doing so, we transfer the linguistic realities into a framework which can be analysed with statistical methods. This approach does not take into account coreferential properties, nor does it solely focus on occurrences of forms within the deictic/anaphoric paradigm. Instead, we favour a mixed approach combining semantic and syntactic properties organised in relation to functional categories. This provides a degree of generalisation which ensures a level of robustness.

By developing an annotation scheme which relies on the aforementioned features, we prepare the corpora for the extraction of variables. In the next chapter, we analyse how an annotation scheme can be developed on several corpora and the type of annotation which needs to be specified.

Chapter 4 Annotation for interoperability

In this chapter, we successively present a state of the art in corpus annotation schemes and examine what kind of multi-layer annotation scheme can be applied to *this* and *that.* The purpose is to study how to put in place such a scheme to make several corpora interoperable and thus allow the large-scale study of the pro-form microsystem.

When dealing with several existing corpora, one problem is that they have been developed with specific annotation schemes and custom-made search tools to leverage their exploitation (Leech 2005, Section 1). Consequently, it is difficult to use the same interface and query syntax to search two or more corpora at the same time. Multi-corpus simultaneous queries are not possible and query results rely on different annotation labels, which makes result comparability weak. The solution proposed in this chapter is to apply the same architecture for the data and to resort to a single annotation scheme used for all corpora. Firstly, the data architecture needs to provide a common framework for storing corpus data including annotation. Secondly, the scheme needs to include several annotation layers that reflect the linguistic features identified in the typology of errors. Using a single multi-layer annotation scheme ensures a high level of interoperability between corpora. We use three corpora of different L1s in order to provide a base for linguistic comparability. We also raise the question of how learner errors need to be addressed to determine the best way to flag them. The large-scale study of all correct and incorrect uses of *this* and *that* linked to several criteria becomes possible.

The development of this chapter is twofold. In Section 4.1, a state of the art in corpus annotation is given in order to understand the variety of annotation schemes and the way to make corpus data interoperable. We assess the specificities of several types of annotation. In Section 4.2, we present three corpora and examine the case of *this* and *that* in terms of possible annotations for these three corpora. We devise a specific annotation scheme which relies on the linguistic features identified in the previous chapter.

4.1 On annotation schemes across corpora

This section is a state of the art in annotation schemes. Firstly, we review the different types of schemes that exist depending on the type of language (learner or native) they have been developed for. We then present the conditions for interoperability to see what is necessary for the use of several corpora. Finally, in Section 4.1.3, we present an overview of tools for two types of annotation schemes, *i.e.* PoS and error annotation. Automation is assessed according to these two types of annotation.

4.1.1 Types and structures of annotation schemes

There are many types of annotation schemes and Leech (2005, Section 2) provides a comprehensive list of the linguistic domains that they may correspond to. At phonetic level, annotation may be chosen to add "information about how a word in a spoken corpus was pronounced" or "information about prosodic features such as stress, intonation and pauses". At syntactic level, annotation may embark information related to the way a sentence is parsed in terms of phrase structure. At text level, semantic annotation informs the system on the specific meaning a multi-sense word can take. It is the principle of named-entity annotation in which category names are assigned to various words mentioned in a text. For instance, the entity *Paris* in a sentence may be assigned the category *city* or *actress* depending on its acceptions in context. Such a type of annotation lays the ground for automated extractions of all entities of a particular kind for example. At context

level, some pragmatic annotation regarding the “kinds of speech act (or dialogue act) that occur in a spoken dialogue” (Leech 2005, Section 2) could be introduced. In other terms, it refers to information on the communication need that is fulfilled with a given utterance. At discourse level, discourse annotation could help in marking coreferential ties between textual entities. In terms of style, annotation can also be added to indicate whether the utterance corresponds to direct or indirect speech or other speech styles. Finally, lemmatised forms of words can also be part of annotation. This consists in “adding the identity of the lemma of each word form in a text” (Leech 2005, Section 2). Gries and Berez (2014, 3–6) give a more detailed taxonomy of annotation schemes in which the previous domains can fit. At a text level, they list lemmas, Part-of-Speech and syntactic parse trees as basic and frequently used types of annotation. Annotation schemes on the meaning of text units such as metaphor, predicate argument information are of the semantic type. Annotation schemes dedicated to phonetic and phonological information or to prosody are part of a more general category dedicated to multimodal corpora which can also include annotation schemes on gestures. Next to these categories, they also distinguish learner-error annotation schemes, some of them covered in section 4.1.1.2. Finally, discourse-pragmatic annotations such as anaphora and coreference resolution information may be applied to corpora.

From the learner corpus researcher's perspective, annotation schemes are divided into two groups depending on the language they target. Schemes are either developed for the annotation of native or learner corpora. This distinction is important because it induces a difference in the way language features are characterised. In the case of native English corpora, annotation schemes initially focused on the description of the language in terms of word classes and later in terms of syntactic structures. As far as learner-language corpora are concerned, the focus was also placed on word classes but with an orientation towards learner error identification. These two kinds of annotation are presented successively in detail.

4.1.1.1 Native-language annotation schemes

Let us start with native English corpora and the two types of syntactic annotation that have been used to process them. The first type corresponds to word class information added to words. The working principle is to indicate the grammatical nature of each word in order to facilitate automatic or semi-automatic syntactic analysis as in the case of the Brown corpus (Francis and Kučera 1964) This corpus was the first of its kind to introduce what is called Part-of-Speech annotation. Once the words are labelled, it is possible to write queries that only retrieve the words of a specified label. This level of generalisation is a powerful tool to extract patterns that support tasks such as term disambiguation, automatic parsing, word frequency computation and so on. PoS annotation relies on a grammatical description of language and, therefore, is subject to discussions as regards the labels that need to be employed and the class boundaries they encompass. For instance, the demonstratives in their pro-form function may not be labelled as pronouns in some schemes while they may in others. In some schemes, they actually receive a label no matter what their function is. This difference reflects a difference in theories and it can turn out to be a hindrance when retrieving and analysing forms from corpora (see Section 2.3.3).

The choices made by researchers on the granularity of PoS annotation also impact the way words are labelled. Some schemes include a much larger number of labels as their purpose is to comprehensively describe the language. The downside of this is that the greater the number of categories, the less common the occurrence of specific categories. Consequently, the risk is that the large variety of labels may impair linguistic observations based on generalisation. One palliative is the development of hierarchical structures for PoS in which case a form is classified as a *noun* and with the assignment of a subcategory such as *common noun*.

The Brown corpus includes 82 tags in its PoS annotation scheme (Francis and Kučera 1964, 22-23)—also called a tagset—and other such schemes were

developed to annotate native corpora. Subsequent projects took two distinct directions. One approach was that of economy. The Penn Treebank was developed for the tagging of the *Wall Street Journal* corpus (WSJ hereafter) (Charniak *et al.* 1987). It is a corpus of 4.5 million words of American English and was collected between 1989 and 1992. It was annotated with 36 PoS tags only. The underlying principle was that of minimising the number of tags. The authors advocate for a simplified version of their tagset that stemmed from the Brown corpus tagset. They argue that tag redundancies can be eliminated on the basis of rationalising rules based on lexical recoverability, consistency for tags of the same nature and correspondence with syntactic function (Marcus, Marcinkiewicz, and Santorini 1993, 314).

A second approach was carried out in the opposite direction and several projects, which also stemmed from the Brown corpus, developed larger tagsets in order to be as comprehensive as possible. The British National Corpus was annotated with the CLAWS PoS tagset (Garside 1987). The corpus now includes more than 100 million words and is made up of samples of spoken and written English of different genres and domains. The tagset has evolved over time and went from 132 tags in the first version to over 130 in its current "C8" extended version. It was implemented in the online version of Brigham Young University COCA corpus (Davies 2009). A smaller version of 60 tags was also designed in order to increase the performance of automated tagging of large corpora. The TOSCA system (Aarts, Van Halteren, and Oostdijk 1998) was used to tag the British part of the International Corpus of English (ICE-GB hereafter) (Nelson, Wallis, and Aarts 1998). The specificity of the system is that it includes a mix of 93 tags that are hierarchically organised according to functional and PoS labels[13]. What is common to all these corpora is that they were tagged automatically with various software tools. This process is not without errors as reported by Dickinson and Meurers (2003). Due to tagging errors

[13] (Gaillat 2013b, 171) gives details on some of the aspects of the annotation scheme employed to tag *this* and *that* in ICE-GB. As well as a functional label such as NPHD for Noun Phrase Head, the two forms can be assigned a grammatical category with labels such as PRON or DTCE which respectively correspond to *Pronoun* and *Central Determiner*. A sub-category is added to provide further details on specific features such as DEM for *Demonstrative*.

the aforementioned corpora were also post-edited manually by experts for error correction.

The second type of native-language annotation that accompanies raw data of corpora corresponds to the way sentences are presented in the form of phrase structures (Chomsky 1957). Conceptually, sentences are split—by using constituency tests—into meaningful units which form constituents such as Noun Phrases and Verb Phrases. Other levels of constituent division can be added to eventually reach the PoS level of each word of a sentence. In practice, this kind of sentence analysis can be carried out with a parsing algorithm which procedurally links a string to its syntactic structure. The parsing process is not without errors as reported for example in the Penn Treebank project[14]. Introducing the phrase structures of sentences in a corpus design shows the constituents of sentences in the form of a tree structure, which in turn can be interpreted as a form of annotation. Many native English corpora rely on tree-like structures whose branch terminations appear to be PoS tags and words.

At this stage in our survey of annotation schemes, it is relevant to provide illustrations of such developments. Two examples of annotation schemes are presented below since some of their characteristics will prove to be useful for our annotation setup. To begin with, the ICE-GB corpus has been developed with both types of syntactic annotation. Data are organised in a tree-like manner to reflect the syntactic structures of sentences. However, at word level, data have been characterised by two levels. Strictly speaking, the PoS level includes 19 tags but each tag may be assigned what the authors call 'features'. These features include information related to number, transitivity, modality and so on. Taking into account these features within the tagset gives 262 possible combinations to tag words in the corpus (Aarts, Nelson, and Wallis 2007, 28). Figure 9 shows the result of the parsing and tagging processes of the phrase “It is the species composition of this layer which determines the treatment efficiency”. First of all, it is important to

[14] See Penn Treebank official site: http://www.cis.upenn.edu/~treebank/ for evaluation campaigns and precision rates (Last access March 31, 2016).

highlight the basic element of the scheme. A tree is made of a series of “nodes” that are tied hierarchically to each other forming the branches of a tree. As shown in Figure 8, “each node on the tree indicates function and category, and may carry additional features such as clause type” (Aarts, Nelson, and Wallis 2007, 29).

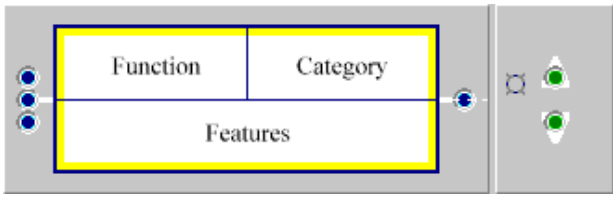

Figure 8: Function, category and features in the nodes used in the ICE-GB annotation scheme

Secondly, the sentence has been automatically parsed by the TOSCA system (Aarts, Van Halteren, and Oostdijk 1998) into a tree-like structure and tagged with PoS labels. On the left of the tree the type of Parsing Unit (PU) is further divided into branches that provide information on the constituents of the sentence with the PoS labels at their extremities. The leaves of the branches show the words. For instance, *this* is assigned the label DTCE for determiner in the function of pronoun. Its features are that it is a demonstrative and singular. This PoS label is dominated by a series of nodes that show how the word is integrated in the sentence. It is part of the Noun Phrase “this layer”, which is itself part of the prepositional Noun Phrase “of this layer”. This constituent is embedded in the NP “the species of this layer” which is a main NP element of the full sentence composed of the following chain: cleft *it* + VP + NP + clause + punctuation.

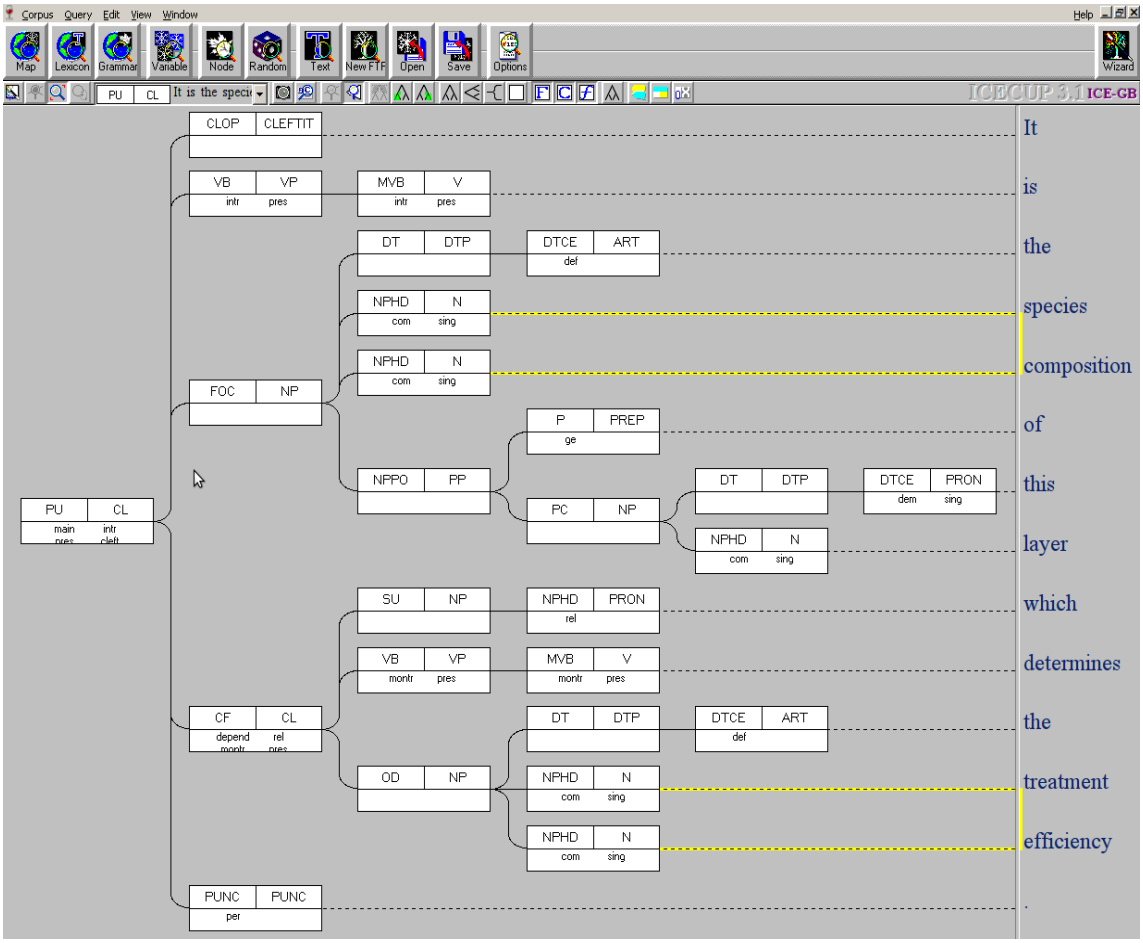

Figure 9: Tree-like structure of the TOSCA/ICE-GB annotation scheme

As a result of this annotation, it is possible to query the corpus on the basis of categorical, functional or featural[15] information that is in fact stored in text files in the format presented in Figure 10. The tree-like structure used in the Graphical User Interface relies on many text files. In the text files, each word form is placed between braces. Features are between brackets and functional and categorial information is displayed in capital letters. Each line corresponds to a text unit and its characterisation. A system of indentation determines the hierarchical organisation between units. In the case of "it is the species composition", *species* is a singular common noun N that is a Noun Phrase Head. The element belongs to a larger structure which is a NP classified as the focus (FOC) of the cleft construction.

[15] Features inform on specific uses of the forms. For instance, depending on their grammatical category, they may be used as comparatives or attributes (for adjectives) or in the singular or plural (for nouns) and will thus obtain corresponding labels. See http://www.ucl.ac.uk/english-usage/resources/grammar/index.htm for a full overview of the TOSCA/ICE Grammar and its label descriptions (Last access March 31, 2016).

This constituent together with other constituents of the same level, such as cleft *it*, form a present-tense intransitive main clause (CL) which is a Phrase Unit (PU).

```
1 [<#9:1> <sent>]
2 PU,CL(main,intr,pres,cleft)
3  CLOP,CLEFTIT {It}
4  VB,VP(intr,pres)
5   MVB,V(intr,pres) {is}
6  FOC,NP
7   DT,DTP
8    DTCE,ART(def) {the}
9   NPHD,N(com,sing) {species
  composition}
10   NPPO,PP
11    P,PREP(ge) {of}
12    PC,NP
13     DT,DTP
14      DTCE,PRON(dem,sing) {this}
15     NPHD,N(com,sing) {layer}
16  CF,CL(depend,rel,montr,pres)
17   SU,NP
18    NPHD,PRON(rel) {which}
19   VB,VP(montr,pres)
20    MVB,V(montr,pres) {determines}
21   OD,NP
22    DT,DTP
23     DTCE,ART(def) {the}
24    NPHD,N(com,sing) {treatment
  efficiency}
25  PUNC,PUNC(per) {.}
```

Figure 10: Example of data structure used in ICE-GB corpus files

The originality of this scheme is that each node embarks all three types of information (function, grammatical category and feature) and thus allows multi-level analysis. It is possible to write a query that searches forms which are pronouns but are not determiners. The relative pronoun *which* as an NP head *(NPHD)* would be a match but not *this* (tagged as a DTCE determiner). As a final comment on this corpus, we can say that the software and the annotation scheme allow powerful fine-grained queries. However, the relative small size of the corpus (one million words) and its commercial licence might be seen as hindrances to research programmes.

The second example of an annotation scheme is the Penn Treebank. It also uses the same two types of annotation as the ICE-GB corpus but it differs in the way the

labels are structured. As we have seen, the ICE-GB annotation scheme organises the data into nodes that may include phrase structure and PoS information. Conversely, the Penn Treebank relies on three independent annotation schemes. PoS tagging, syntactic bracketing and disfluency annotation are applied to texts from the Wall Street Journal and the Switchboard corpus. The three schemes are processed independently but the resulting data can be merged together (Taylor, Marcus, and Santorini 2003, 16). Figure 11 shows data including both PoS-tagging and syntactic-bracketing information. In terms of PoS, this information is put between brackets with the word, *i.e.* (VBZ is) for *is* as a third person singular verb. Concerning syntactic bracketing, sentences are organised hierarchically according to their NP, VP or other type of constituents. When constituents belong to the same level, they are indented identically with opening brackets. The resulting indented hierarchy represents a tree with branches made of dependent constituents and whose leaves correspond to PoS labels. For instance, an adverbial clause (S-BAR-ADV) and an infinitive clause (S) are positioned at the same level in the first sentence of Figure 11: “Mr Hahn is trying to entice Nekoosa into negotiating a friendly surrender while talking tough”. Each of these clauses dominates a series of other constituents. So the hierarchical structure is a representation of the domination ties that exist between constituents.

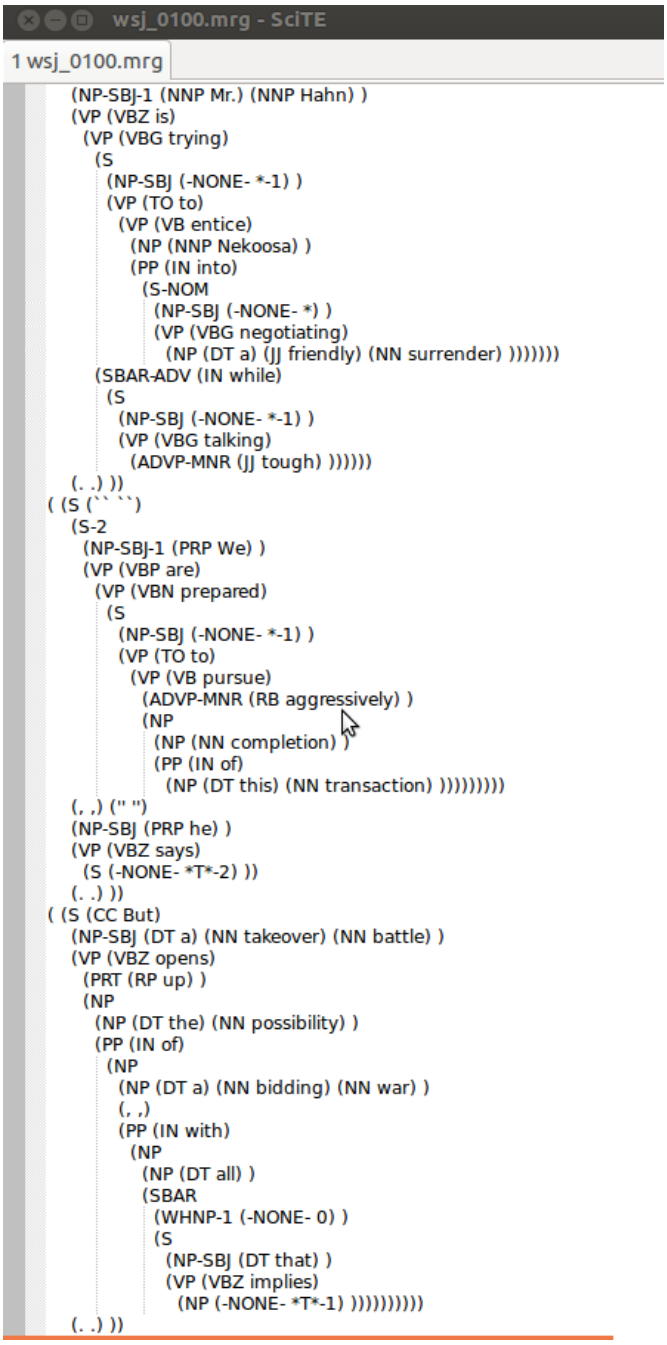

Figure 11: Extract from WSJ corpus with PoS-tagging and syntactic-bracketing information

The Penn Treebank combines a limited number of PoS labels together with hierarchically organised constituents, which enables the analyst to query the structure syntagmatically and paradigmatically. For instance, it is possible to extract all forms of a particular PoS label that only belong to a specific type of clause. It is also possible to identify specific forms according to their syntactic path. This can be useful for finer identification of the forms in a functional perspective.

For instance, the functional distinction of *that* can be addressed thanks to the identification of the syntactic paths. The pro-form *that* is dominated by an NP whereas the complementiser *that* is under the direct dominance of an SBAR clause. This process is detailed in Section 5.1.1.

This type of annotation scheme is very important for the purpose of our project. We are interested in the syntactic annotation of corpora and we intend to use automatic methods to apply functional annotation. At functional level, the use of automated tools can be envisaged to mark the pro-form or determiner functions of the forms. Our guiding principle is that functional annotation should stem from the functional linguistic features described in the previous chapters. Since these features are interpreted at context level our position is to hypothesise that co-text plays a role in the construction of an utterance containing *this* or *that* (Cornish 1999, 69-70; see also Section 2.4.1). Consequently, we propose a setup in which functional information will be assigned to word forms automatically. Several corpus projects such as the ICE-GB (Nelson, Wallis, and Aarts 1998) or the BNC (Burnard 2007) have proved the compatibility of a functional approach of word annotation with a PoS tagset. It appears reasonable to envisage the use of such a tagset for our purposes. Section 5.1.1 is devoted to this issue.

Syntactic parsing and tagging are not the only methods which are used to add syntactic and PoS analyses as annotation to native corpora. Much work has also been placed on other types such as phonology and discourse. In Chapter 2, we adopt a discourse functional approach for the analysis of referential forms and more particularly *this* and *that*. It is thus relevant to broach this issue as far as annotation schemes in computerised corpora are concerned. In this respect, discourse annotation needs to be investigated to see if there exists annotation schemes that would comply with the view expressed above.

The CESAX project (Komen 2012) appears to be a good candidate for this purpose. CESAX is a coreference editor for syntactically annotated XML corpora. Its main

purpose is to assist the linguist in the making of coreferential links within texts. To do so, it relies on the introduction of discourse rich information in syntactically parsed components of corpora. The tool offers a variety of functionalities to edit and format corpus files of different structures. On the whole, it takes charge of several annotation layers such as PoS, syntactic parsing and coreferential information. It also provides the user with conversion capabilities to transform Penn Treebank formatted corpus files into *XML* files which mirror the original syntactic trees and PoS, whilst fostering data exchange thanks to the encoding language. The working principle of the application is twofold. Firstly, the corpus is parsed syntactically and morpho-syntactically. Secondly, the file is processed for coreference information. Figure 12 gives an insight into the structure for the first phase.

b

```
<forest forestId="9" File="long-1866" TextId="long-1866" Location="2,76.9">
    <eTree Id="137" Label="IP-MAT" from="1" to="32">
      <eTree Id="138" Label="NP-SBJ" from="1" to="8">
        <eTree Id="139" Label="NPR" from="1" to="8">
          <eLeaf Type="Vern" Text="Diodorus" from="1" to="8" />
        </eTree>
      </eTree>
      <eTree Id="140" Label="BED" from="10" to="12">
        <eLeaf Type="Vern" Text="was" from="10" to="12" />
      </eTree>
      <eTree Id="141" Label="NP-OB1" from="14" to="31">
        <eTree Id="142" Label="D" from="14" to="14">
          <eLeaf Type="Vern" Text="a" from="14" to="14" />
        </eTree>
        <eTree Id="143" Label="N" from="16" to="21">
          <eLeaf Type="Vern" Text="native" from="16" to="21" />
        </eTree>
        <eTree Id="144" Label="PP" from="23" to="31">
          <eTree Id="145" Label="P" from="23" to="24">
            <eLeaf Type="Vern" Text="of" from="23" to="24" />
          </eTree>
          <eTree Id="146" Label="NP" from="26" to="31">
            <eTree Id="147" Label="NPR" from="26" to="31">
              <eLeaf Type="Vern" Text="Sicily" from="26" to="31" />
            </eTree>
          </eTree>
        </eTree>
      </eTree>
    </eTree>
  </forest>
```

a

```
( (IP-MAT (NP-SBJ (NPR Diodorus))
          (BED was)
          (NP-OB1 (D a) (N native)
                  (PP (P of)
                      (NP (NPR Sicily))))
          (. .))
  (ID LONG-1866,2,76.9))
```

Figure 12: XML structure of a CESAX annotated corpus (b) compared with its original Penn Treebank source (a) (Komen 2012)

It hinges on a set of XML elements that structure the annotation data which compose the corpus. Tag elements such as <*forest*>, <*eTree*> and <*eLeaf*> comprise attributes that allow the precise identification of the elements both in terms of grammatical nature and text position. For instance, the word *native* belongs to several indented <*eTree*> tags whose successive attributes indicate its

dominance path that can also be read in the Penn Treebank format (see frame (a) in the figure). As a result, *native* is a Noun and belongs to the end of the path as a <*eLeaf*> tag or element with respect to the XML terminology. The attributes of this tag give type information and also the unique location of each of the characters that compose the noun in the sentence. The sentence is also uniquely identified and the *to* and *from* attribute-value pairs show the ordinal positions of the characters that it spans. The six letters of *native* come between the 16th and 24th position of all the characters (including spaces) that compose the text and they fall within the 1st to 32nd position of the <*eTree*> number 137. Consequently, every word of every sentence is uniquely identifiable in a sentence which is uniquely identifiable in a text.

After detailing the initial XML structure of the corpus files, it is relevant to explore the way referential annotation is written within the files. After the semi-automatic or fully automatic coreferential process, an XML file is created by the program. The initial tree-based structure is repeated and extra information is added in relation to coreference.

Figure 13 shows an example of such an annotation on an extract of the Diderot-LONGDALE corpus (DID004-S001). The figure gives the XML structure of the sentence “So you chose topic number two”. <*eTree*> and <*eLeaf*> tags are the backbone of the structure used to tag the phrase. They maintain the hierarchical dominances between the syntactic constituents. Within the syntactic XML structure, a specific set of tags potentially related to coreferential annotation are placed. This set comes embedded within the <*fs*> tag for *feature set*. <*f*> elements compose the set and include a series of attribute-value pairs that provide information on the type of information related to the type of reference. For instance, referential information is added for the *topic number two* entity thanks to the code (in bold) that spans from line 28 to 33 in Figure 13. The code in bold indicates that the *topic number two* NP (whose Id=30) is a potential coreferential entity. What is relevant in terms of discourse is that the <*f*> element comprises linguistic features such as

the *refType* and *identity* values of two attributes that are part of the feature set. These features, together with other features applied to the noun phrase (see lines 35-39), enable the coreferential resolution algorithm to process the various entities represented as syntactic constituents in the structure.

"So you chose **topic number two**"

```
1 <eTree Id="23" Label="SBAR" from="66" to="94" IPnum="1">
2           <eTree Id="24" Label="IN" from="66" to="67" IPnum="1">
3             <eLeaf Type="Vern" Text="so" from="66" to="67" />
4           </eTree>
5           <eTree Id="25" Label="S" from="69" to="94" IPnum="1">
6             <eTree Id="26" Label="NP" from="69" to="71" IPnum="1">
7               <fs type="coref">
8                 <f name="history" value="Erwin R.
  Komen:AutoSusp(3/14/2014)" />
9                 <f name="IPdist" value="0" />
10                <f name="RefType" value="Identity" />
11                <f name="NdDist" value="6" />
12              </fs>
13              <ref target="20" />
14              <fs type="NP">
15                <f name="GrRole" value="Oblique" />
16                <f name="PGN" value="3" />
17                <f name="NPtype" value="unknown" />
18              </fs>
19              <eTree Id="27" Label="PRP" from="69" to="71"
  IPnum="1">
20                <eLeaf Type="Vern" Text="you" from="69" to="71" />
21              </eTree>
22            </eTree>
23            <eTree Id="28" Label="VP" from="73" to="94" IPnum="1">
24              <eTree Id="29" Label="VBD" from="73" to="77"
  IPnum="1">
25                <eLeaf Type="Vern" Text="chose" from="73"
  to="77" />
26              </eTree>
27              <eTree Id="30" Label="NP" from="79" to="94"
  IPnum="1">
28                <fs type="coref">
29              <f name="history" value="Erwin R.
  Komen:AutoSusp(3/14/2014)" />
30                  <f name="IPdist" value="0" />
31                  <f name="RefType" value="Identity" />
32                  <f name="NdDist" value="22" />
33              </fs>
34              <ref target="8" />
35              <fs type="NP">
36                <f name="GrRole" value="Oblique" />
37                <f name="PGN" value="3s" />
38                <f name="NPtype" value="unknown" />
39              </fs>
40              <eTree Id="31" Label="NN" from="79" to="83"
```

```
  IPnum="1">
41                    <eLeaf Type="Vern" Text="topic" from="79" to="83"
  />
42                  </eTree>
43                  <eTree Id="32" Label="NN" from="85" to="90"
  IPnum="1">
44                    <eLeaf Type="Vern" Text="number" from="85"
  to="90" />
45                  </eTree>
46                  <eTree Id="33" Label="CD" from="92" to="94"
  IPnum="1">
47                    <eLeaf Type="Vern" Text="two" from="92"
  to="94" />
48                  </eTree>
49                </eTree>
50              </eTree>
51            </eTree>
52 </eTree>
```

Figure 13: CESAX XML output file after the coreference resolution process of a sentence

In terms of discourse annotation, the *refType* name and the corresponding values of a feature element are very interesting since they could be applied to any component in the structure, namely indexicals in the sense of Cornish (1999). In Chapter 2, we endorse the discourse-functional view of reference in which deictic and anaphoric procedures are grounded in the principle of new or given information. We defend the idea that it is not a decisive element in the interpretation of discourse entities to identify some expressions as coreferential or not. Instead, we consider it essential to correctly identify the discourse referents that are targeted in what Cornish calls the "discourse model representation" (Cornish 1999, 22).

The aforementioned *refType* feature shows that CESAX embarks an entity recognition system that identifies the type of entity in terms of reference. The referential status that can be assigned to the entity is of crucial importance to see if it complies with the discourse functional view. Komen (2012, Section 2.2) follows this line of research in which the new/given paradigm serves as a guiding principle. Nevertheless, it still needs some refinements similar to those proposed by Ariel (1994). Komen's approach consists in assigning a reference type to syntactic constituents in a semi-automatic manner. This means that the user must ultimately

make a decision. (Komen 2012) defends the view of a simplified version of the various degrees of new/givenness information. Five degrees appear to him as sufficient to serve as primitives from which finer-grained distinctions can be made (Komen 2012, Section 2.2). *Identity*, *inferred*, *assumed*, *inert* and *new* compose the different values of the reference type feature. In the case of Figure 13 line 31, the *topic number two* NP is assigned an *identity* reference type. So what do these levels indicate?

The identity value indicates that the referent is identical to that of a previously mentioned entity in the text. The use of a personal pronoun such as *me* would most probably have a textual antecedent in the form of *I.* Their referent would coincide completely. The *assumed* value postulates that, for instance, in the case of *you* in Figure 13, the referent is common knowledge to both speakers, just like the name of a country would be. *Inferred* is a value that approaches Kleiber's associative anaphora. Part-whole relationships between entities allow bridging strategies in terms of word associations (Kleiber 2001). The *inert* referential status determines entities that have no referent by themselves. They are part of a construction in which they appear as a simple attribute to the main entity mentioned in the construction. Komen gives the following example: “[…] when H. M. S. Defence foundered, with all hands, in a gale of [np **wind**] in the Baltic in 1811”. In this example, *wind* is interpreted as an attribute of *gale*. Finally, *new* is the value applied in case an entity refers to a referent which is mentioned for the first time in the discourse.

As already explained, we consider that the foundation of reference does not lie within intratextual linkage between expressions (see Section 2.3.1). Instead, discourse functions are used to “ensure that the speech participants are on the 'same wavelength' with respect to their focus of attention at any point in the discourse” (Cornish 1999, 24). Focus of attention, in the interlocutors' minds, is the point of articulation of the referential processes. All this means that we distinguish discourse from textual features according to their focus function in discourse. It

appears that CESAX complies with this view when it deals with the assignment of the referential status of each entity. The only reservations we have would be for the *identity* feature since it indicates an antecedent in the discourse and this comes as a breach of the focus-centred view. Nevertheless, it still indicates the fact that an entity is already known and this is essential. The fact remains that the attempt to link up a given text entity with its antecedent is a point of debate. To summarise, the tool is very promising as it automates the assignment of values for referents. CESAX enables the user to quickly enrich corpora with focus-related annotation that can later be exploited in multi-factorial analysis. In section 4.2.2.5, we assess the possible use of this application on a learner corpus and the annotation of *this* and *that*.

Having reviewed the technical possibilities for the annotation of native speech, we now turn to non-native speech and the issues that it raises in terms of annotation.

4.1.1.2 Learner-language annotation schemes

As put by Meurers: "The purpose of annotating learner corpora is to provide an effective and efficient index into relevant subclasses of data" (Meurers 2015, 557). Thanks to this index, further processing and analyses are possible and the use of subclasses can help extract specific forms. In the following paragraphs, we focus on the types of annotation used for learner corpora.

The first type of annotation also used on learner language is PoS annotation. It has been applied to learner language with some significant levels of accuracy. (de Haan 2000) reports an experiment in which the TOSCA system was used with 95% accuracy for a corpus of advanced Dutch learners of English. (van Rooy and Schafer 2003) also evaluated the possibility of PoS annotation on learner corpora by comparing three specific tagsets (CLAWS7, Brill, and TOSCA). Overall accuracy levels range from 86.34% to 96.26%. In a study on specific functional PoS labels for *this* and *that* in a learner corpus, (Gaillat 2013a) reported overall accuracy of 91% when tagging a spoken corpus of French learners of English with a modified

version of the Penn Treebank. Overall, PoS tagging of learner corpora has revealed its interest and feasibility but experience shows that tagging error rates are higher than with native corpora. Post-editing is therefore necessary. Further research on how to deal with tagging errors is underway. (Díaz-Negrillo *et al.* 2010) propose the distinction of three layers of PoS annotation in order to encode the distributional, morphological and lexical aspects that are specific to learner language. As mentioned before, Lüdeling's (Lüdeling *et al.* 2005; Lüdeling and Hirschmann 2015) work on the FALCO corpus relies on a multi-level annotation scheme. She advocates for the use of several layers when interpreting the target hypothesis, *i.e.* the underlying assumption of the alternative and correct form of an error.

There is a second group of annotation schemes which is specific to learner language. One of the main specificities of learner corpora is the fact that they include a certain degree of errors committed by learners. These errors can be related to word morphology, syntax and lexis. Several research projects have been dedicated to the construction of specific learner-error annotation schemes. As argued by (Granger 1994, 26), computerised learner corpora allow for a more consistent approach to the analysis of errors by applying error specific tags to the raw data. A number of projects have focused on the matter by devising typologies of errors (see de Haan 2000; Nicholls 2003; Izumi, Uchimoto and Isahara 2005). They all tackle the same issues that are related to the choice of classes and subclasses to categorise errors, the need to insert a correct version of the erroneous form and the way to insert the annotation in the corpus. One of the most prominent of these projects is the ICLE corpus (Granger 1993), which relies on a specifically developed tagset known as the Louvain tagset. This tagset is organised hierarchically and uses a series of seven main tags that are assigned a variety of sub-tags to provide more details about errors. For instance, an error on a auxiliary might be tagged *GVAUX* because it is considered to Grammatically affect a Verb which is an AUXiliary. In total, there are 40 different tags that can be assigned to errors. This assigning process is carried out manually with a special editor and

under strict guidelines published in the accompanying manual. The tags are embedded into the texts that compose the corpus. Figure 14 shows how each learner error is accompanied by a typological classification and a proposed correction between two dollar signs. The advantage of such a system is that it "is possible to search for any error category and sort them in various ways" (Granger 2002, 15).

There was a forest with dark green dense foliage and pastures where a herd of tiny (FS) braun $brown$ cows was grazing quietly, (XVPR) watching at $watching$ the toy train going past. I lay down (LS) in on the moss, among the wild flowers, and looked at the grey and green (LS) mounts $mountains$. At the top of the (LS) stiffest $steepest$ escarpments, big ruined walls stood (WM) 0 $nsing$ towards the sky. I thought about the (GADJN) brutals $brutaI$ barons that (GVT) lived Shad lived$ in those (FS) castels $castles$. I closed my eyes and saw the troops observing (F$) eachother $each others with hostility from two (FS) opposit $opposite$ hills.

Figure 14: Sample of error-tagged text in ICLE

The Louvain tagset relies on a limited number of tags and its lack of granularity in terms of error description could be seen as an impetus to create a finer-grained tagset. Negrillo's EARS tagset (Díaz-Negrillo 2009) used for the NOCE corpus claims exactly that. "A finely developed taxonomy saves the stage of further categorization that generic tagset users should undertake if precise description and explanation of errors is aimed at" (Díaz-Negrillo 2007, 81). The Non-native Corpus of English (NOCE hereafter) is a corpus of Spanish learners of English composed of written productions of first year students at the universities of Granada and Jaén. It consists of over 300,000 words. Another claim for the development of EARS relates to the fact that some tagsets are multilingual in nature and are not specific enough in terms of errors related to English and to the learners' L1s (Díaz-Negrillo and Garcia-Cumbreras 2007, 198). To comply with these claims, the tagset was developed following a corpus-driven approach in which the taxonomy stemmed from the errors found in the corpus. It contains 614 tags which "are a representation of all possible error category combinations used for the description of the errors found in the corpus" (Díaz-Negrillo 2007, 81). Error categories are organised according to six levels of linguistic description:

- Spelling
- Punctuation
- Word Grammar
- Phrase Grammar
- Clause Grammar
- Lexis

Each level may be characterised by up to four layers that correspond to a particular aspect of the error:

- Unit identification for the type of unit involved in the error
- Error focus for the linguistic details
- Error scope for forms that are either well-formed but wrongly positioned or ill-formed corresponding to errors resulting from creative processes
- Error type for surface structure classification such as omission, misselection, ordering and overinclusion (Díaz-Negrillo 2007, 88).

Figure 15 presents an example of an error from the NOCE and how it is classified according to the tagging scheme. The error on the comparative form *happier* can be seen in different ways and EARS provides a comprehensive characterisation of this error.

```
[…]          they          are          <WG.AD.DG.CV.IT.MS>more
happy</WG.AD.DG.CV.IT.MS> […] GR-1-B-EN-027-Y
```

Layer	Category	Code
Linguistic Level	Word Grammar	WG
Unit Identification	Adjective	AD
Error Focus	Degree, Comparative	DG.CV
Error Scope	Internal	IT
Error Type	Misselection	MS

Figure 15: An occurrence of an error and its tagging with EARS tagset

From this example, the fine-grained aspect of the tagging process is clear. Furthermore, the analyst can retrieve data according to several parameters that span various facets of learner errors. For instance, it would be possible to retrieve all the errors of the corpus related to pronouns.

All in all, there are two sides to the coin. On the one hand, the information provided by PoS tags provides detailed grammar classes for words but taggers do not handle learner errors. On the other hand, the information retrieved by way of error categorisation provides in-depth knowledge of learners' difficulties but it may also provide a distorted image of what the learners' actual capabilities are since only errors are reported. They present a partial view of interlanguage (see Section 3.1.1). Should the interest of the analyst be on developmental patterns, errors would only provide information on the erroneous versions of the patterns used by learners and hide the correct uses that also occur due to learner language variability. It is therefore paramount to have annotations that support the analysis of developmental patterns. In this respect, error tagsets and PoS tagsets can be seen as complementary since the study of learner language requires a comprehensive view that provides linguistic and error-related information.

4.1.2 Interoperability

It is a broadly accepted fact that annotation enhances corpora as it provides a wealth of information about any raw corpus and its text-unit components. For Leech “adding annotation to a corpus is giving 'added value', which can be used for research by the individual or team that carried out the annotation, but which can also be passed on to others who may find it useful for their own purposes” (2005, Section 1). Leech points out two essential aspects, of which the second one, reusability, must not be underestimated. As more and more corpora have been developed, many annotation schemes of different types have been created and applied to corpora. Due to the diversity of approaches, the question of the reusability of these corpora and their annotation, outside of their initially intended

use, is raised. In other terms, for a given annotation type, such as PoS tagging, it is cumbersome to use several annotated corpora in the same survey, due to the impossibility of combining results that rely on text units and/or segments which have been annotated differently. Leech (2005, Section 4) points to the lack of agreement that may exist between annotation standards. Even though some level of consensus exists between annotation schemes, different codings may be used for text segments that are similar but whose boundaries may vary. For instance, some annotation schemes consider anaphora as a tie between a pronoun and its NP. However the tagging process of some NP boundaries may be unclear due to the order-entity category the NP belongs to, thereby the difficulty to clearly mark the elements referred to (see Section 2.2.1).

As well as the variety in annotations for one given type, the multiplicity of annotations also may appear as a hindrance to reusability since several corpora may embark different types of annotation. In such a case, their combination entails technical issues with regard to the requirement of making all these annotations accessible for queries. If, for instance, a corpus includes some discourse annotation on anaphoric relations between textual elements and another corpus includes annotation of the same level but with information on givenness, then both annotations should be made accessible to the researcher.

In fact, the two limitations described above and the questions they raise find their answers in the issue of interoperability. Interoperability can be defined in general terms “as the capability of language resources to interact or work together” (Sérasset *et al.* 2009, 5). In their paper, Sérasset *et al.* describe various types of language resources that can be classified multi-dimensionally. Resources can be static or dynamic. For instance, automatic PoS-taggers produce data, which makes them dynamic whereas corpora can be classified as static. Resources can also be either text-based or item-based. Taggers, for instance, focus on collections of individual items, *i.e.* tokens, whereas corpora focus on entire texts. Finally, resources can also be distinguished on their interpretative value. Raw data made of

handwritten material, for instance, do not result from any interpretation whereas primary data are the produce of some kind of interpretation, *e.g.* oral speech transcription. So taggers can be classified as dynamic, item-based, interpretative resources while raw corpora are text-based, interpretative (only if transcribed) and static.

In theory, when dealing with corpus interoperability, there are two kinds of interoperability to achieve. Firstly, corpora of different origins need to be interoperable at annotation level. The challenge is to make an annotated corpus interoperable with another annotated corpus considering their static nature. There are two alternatives. If both annotation schemes are of a different type, one solution is to keep them both in different annotation layers, much in the way the NITE object model does (Carletta *et al.* 2003). If both annotations are of the same type, *e.g.* PoS, then it is a case where one annotation scheme needs to be selected and the other one superseded or converted. Secondly, interoperability can be seen from the point of view of dynamic resources such as software tools and the standards that are used to process and structure data. This is a technical aspect as the discussion refers to the formalisms, *i.e.* the standards, that are adopted in the creation of a data structure. Indeed, when the data are being structured, it follows rules that describe the data. In the case of interoperable linguistic tools, the formalisms must be understood by the programs and data structures must be organised according to common standards if the tools are to be able to communicate with each other. In this context, TEI-XML compliant structures provide a common framework for the representation of annotated corpora. In addition, linguistic formalisms must also be taken into account. (Sérasset *et al.* 2009, 11) describe the Resource Description Framework (RDF/RDFS) that introduces the notion of classes and subclasses which thereby allow the hierarchical organisation of linguistic knowledge.

In practice, making several corpora interoperable requires first an assessment of what can be done, which is the purpose of subsequent sections of this chapter. We

address the question of interoperable annotation by assessing tools and annotation types used to produce different types of annotation. Secondly, in Chapter 5, we show how we implement an interoperable annotation system between corpora. We also explain what kind of XML structure we apply to ensure the same formalism across several corpora.

4.1.3 Available tools to apply annotation schemes

One way of reaching interoperability is done through the structure that governs annotation layers. Another complementary method is to automate the processes. Indeed, automaticity brings consistency in processing the data. Depending on the type of annotation, there are several kinds of solutions that can be used. In this section, we only report on some NLP tools that could be used for PoS and error annotations. It must be recalled that other tools exist for other types of annotation. Section 4.1.3.1 shows how PoS annotation can be automated. Section 4.1.3.2 covers two different approaches in the handling of errors, *i.e.* manual and automatic detection.

4.1.3.1 Automatic PoS annotation

For PoS annotation, each annotation scheme is closely linked to an automatic tagger. This type of tool is used to assign word-class categories to tokens in a corpus. Automatic PoS taggers rely on algorithms that implement different methods that are either rule-based or probabilistic. The challenge of PoS taggers is to disambiguate the different word classes that could be applied to the same word. For instance, the word *can* can be interpreted as a noun or a modal verb. Assigning the right word class depends on a series of features found in the context. In the case of rule-based systems, already-tagged words are used to train the tagger that produces a list of word tag values. Ambiguous cases are then processed in a series of phases that apply rules to disambiguate them. In the case of the Brill tagger there is a rule that modifies the verb into a noun if one of the preceding words is tagged as a determiner (Brill 1992). This process of modification leads to the

assignment of a unique word class to each word of a corpus. Performance results show good levels of robustness as taggers such as Brill's fare close to 95% in overall accuracy with a 5.1% error rate.

Other taggers rely on probabilistic methods for the disambiguation process. The principle is to survey what is called a training sample of a tagged corpus and to compute metrics that are later used on a test corpus to assign tags. The computation of the metrics varies and it is outside the scope of this thesis. However, it can be said that the process relies on the calculation of probabilities of sequences of words and tags. In doing so, a tagger looks at what are called n-grams and statistically establishes the likelihood of specific sequences. Specific metrics such as information gain are measured in order to facilitate tag-assignment processes. One such approach in machine learning is the one adopted by the CLAWS disambiguation module. It consists in finding pairs of ambiguous tags in their close context and calculating their probabilities of occurring. During the training phase, the CHAINPROBS module (Garside *et al.* 1987, 39), as it is called, builds a matrix of probabilities of a tag occurring, given its immediately preceding "sister". Each pair of tags is assigned a probability which is later used to disambiguate tags in the tagging phase.

A second type of approach, also based on probabilities, involves the construction of decision trees following the ID3 learning algorithm (Quinlan 1986, 88). In the training phase, the algorithm builds a tree composed of branches leading to elements, such as PoS tags, depending on specific features such as previous PoS tags in the context. The most informative element—measured with Information Gain value—is positioned at the root of the tree and a branch is created for each possible subsequent element. Elements are tested statistically to evaluate how well one alone classifies the training examples. The criterion used to compare the tests is grounded in the measure of information gain defined by Mitchell (Mitchell 1997, 55) as "the expected reduction of entropy caused by partitioning the examples according to this attribute", and where entropy is "a measure that characterises the

(im)purity of an arbitrary collection of examples of some target concept". In other terms, potential tags are weighted in relation to their statistical significance in the assignment of tags in the training sample. TreeTagger (Schmid 1994) is a tool that partly applies this approach as far as building trees is concerned. It was developed at the Center for Information and Language Processing[16] and implements the Penn Treebank tagset. In terms of algorithm, TreeTagger (Schmid 1994, 15) is a probabilistic tagger. Its algorithm computes the probabilities of sequences of three tags, *i.e.* trigrams, but it also includes decision tree abilities by estimating transition probabilities between tags with the construction of a binary decision tree. The probability of each tag is based on a decision tree that includes a series of tests on the two preceding tags. Each test checks whether the path leads to the tag and if the tests are completed successfully, the frequency of the last tag of the tree is assigned to the trigram. This process allows for ungrammaticalities to be taken into account even though such sequences might record null frequencies. In conventional taggers only based on n-gram frequencies, a null frequency is problematic as the probability calculation cannot be performed, giving way to various adjustments that ignore the occurring reality of these sequences (Schmid 1994, 15). PoS tagging can include ungrammatical sequences, especially in the case of tagging learner English. This is one reason that makes TreeTagger a good candidate for learner corpus PoS annotation.

Another probabilistic approach, named instance-based learning, was also tested on tagging tasks. The algorithm memorises classes matched with sequences of features. Statistical calculations are performed to measure the distance that separates a new sequence of features from a k number of most similar and already-known sequences of features and to assign a class to the new sequence (Mitchell 1997, 232). The classifying process can be applied to PoS tagging, but not only, as the next section will show.

[16] http://www.cis.uni-muenchen.de/ (Last access March 31, 2016)

4.1.3.2 Handling learner error annotation

In this section, we present two ways of handling learner errors. The first one is mainly used in the domain of SLA and its community and annotating is mainly a manual process. The second one cannot strictly be assimilated to annotation but rather detection and it is based on automatic processes.

Regarding manual error annotation, there are no agreed annotation schemes as seen in Section 4.1.1.2. Through the building of various corpora, a variety of software tools have been developed to facilitate the work of annotation carried out by trained experts. For instance, ICLE was tagged with the Louvain tagging tool called *UCLEE tag editor* (Dagneaux, Denness, and Granger 1998, 167). It consists of a Graphical User Interface (GUI) made up of a main window with a menu bar. The text with errors, to read and annotate, appears in the window and the menu bar provides access for the selection of error types. Error types are classified and the menu makes the chosen tag accessible via a global to specific path. The selection of each error type triggers the insertion of an error tag into the text. This type of GUI facilitates the tagging process and allows for a quick retrieval of all instances of a particular error type. However, it is not without caveats as discussed in Section 4.2.2.2.

Another line of research on errors is that of automatic error detection. This line of research can also be seen as some sort of annotation as NLP tools are used to tag learner output as erroneous or not. A number of studies based on NLP technologies have opened new horizons in the handling of errors. Leacock reports on a series of studies carried out on the automated detection of articles and prepositions in learner language (Leacock 2010, 47). As difficult a task as it might be, they manage to group studies in three categories. They report the use of automated systems, among which, the most recent ones are statistical classifiers. The classifiers process four types of information. Firstly, they might rely on source-language information to detect erroneous uses of articles or prepositions in English for instance. The second group of systems corresponds to those based on token context. Classifiers

are used to operate on the surrounding contexts of the forms to elaborate metrics, such as information gain, to learn and then classify new occurrences. The third group of systems corresponds to those based on syntactic context information such as PoS and parsed annotation. Finally, semantic information might also be used in order to train classifiers. One such study was conducted by (Pradhan *et al.* 2010) on the automatic classification of article errors in L2 English. Researchers used an error-annotated and PoS-tagged learner corpus to train several classifiers on the basis of PoS tag sequences. The objective was to predict the occurrence of error in article usage based on context features made of the PoS tags. The best accuracy level reached was 70%. This project was only one of many that sprouted in the last decade. See Na-Rae Han *et al.* (2006) for the same detection of article errors but based on context information and trained on native English. See also Leacock (2010, 52-54) for a comprehensive overview of studies on the matter.

4.2 Annotation for *this* and *that*

In this section, we cover the annotation framework which needs to be put in place to annotate *it*, *this* and *that* in their contexts on three specific corpora. After examining annotation schemes from a global point of view, *i.e.* their design principles, in Section 4.1, we take a closer look at the way *it*, *this* and *that* need to be handled in terms of annotation. We describe the corpora used in subsequent experiments. We also show the need to implement a multi-layer annotation scheme, including the fact that the PoS annotation layer requires finer-grained labels to support our research on the distinction between the pro-form and determiner functions. (see Chapters 6 and 7). In Section 4.2.1, we present the three corpora which are used in our annotation setup. Section 4.2.2 covers and assesses the various types of annotation which may be applied on the forms. In Section 4.2.3, we discuss the requirements of the annotation setup that will be implemented in the next chapter.

4.2.1 The corpora

In this section, we describe the three corpora with which the experiments are carried out (Annex B gives a synthetic view of the specifications of the corpora). In reference to learner corpus research methodologies covered in Section 3.1.2.2, it is necessary to combine native and non-native corpora in order to make comparisons and measure differences and similarities between learners and natives. The introduction of several learner corpora in the setup is also important as it will allow a comparative approach between learners of different L1s. L1-specific strategies such as transfers might be isolated from general learner patterns giving clues on stages of acquisition of English as an L2. Overall, three corpora are used.

The *Wall Street Journal* subset of the Penn Treebank corpus (Marcus, Marcinkiewicz, and Santorini 1993), as far as native English is concerned, is the corpus we chose for comparisons between NS and NNS. There are multiple reasons for this choice. First of all, it is a large subset that consists of one million words of American English and which includes press articles published in the *Wall Street Journal* in the early nineties. As presented in Section 4.1.1.1, it is PoS tagged and parsed in a tree-like manner and verified manually, which makes it a gold standard. It is necessary to mention at this point that only the PoS tag annotation is used for the annotation setup of our work. The second reason for choosing this corpus is that it is PoS annotated with the tagset that is also used for the automatic annotation of the learner corpora, making them interoperable. It is worth mentioning that the Penn Treebank PoS tagset also reflects a militant approach for the significance accorded to syntactic context because words' syntactic functions are encoded in PoS labels whenever possible (Marcus, Marcinkiewicz, and Santorini 1993, 316). In spirit, this approach matches the fine-grained distinction that we intend to introduce in the tagset for the tagging of the Diderot-LONGDALE corpus (see below for the description of this corpus). Finally, the fact that this corpus represents only one genre provides an answer to the problem of corpus heterogeneity and the variations it entails. It can be argued that press articles do not constitute a perfect base for comparison with learner English. However, the fact

that the corpus is composed of one genre, which incidentally happens to be the prototypical target language in L2 classes, makes it a stable point of comparison. Barlow points out that "the combination of genres in the general corpus does not provide a good reference point for the learner corpus, which invariably consists of a single genre" (Barlow 2005, 345). The advantage of a single genre NS corpus is that observations of variations or discrepancies, resulting from NS and NNS comparisons, will not find their source in genre variations. Instead, they can be interpreted as interlanguage variations (see Section 3.1.1) even though learner language is far from being identical to that of the Wall Street Journal.

The second corpus to be used is the NOCE corpus developed by Díaz-Negrillo (2007) at the university of Granada. The samples were collected between 2003 and 2004 from first year Spanish students of two compulsory English classes for specialists of English. It is composed of 39,015 words (Díaz-Negrillo 2007, 16) that are part of handwritten texts which were later typed. The texts correspond to one of four thirty-minute tasks that consisted in writing a descriptive, narrative, argumentative or free-writing essay (Díaz-Negrillo 2007, 8). The corpus was annotated with a specifically designed error tagset described in Section 4.1.1.2 of this chapter. In addition to error annotation, the corpus was also PoS annotated with the Penn Treebank tagset. Two annotation layers can be extracted from this corpus. The file structure is described in (Díaz-Negrillo 2007, 8-9) and each file name gives a precise identification of each sample of the corpus. In our study, we use a subset which has been specifically annotated and manually checked. It is made up of 13,143 tokens.

The third corpus to be used is the Diderot subset of the LONGDALE corpus (Meunier *et al.* 2008) collected at the University of Paris-Diderot. It is a longitudinal corpus of spoken English collected from the same students of L1, L2 and L3 levels. Recordings were carried out in waves at several points in time: October 2009, June and October 2010 and March 2013. For each wave, a total number of 22 recordings were made, resulting in a corpus of 66 recordings. The

length of time of each recording varies from around 2 minutes to over 10 minutes. For our study, we had to carry out annotation which needed to be manually checked. Consequently, we worked on a subset of the corpus made up of 36 transcribed recordings totalling 64,858 tokens. The recordings were transcribed by Master's students in Linguistics from the University of Paris-Diderot. Proofreading, coupled with listening, was carried out on the transcriptions by Professors and PhD students of the CLILLAC-ARP research team. Transcriptions adhere to the LINDSEI guidelines[17] presented in Annex C. The transcriptions include meta-data in the form of codes on such elements as the interview identification, speakers' turns or overlapping speech. There are external meta-data such as age and gender.

Each wave of recordings is dedicated to two types of task. The first type is characterised by spontaneous production of free speech captured in conversations and monologues. The second type of task is a reading of an extract from literature or from a political speech. Annex D gives the full description of each of the tasks for each recording. However, as an example for illustration, it can be mentioned that the first task of the first wave of recordings was about discussing one of the following topics and, after a first monologue-type part, the native language assistant asked questions to create interaction:

- Topic 1: A good or a bad experience
- Topic 2: A country you visited
- Topic 3: A film, a play, a sport event, a concert you saw

The second task of the first wave involved the reading of an extract from Oscar Wilde's *The Selfish Giant*. This activity was considered unplanned since one minute preparation was given to the student.

In terms of file structure, the recordings were saved in the WAV format, which corresponds to raw and uncompressed audio. The transcriptions were saved in file names adhering to a strict naming policy in order to ensure precise identification of

[17] http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Lindsei/transnew.htm (Last accessed March 31, 2016)

the recording. The file-name structure is made up of three parts separated by a hyphen and corresponds to the following pattern:

```
DID0000-S00X(-R)
```

A number of comments must be made to explain the meaning of the components. The first part is made up of two components. The DID component corresponds to the identification of the Paris-Diderot subset of the LONGDALE project because there are other subsets from other universities in Europe[18]. The second one indicates the student's unique identifier. The second part after the hyphen is also made up of two components. The S indicates the Spoken mode[19] and the 00x identifier determines whether the recording corresponds to the first, second or third wave. The last optional part, between brackets, indicates the reading task to be found in the recording.

One important characteristic of the corpus is that it is longitudinal as it includes audio recordings—and their transcriptions—of learners over a three-year period. To the best of our knowledge, this makes it the first learner corpus of its kind ever compiled. The Diderot subset will be made available as part of the LONGDALE corpus. So far, two full cohorts of data have been collected over a six-year period between 2008 and 2012.

The corpus is not annotated. Nevertheless, it is worth mentioning that unofficial and experimental PoS and parsed annotation has been achieved with CESAX (see page 134) by Erwin Komen[20] from the University of Radboud. Komen's implementation of an annotation scheme on the Diderot-LONGDALE corpus relies on the Penn Treebank annotation scheme. As we shall see, we also use the Penn

[18] For a full listing and details on the project: http://www.uclouvain.be/en-cecl-longdale.html. See also (Goutéraux 2013) (Last access March 31, 2016)

[19] We use the term 'mode' to refer to Biber *et al.*'s 'physical mode'. In their work, they posit that 'registers' have a number of situational characteristics, including 'physical mode' which can be 'oral' or 'written' (Biber *et al.* 1999, 15). In our analysis, this notion of 'mode' provides a neutral variable which applies to all our corpora. Using this concept adds an extra paradigm to help us compare corpora.

[20] See the LONGDALE files used in the 2014 Nijmegen workshop: http://erwinkomen.ruhosting.nl/eng/ (Last access March 31, 2016)

Treebank to PoS tag the Diderot-LONGDALE. Our annotation setup differs from Komen's insofar as different types of annotation are used, including a modification of the Penn Treebank so as to provide finer-grained functional tags for *it*, *this* and *that*.

4.2.2 What kind of annotation for *this* and *that*

As a reminder, the problem we need to address is to search and extract occurrences of *it, this* and *that* according to different levels of interpretation in several corpora. There are different interpretation levels of the pro-form microsystem—modelled in Section 3.3—such as function, position, context and discourse. As mentioned before, these levels of interpretation convert into a diversity of annotation types. In this section, we assess the particular types of annotation that could be introduced in the annotation setup used for the corpora. We also decide whether each type is going to be implemented or not. In Section 4.2.2.1, we deal with the PoS-functional annotation layer that needs to be introduced and adapted to *this* and *that*. In Section 4.2.2.2, we discuss the relevance of an error annotation layer on *this* and *that* in our setup. Section 4.2.2.3 covers the annotation layer which deals with the syntactic position of the forms in context. In Section 4.2.2.4, contextual annotation is discussed in light of the types of endophoric or exophoric contexts that are attributable to each occurrence of the forms. In Section 4.2.2.5, we finish with discourse level information, *i.e.* the information givenness status of the referential expressions.

4.2.2.1 PoS-functional annotation

Considering the functional approach of our analysis, it has been shown that *this* and *that* operate in two microsystems that are grounded in a functional distinction (see Section 3.2.2.2). The determiner or pro-form function determines which kind of forms *this* and *that* compete with in learner productions. Consequently, the PoS tags must reflect this distinction and this is the reason why we use the term PoS-functional annotation. In this section, we examine whether the distinction appears

in various PoS annotation schemes and we show the need to introduce a finer-grained tagset.

Table 10 shows how the forms are tagged by several tagsets. The annotation process of *this* and *that* also leads to the side exploration of other realisations of *that*. The table is to be read horizontally. For instance, the first line corresponds to *this* as a determiner. It receives the *DT* tag with the Penn Treebank tagset. It receives the *DD1/2* tag with CLAWS7. However, it does not receive a determiner label with TOSCA/ICE as it is dealt with as a pronoun. *That* as a pro-form is still tagged as a determiner in the Penn Treebank and CLAWS7. Conversely, the TOSCA/ICE tagset clearly labels the pronoun *that* but it uses the same label for determiner *that*.

If we observe the data in Table 10 in relation to *this* and *that,* and the tag accuracy to describe them, there are recurrent occurrences which may be non-distinctive. For example, the fact that the pro-form and the determiner functions receive the same tag (*DT*) in the Penn Treebank tagset shows that automated PoS tagging would not allow in-depth research on the issue. The same applies to the CLAWS7 and TOSCA/ICE tagsets where tags are used identically for different functions (*i.e. DD1/2*).

	Functions	**Penn Treebank native POS tags**	**CLAWS7 native POS tags**	**TOSCA/ICE native POS tags**
this	Determiner	DT*	DD1/2	PRON(dem,sing)
this	Pro-form	DT**	DD1/2	PRON(dem,sing)
that	Determiner	DT*	DD1/2	PRON(dem,sing)
that	Pro-form	DT**	DD1/2	PRON(dem,sing)
that	complementiser	IN	CST	CONJUNC(subord)
that	Relative pronoun	WDT	CJT	PRON(rel)
that	Adverbial	RB	RG	ADV(inten)

* This category includes the articles *a(n)*, *every*, *no* and *the*, the indefinite determiners *another*, *any* and *some*, *each*, *either* (as in *either way*), *neither* (as in *neither decision*), *that*, *these*, *this* and *those.*

*** When determiners are used pronominally, i.e. without a head noun, they should still be tagged as Determiners (DT) - not as common nouns (NN), e.g. I can't stand this/DT.*

Table 10: This and *that* in tagsets

As it appears, functional distinction is an issue. The pro-form and determiner functions of *this* and *that* do not appear in the predetermined tagsets used in automated tagging processes. In the current situation, no scheme is satisfactory as far as the functional distinction is concerned. Consequently, the choice of a scheme must be made on the basis of the possibility of modifying the relevant tags in the tagsets. This raises the question of which tagset is modifiable and how.

We choose the Penn Treebank as it is implemented in TreeTagger (Schmid 1994), the open source software tool, already mentioned, and with which the tagset can be modified (see Section 5.1.1 for details on the procedure). The tagset is very well documented (Taylor, Marcus, and Santorini 2003) and already includes most of the distinctions for *that* even though it lacks accuracy for the determiner and pro-form functions. For the annotation setup is possible to modify the *DT* label (see Annex E for a description of the entire tagset) that is used in both cases to create two new labels: *DT* to indicate the determiner function, *TPRON* to indicate the pro-form function. To conclude, the choice of the Penn Treebank PoS tagset is justified by

the fact that it is modifiable to obtain finer-grained tags for the specification of the functional distinction of *this* and *that*.

Another advantage of this tagset is that it can also be implemented automatically with TreeTagger. This raises the question of reliability. In spite of errors, PoS-tagging provides good levels of accuracy on native English. On learner corpora, it has been shown that accuracy deteriorates (Gaillat 2013c). Our approach is to use automatic PoS tagging on all types of corpora and to post-edit the learner corpora. As it is costly in terms of labour, only the tags related to *it*, *this* and *that* are manually verified and corrected.

4.2.2.2 Error annotation

Error annotation is a field that has been under close scrutiny over the past two decades as computerised learner corpora emerged. As mentioned in Section 4.1.1.2, there are two types of manual annotation schemes (general or fine-grained) that have been developed over the period. Table 11 shows two examples of such types: the Louvain tagset and EARS tagset. Both annotation schemes differ in the granularity of their tagsets but they both rely on manual processes of annotation. The table shows what tags are used for the annotation of errors made on *this* or *that*.

	Functions	**NOCE/EARS error tags**	**Louvain error tags***
this	Determiner	WG.PO.DM...	GP
this	Pro-form	WG.PO.DM...	GP
that	Determiner	WG.PO.DM...	GP
that	Pro-form	WG.PO.DM...	GP
that	complementiser	CG.PC.SD.IT...	LCS OR XCONJCO
that	Relative pronoun	WG. PO.RL.CA...	GP
that	Adverbial	WG.AV.DG.SV.ER.MS	GADV OR GADVO OR LS

* In the latest version of the error tagset a distinction is made between pronouns and determiners, as well as between different types of pronouns and determiners (demonstrative, relative, etc.)

Table 11: *This* and *that* in two error annotation tagsets

The table reads horizontally and each form is matched with a possible function it can take. Depending on the function, each form is assigned a particular error tag which is more or less complex depending on the tagset.

In the case of the EARS tagset, errors are specified from a more generic to a more specific path as described in Section 4.1.1.2. Errors on *this* and *that* can be marked differently depending on their grammar category, on their syntactic category, their internal or external nature (whether the learner ill-formed or misused the form), and the surface structure realisation such as misselection, omission or overinclusion.

As far as the pro-form/determiner distinction for pronouns is concerned, the taxonomy does not take it into account. Díaz-Negrillo (2007, 85) specifies that "the category Pronoun, available in Word Grammar and Lexis, groups units [...] may take the function head of the noun phrase and also determiner". Errors may be tagged with different paths with a common initial part. The following paths provide details for the classification of pronoun/determiner-related errors and other types to which the forms might correspond. Indentations show forks in the paths.

For pronouns and determiners (Díaz-Negrillo 2007, 268-275)

WG.PO.DM... - WG Word Grammar; PO pronoun; DM demonstrative
PX Proximity; DL Proximal or RX proximal
NB number
LU Plural
SG Singular
IT Internal
ER External
MS Misselection

For *that* complementiser:

CG.PC.SD.IT... - Clause Grammar, PC Processes; SD Subordination; IT Internal
OM Omission
OV Overinclusion

For *that* relative pronoun:

WG. PO.RL.CA... - Word Grammar; PO Pronoun; RL Relative; CA Case
AC Accusative
NT Nominative
ER External
IT Internal
MS Misselection

For *this* or *that* adverbials

WG.AV.DG.SV.ER.MS – WG Word Grammar; AV Adverb; DG Degree; SV Positive; ER External; MS Misselection

For the sake of illustration, the following examples of error tagging on occurrences of *this/these* can be found in (Díaz Negrillo 2007, 148)

36) [...] <WG.PO.DM.NB.LU.ER.MS>these</WG.PO.DM.NB.LU.ER.MS> type of films. GR-1-A-EN-001-F

37) I really like <WG.PO.DM.NB.LU.ER.MS>these</WG.PO.DM.NB.LU.ER.MS> life. GR-1-A-EN-017-F

38) […] the answers for <WG.PO.DM.NB.SG.ER.MS>this</WG.PO.DM.NB.SG.ER.MS> questions could be summarised [...] GR-1-C-EN-045-Z

39) [...] the mental health of <WG.PO.DM.NB.SG.ER.MS>this</WG.PO.DM.NB.SG.ER.MS> people is not very good. <ICLE-SP-UCM-0036.4>

As already said, the Louvain tagset is generic in its approach to errors. Under the Louvain taxonomy *this* and *that* can be assigned the same non-distinctive label for their pro-form or determiner functions and even the relative pronoun functions. The *GP* label, for Grammar, Pronoun is used as a path to point out the type of error. In its complementiser function, depending on the error context, *that* could be assigned the LCS or XCONJCO labels for Lexis, Conjunctions, Subordinating and LeXico-Grammar, Conjunctions, Complementation, respectively. As adverbials, the forms can be classified with GADV(O) for Grammar, Adverbs, Order or LS or lexical, Single.

So the two tagsets describe two manual approaches in the description of errors that may be made on the forms. Their level of granularity is intended to encompass all possible errors that may be found in a corpus. The advantage is that of qualitative accuracy since all forms are manually tagged by trained experts and this ensures a high level of correctness in the chosen tags. Manual error annotation has appeared to be the most robust line of conduct when it comes to learner language given its high variability. So far, it has appeared obvious that it could only be achieved manually. However, manual error-tagging is not without flaws as Dagneaux *et al.* put it:

> "The problem is compounded by the error categories used which also suffer from a number of weaknesses: they are often ill-defined, rest on hybrid criteria and involve a high degree of subjectivity. Terms such as "grammatical errors" or "lexical errors", for instance, are rarely defined, which makes results difficult to interpret, as several error types—prepositional errors for instance—fall somewhere in between and it is usually impossible to know in which of the two categories they have been counted. In addition, the error typologies often mix two levels of analysis: description and explanation." (Dagneaux, Denness, and Granger 1998, 164)

As clearly expressed, manual error annotation can lead to differences in the interpretation of certain categories to be assigned to certain learner errors. Consequently, inter-rater agreement between annotators of the same texts might be

low enough as annotations on errors vary drastically (Leacock 2010, 98). Hence, should a tagging error be discovered for a given occurrence in a corpus, it is not possible to systematise its detection in the entire corpus. In addition, manual tagging is a labour-intensive solution that requires a team of experts to be set up, which is not the kind of resource we have. This is the reason why a solution based on automation needs to be found to handle learner errors on *this* and *that*.

Our solution is to avoid the error annotation process *per se* and to focus on error detection. The principle is to use neutrally robust annotation on texts to train tools which can then be used to detect errors in other annotated texts. In this context, Díaz-Negrillo *et al.*'s (2010) proposal of a tripartite interlanguage PoS annotation for learner corpora could be seen as support for better error detection. They note that traditional PoS tags for native English are the point of convergence of three types of evidence which are distribution, stem and morphology. These types could be converted into three levels of annotation. Each level could be used as a criterion by tools that could automatically analyse the consistency of the forms and reject forms whose distribution, stem and morphology are not consistent with each other.

To understand how this tripartite PoS annotation system can work with *this* and *that*, let us consider the following utterance from the Diderot-LONGDALE corpus.

40) it’s more cool there's so different and (er) out of the language it’s very I prefer the American actually but that special accent *that*’s very complicated - DID0118-S001

The error on *that* as a pro-form refers to the microsystem described in Section 3.2.2.2.1. It corresponds to a substitution with the pronoun *it*. Since this error is related to the subject position of the form in the sentence, it is relevant to consider it as distributional evidence. Another element which helps identify the error is its functional realisation as a pro-form. So, it is precisely when the form is found in the pro-form function as a subject that learners may substitute *that* with *it*. As shown previously, native PoS schemes, including the Penn Treebank, only apply one type

of PoS for the determiner and pro-form functions. Therefore, a finer-grained annotation is required for *this* and *that* to link the error to distribution.

Another example shows how the distribution annotation, based on functional realisation, helps detect another error with *this*. Example 41 from a learner shows a case of distribution—morphology incompatibility.

41) we didn't (er) seen we haven't seen the car that (er) was a= arriving by (er) our left so the car (er) (em) shooted you (er) us . shooted us and (em) and (er) I didn't I don't remember (er) the the next ten minutes about *this* events . DID0080-S001

In this case, the functional realisation as determiner unambiguously links the determiner to the following noun and thus shows the incompatibility between the singular form (morphological feature) of the determiner and the plural form of the *events* noun. The functional distinction is a useful piece of evidence to identify the error, which supports the need to introduce it in a tagset. It helps detect the incompatibility between distribution evidence and morphology.

In both examples, the detection of errors relies on one or two types of distributional information. The functional distinction (Section 4.2.2.1) and the position (see Section 4.2.2.3 for details of its requirements) help identify whether learner occurrences are coherent or not. The introduction of such information in a distribution type of annotation could be a step in the direction of Díaz-Negrillo *et al.*'s (2010) tripartite PoS annotation. Functional realisation and position could be considered as features of *this* and *that* within the distribution annotation layer of a scheme. Adding a determiner/pro-form distinction as well as a nominative/oblique distinction in the distribution annotation type provides neutral linguistic evidence which can subsequently be used for automatic analysis. Our assumption is that this information, combined with other PoS information from the surrounding context, can pave the way for the automatic identification of errors.

Automated error detection must rely on tools that help compare evidence collected about the forms with previously validated features of the same forms. In Section 7.2, we report on an experiment carried out on a sample of the Diderot-LONGDALE corpus. We show that by collecting distributional evidence of a set of forms in learner output, it is possible to compare them with previously collected correct forms and deduce whether they are erroneous or not.

For error detection, a classifier—a tool that belongs to the machine learning domain—called TiMBL (Daelemans *et al.* 2010) can be used. TiMBL is a Memory-Based Learner (MBL hereafter) developed at the University of Tilburg by the ILK research group[21]. The tool memorises lists of instances and their matches with classes before applying a classification process to new instances. The tool is launched with a command line in which the program is called together with a training file and a test file. Options can also be passed on to determine the algorithm and metrics to be used. TiMBL's principle is to store many instances of a phenomenon in its memory. When a new instance needs to be classified, it is compared to those already stocked in memory. Like all Memory-Based Learners, TiMBL is made up of two modules. The first one is designed for the training phase and the second one is in charge of the actual performance, that is the classifying task. An instance is represented by an array of features that correspond to a specific class. The training phase consists in collecting a high number of arrays and their respective classes. In the classifying phase, a new instance is presented to the classifier and its similarity with all the instances in memory is computed. TiMBL implements several entropy-based metrics such as information gain ratio to measure the distance that separates an instance from the k-nearest neighbour examples in memory. In the case of error detection on *this* and *that*, we convert the contexts found around each occurrence into sequences of features based on words, PoS and context information. Each array is assigned either a correct or error class. The classifier can then be used to classify new occurrences, hence performing an

[21] https://ilk.uvt.nl/timbl/ (Last access March 31, 2016)

error detection task. An experiment on such a process will be documented in Section 7.2.

4.2.2.3 Syntactic position annotation

As seen in Chapter 2, learner output seems to be specific as far as pro-forms are concerned. Learners tend to have difficulties to differentiate between *this, that* and *it* in their pro-form function and the positional criterion is to be taken into account. Depending on its position as object or subject, the pro-form might be used differently by speakers of different L1s. Consequently, specific positional annotation must be added in order to provide information on nominative and oblique cases. We define nominative cases as those syntactic configurations where the form is the subject of the verb. The other argument of the verb can correspond to obliques or objects: "With NPs, we have a clear distinction between object (related directly to the verb) and an oblique (related via a preposition)" (Huddleston and Pullum 2002, 1207). However, further analysis on infinitive and content clauses in their work shows that the distinction is not always easy. Therefore, a binary distinction between nominative (NOMI) and non-nominative cases (OBLI) might appear as sufficient for the analysis of the pro-form microsystem.

Another requirement for the syntactic position annotation is that it must be carried out automatically. Due to the deterministic nature of the process, it is possible to set up rules that rely on the close environment of the forms for the assignment of the labels. An algorithm is developed to carry out this task and it is presented in Section 5.1.2.

4.2.2.4 Context annotation

In the previous chapter of this thesis, we described semantic linguistic features to take into account in the analysis of *this* and *that* in corpora (see Section 3.3.1). Endophoric and exophoric characteristics of context represent two contextual elements that participate in the identification of referential processes. We call them

context features as opposed to discourse features such as the deictic anaphoric distinction performed at discourse level. In terms of corpus annotation, these features should ideally be converted into some kind of labels.

The context annotation is introduced for the exophoric and endophoric values of *this, that* and *it*. This type of annotation is applied to words as it involves the identification of the senses of the word forms. This kind of annotation can be seen as a form of word sense disambiguation. Automating this process may be envisaged with a classifying algorithm but this would require a series of programming operations which have not been developed in the course of this research. Consequently, manual annotation is carried out for the learner and native corpora used in this study.

The manual tagging of the corpus is carried out after the PoS tagging phase performed with TreeTagger. The output files lay out the data in a two-column format as seen in Section 4.2.2.1. The endophoric and exophoric tags are added in a third column for each *this*, *that*, and *it* form. Two tags are chosen: *ENDO* and *EXO*. Figure 16 shows how the token *this* has been functionally tagged *DT* in the Penn Treebank and then manually tagged as exophoric. All corpus data files are fashioned in this manner as a result of the annotation-oriented stage of our setup. In Chapter 5, we show how the files are further handled to prepare the data for analysis.

They	PRP	
make	VBP	
the	DT	
argument	NN	
in	IN	
letters	NNS	
to	TO	
the	DT	
agency	NN	
about	IN	
rule	NN	
changes	NNS	
proposed	VBD	
this	DT	EXO
past	JJ	
summer	NN	
that	TREL	
,	,	
among	IN	
other	JJ	
things	NNS	
,	,	
would	MD	
exempt	VB	
many	JJ	
middle-management		JJ
executives	NNS	
from	IN	
reporting	VBG	
trades	NNS	
in	IN	
their	PRP$	
own	JJ	
companies	NNS	
'	POS	
shares	NNS	
.	.	

Figure 16: WSJ sample after PoS-functional and context annotation processes

Section 5.2 shows how the output files are then integrated within a broader XML structure and how they are converted into sequences of linguistic features for distribution analysis.

As it is a manually operated annotation process, rules must be laid out to describe how to apply context annotation in line with Burnard's recommendations on the elaboration of corpus metadata (Burnard 2005, Section 3). As a preliminary remark, it is important to state the fact that we annotate the language as it is and not as it is meant by the speakers. In other terms, if speakers commit errors and

thus produce specific referential meaning effects, we do not try to annotate what they actually mean but we annotate the word as it is and its function as it is.

For endophoric and exophoric contexts, we consider whether the referent's accessibility is performed via the text, *i.e.* the spoken speech or the stream of written words or the situation, *i.e.* an object/date/place present in the situation of utterance. In case of a referent present as an object in the communication situation (which is the case in some utterances of the Diderot-LONGDALE), we consider it exophoric in all its instances even though it is arguable that after the first mention, it is part of the speech and so has become endophoric. However, the fact is that sometimes full accessibility of the referent is clearly exercised via the visual channel. The fact is that situational details might be necessary to understand what is meant exactly. In recordings S003 of the Diderot-LONGDALE corpus there is a question on several paintings that are presented to the students. They give their comments and refer to the painting in their hands with *it*. Sometimes, the full interpretation of their comments can only be achieved thanks to details seen in the situation of communication. This advocates for the tagging of *it* as exophoric even though it also has an endophoric value. See the following commented examples from the Diderot-LONGDALE contextual annotation:

42) “so do you just take the pictures for yourself or do you show them to </A><B>oh I don't show *it*” (*PRP, ENDO*) - DID014-S002

In this example, *it* is endophoric because access to the referent is done via the text, hence the repetition of the verb show.

43) I like to think what's goin=what's happened in my body what happened in my head what's happened and .. I I like *those* (*DT, ENDO*) things actually so - DID014-S002

This reference is endophoric again because it is made to elements previously mentioned in the utterance.

44) (Speaker A is showing a picture to Speaker B)

A. And does *it* (*PRP, EXO*) give you any emotional response

B. Well I think *it* (*PRP, EXO*). *it* brings peoples back to their childhood. DID014-S003

The first *it* is marked as exophoric as the reference is made via the situation. The second one is endophoric as speaker B most likely refers to the object placed on the table.

45) DID014-S003 <B> […] and besides the :: the first and the second one I'm the one who liked the painting but on the third I'm the one to be forced to like the painting <laughs> </B> <A> interesting and what do you think about the painting that makes *that* (DT, ENDO) difference</A>

The determiner *that* is endophoric because speaker A refers to a distinction made by speaker B in his speech.

4.2.2.5 Discourse annotation

The discourse type of annotation could determine the givenness of discourse entities. As opposed to the PoS-functional annotation described in Section 4.2.2.1, it must be recognised that it is so semantically complex that the automation of such processes seems to be a difficult task. As shown in Section 2.2.1, most experiments intend to identify coreferential items at text level and few focus on the classification of indexicals in relation to their value at discourse level. Our theoretical background relies on a discourse-functional approach of the problem. Consequently, it is important that the codings chosen for the discourse annotation type include information related to Cornish, Kleiber and Ariel's central paradigm of topic information with its *given* or *new* status. In this respect, a solution might be envisaged with CESAX (Komen 2012). However, compatibility with our analysis is an issue. CESAX partly implements coreferential link annotation and antecedent identification, which is not our focus. It also uses a standard version of the Penn Treebank tagset, which means that there is no distinction between pro-forms and determiners for *this* and *that*. In spite of these issues, the following paragraphs are

devoted to a brief assessment of this tool, because it looks very promising in its features and applications. Its technical functionalities have to be assessed for the automatic analysis of coreferentiality.

In this paper, the author reports a semi-automatic method to assist with the linguistic annotation process prior to data analysis. As seen in Section 4.1.1.1, CESAX's annotation scheme supports the introduction of referential annotation features that match the discourse-functional view. What is more, the CESAX tool supports the automated assignment of tags defining several types of cognitive status of information. It automatically scans NPs in a corpus text and marks them according to 5 classes of referents: *new*, *assumed*, *identity*, *inferred* and *inert* (see Section 4.1.1.1). The way it does this is grounded in grammar features assigned to NPs in the parsed corpus.

This classification has been used in an experiment to provide support for coreference resolution (Komen 2012, Section 2.3). As far as performance is concerned (Section 2.4 of the publication), a 28% error rate is reported for the automatic coreference resolution process, which might be viewed as sufficient if we consider the tool as an assistant to the linguist involved in the annotation process of a corpus. This method can drastically accelerate the manual processing of a corpus. Nevertheless, it must be stated that coreferential resolution is the second phase of a previous one whose purpose is to carry out the classification of NPs according to the aforementioned classes. It is precisely this first phase that is of interest for the purpose of the discourse-functional annotation. It is therefore relevant to explore how well such a discourse annotation process can be carried out.

In a series of experiments on referential state prediction, and more specifically the second one, Komen (2013) shows that a memory-based learning approach with a classifier yields 92.5% performance (F-Score) in assigning information topic classes to NPs in a corpus. The classifier is trained on a corpus of English from which 27 features, such as grammar role or negation (for a detailed listing, see Table 1 in the

aforementioned article), are used to predict whether a particular entity in an NP is new, linked with a previously mentioned entity (named *link*) or inert (see Section 4.1.1.1 on CESAX's annotation scheme for a description of this feature). This experiment uses three classes as Komen encapsulates the *assumed*, *identity* and *inferred* classes into a unique *link* class. Detailed results show variations in performances between the features. The *link* feature fares higher than, in descending order, the *new* feature and the *inert* one (F-scores are respectively: 90.2%, 82.3% and 74.1%).

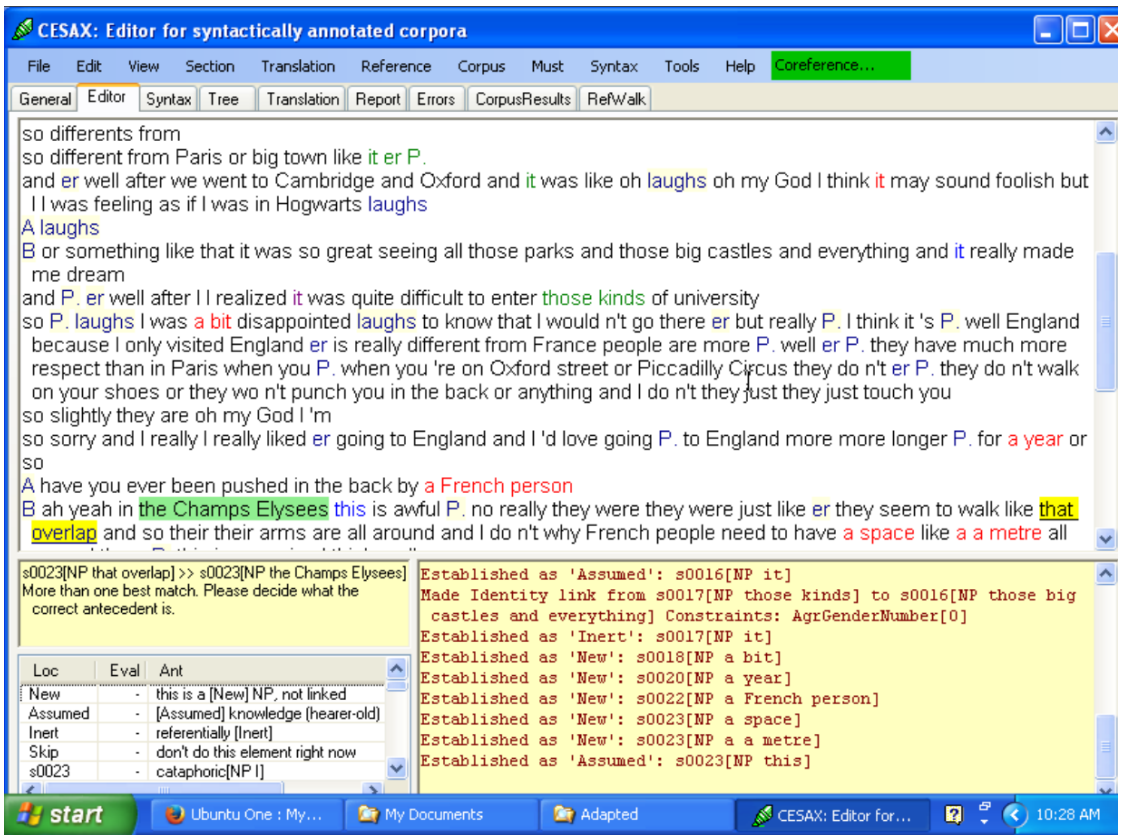

Figure 17: CESAX's editor view during semi-automatic coreference resolution process

CESAX implements this type of classification method. The fully-automated version sets reference types for all NPs whilst the semi-automatic version scans each NP. In case of several possible coreferential links, the program stops to let the user decide. In doing so, the user triggers the assignment of the reference type as well as the textual antecedent if it exists. The bottom right window, in pale yellow colour in Figure 17, shows the list of referent status assignments performed during a semi-automatic use of the tool. In order to assess the tool for further use, one of the

Diderot-LONGDALE corpus file is successively fully-automatically and semi-automatically annotated (Filename: DID0108-S001). The objective is to provide an introductory insight into what discourse annotation looks like with CESAX in order to give a first assessment on the feasibility of deploying or merging the solution with our fine-grained functional annotation of *this* and *that*. As time is a limited resource, the test is carried out on one file only in order to measure the necessary time of labour to process it. Another purpose is to see how the automation performs. In order to accelerate the process, and in compliance with the need to explore the microsystem of reference that learners exploit (see Section 3.2.2.2), only *it*, *this* and *that* are targeted for tagging. To assess both annotation processes, a gold standard is manually created in which all the *it*, *this* and *that* determiners and pro-forms are manually classified with the same categories as the tool uses (see Section 4.1.1.1 page 139). It must be noted that the *inferred* class is not available in the installed version CESAX. We also consider the *identity* and *assumed* categories interchangeable due to the fact that they refer to something already known in the discourse.

To assess the fully automatic process, the results are compared to the gold standard. All together, 22 determiners or pro-forms are detected by the system. The results are presented in the series of confusion matrices (see Tables 12, 13 and 14) for each relevant form.

		Automated annotation			
		Assumed / identity	New	Inert	unmarked
Gold standard	Assumed or identity	4			
	New				
	Inert				

Table 12: Confusion matrix for tagged forms of *this* with CESAX

		Automated annotation			
		Assumed / identity	New	Inert	unmarked
Gold standard	Assumed or identity	0			1
	New		1		
	Inert				

Table 13: Confusion matrix for tagged forms of *that* with CESAX

		Automated annotation			
		Assumed / identity	New	Inert	unmarked
Gold standard	Assumed or identity	11			
	New	2	1		
	Inert	1		1	

Table 14: Confusion matrix for tagged forms of *it* with CESAX

To interpret these results, it is important to recall that *assumed* and *identity* states are used interchangeably in the gold standard file. It appears that referential states are selected with excellent accuracy for *this*. *It* and *that* appear to cause difficulties as some of the forms are wrongly tagged or even not tagged at all. However, the sample is too small to be representative, further testing would be required to have significant results. Should it be conclusive, the deployment of the fully automatic solution could be envisaged.

As far as the semi-automatic process is concerned, its principle relies on the possibility to automatically scan the text, to stop at each NP candidate and to manually select from a list of reference candidates. As we are only interested in three forms, we skip all the other forms of NPs. For each form a type of reference can be quickly selected by pressing a key (A for assumed, N for new for instance). It is important to mention that, again, we use the *assumed* and *identity* tags interchangeably. *Identity* is in some cases quickly identifiable in the candidate list and thus selected. When candidate selection requires a search in the list of possibilities, the *assumed* state is preferred.

The file is processed within 15 minutes (on a PC with a Core i3 CPU and 4Gb of memory) and an XML file is generated, which includes newly added reference-related feature sets as in Figure 13 page 138. After proofreading, a few errors are reported. 3 *assumed it* forms are reported to be unmarked and one *that* form is missed due to its being tagged as a complementiser and so not part of an NP. *Those* forms are also missed and thus corrected with the *assumed* status.

A number of issues are encountered in the semi-automatic process. Firstly, the performance of the process depends on the quality of parsing and parsed errors lead to misselection of forms. For example, markup tags such as *<laughs>*, *<overlap>* and *<P>* are present within texts and are not distinguished with the lesser or greater than signs. Consequently, the tool takes them into account when computing coreference. For instance, in an utterance such as "... they seem to walk like that overlap and so so their arms are all around..." (DID108-S001), the markup element <overlap> does not appear as such. This leads the tool to consider it as an NP whose determiner is *that* instead of being identified as the head of the NP. Consequently, it alters grammar features such as *HeadText* which is one of the parameters used to identify the antecedent. CESAX's coreferent candidate selection is thus slowed down.

Secondly, because the final purpose is to assign a textual antecedent, another issue must be pointed out. The list of antecedent candidates in the lower left window of the application (see Figure 17) keeps increasing, which slows down the process of selection by the user. We choose to select the *identity*, *assumed, inert* and *new* features to accelerate tagging. In our terms, *identity* and *assumed* belong to the same class. Selecting one or the other depends on convenience in the use of the interface. Choosing the *identity* feature sometimes forces the user to read the long list and, yet, only serves the purpose of recreating the tie with the textual antecedent. Consequently the *assumed* tag is preferred in such cases.

Thirdly, the fact that every NP is scanned, added to the fact that some NPs are children of other NPs, leads the application to loop back on the same form more than twice at times. This, again, slows down the tagging process. For example, the following string: "<B> or something like that it was so great seeing all those parks and those big castles and everything and it really made me dream" is presented in the first place. In the second place, "those big castles" is proposed. This is due to the syntactic structure of the phrase which embeds an NP within an NP but its looping requires further checking.

The aforementioned issues seem to be the source of a number of errors which are reported in the following tables.

		Semi-automatic annotation			
		Assumed / identity	New	Inert	unmarked
Gold standard	Assumed or identity	4			
	New				
	Inert				

Table 15: Confusion matrix for semi-automatically tagged forms of *this* with CESAX

		Semi-automatic annotation			
		Assumed / identity	New	Inert	unmarked
Gold standard	Assumed or identity				1
	New		1		
	Inert				

Table 16: Confusion matrix for semi-automatically tagged forms of *that* with CESAX

		Semi-automatic annotation			
		Assumed / identity	New	Inert	unmarked
Gold standard	Assumed or identity	8			3
	New		1		2
	Inert	1		1	

Table 17: Confusion matrix for semi-automatically tagged forms of *it* with CESAX

There clearly is a number of incorrect tags assigned to the forms especially concerning the *it* form. These may be due to manual selection confusions. The number of possibilities provided in the special selection window might be counter-productive.

This solution is of interest considering its implementation of the new/given type of information. However, the output files also include non-pertinent coreferential annotation and lack determiner and pro-form fine-grained tags. Deploying the solution on our corpora at this stage (we received this annotation tool towards the end of our PhD) might still be considered as an option, but time remains a strong limitation. The main difficulty concerns the compatibility of our PoS-functional scheme with CESAX's parsing scheme. A number of stages need to be undertaken. First, our corpus files require tagging with CESAX (with is a time consuming semi-automated solution). Then, the givenness information of each form in the output files needs to be extracted before being merged with our already annotated corpora. This means that CESAX output files need to be aligned with our corpus files so that the tokens of both types of output match each other. Aligning the texts requires further programming as tokens of both types of output files need to be indexed to be merged. In sum, operating CESAX manually for each occurrence and file alignment programming amounts to many more hours of development, which we could not afford at the stage of our PhD.

Merging the PoS functional, contextual and positional annotation layers within a CESAX parsed structure would provide a scheme that would include PoS, functional, syntactic and discourse annotation. The prospects of such a possibility are promising and we shall keep this in mind for another phase of our work. Indeed, albeit extremely tempting to follow that thread, it appears more reasonable to focus the present work on the PoS-functional and context types of annotation described in Sections 4.2.2.1 to 4.2.2.4. At this stage in our research, this solution will not be implemented in our protocol for the aforementioned reasons. It is therefore outside the scope of the work reported in this thesis.

4.2.3 The annotation setup

In this section, we identify our requirements for the annotation setup which is implemented on the three corpora. In Section 4.2.2, two problems have been addressed: inter-annotation layer diversity and intra-annotation layer diversity. As mentioned before, the pro-form model, characterised by several levels of interpretation, requires the introduction of several types of annotation. Consequently, we have chosen to create three annotation layers in order to comply with Leech's rule on the necessity to always keep annotation separable (Leech 2005, Section 4). We have assessed the possible levels of the annotation scheme to implement on the corpora and three levels have been retained: PoS-function, position and context. The second problem is related to the diversity of possible tagsets within a level (intra annotation layer diversity). We have decided that only one specific annotation type of a given level must be applied to the corpora. For each annotation level a tagset has been identified. For PoS annotation, the tagging process is automated with the use of a PoS-tagger based on the Penn Treebank. Position annotation is also applied automatically with a specific PERL program. Context annotation is applied manually. In order to prepare the corpora for automatic data analysis, we adopt two data structures which ensure the co-existence of several annotation layers: the NITE data set Carletta *et al.* (2003) and what we call sequenced data structures (SDS)(see Chapter 5.1.1 for its implementation).

The NITE data model adopts a multi-layer approach. It relies on an XML-based data storage format aimed at supporting heavily cross-annotated data sets. The NITE data set model splits the data in several files, each including a single-type annotation and is called a 'layer' (2003, Section 4). All the information on the multiple-layer structure is stored in some sort of index file which includes information of all the codings used in all the annotation layers, *e.g.* the PoS tagset. Within each layer, the XML data structure annotations are organised according to a parent-child type of relationship. Time and structural layers allow the data to be organised in such a way that some lower-level layers can be dependent on higher-

level ones and “timings can percolate up structural layers from a time-aligned layer at the base” (Carletta *et al.* 2003, Section 4). Some attributes such as time may be found in several annotation files, hence allowing alignment of the annotation layers. At operational level, corpora consist of sets of files matching their annotation type. Corpus identification is stored in each of the annotation attributes to allow clear identification of each query result. This setup is compliant with Leech's requirements on annotation (Leech 2005, Section 3). It supports automatic analysis. It is reusable in the sense that annotated text units can be extracted for integration in databases or for exploitation by search tools. It is multi-functional since the structure of files are not dedicated to the sole study of *this* and *that*. For instance, existing annotations could come in useful for other research such as the identification of completive clauses in learner and native corpora. Another advantage of the NITE XML toolkit is that it can be used with multimodal corpora, which means that our annotation scheme could be used with corpora including videos.

In the sequenced data structure, the data are stored in tables whose lines represent single occurrences of each form in the corpora. Each form is characterised by a line of features that provides information sourced from annotation. Our requirement is to have a consistent structure that allows any sort of data (texts and annotations) to be exchanged. In computer science, data exchange is defined as “data structured under one schema (which we call a source schema) [which] must be restructured and translated into an instance of a different schema (a target schema)” (Fagin *et al.* 2005, 90). Indeed, with this structure, NLP tools can be used to explore and manipulate the data with a view to understanding how the aforementioned linguistic factors interact during the selection process of a form as the speech unfolds. The data structuring phase is the condition for subsequent processing that could be carried out by other researchers.

All in all the setup is designed to ensure that one native and several learner corpora are on par in terms of annotation schemes and data structures. As a result, it is

possible to develop scripts that query the corpora simultaneously and extract lists of occurrences corresponding to combinations of features that span several annotation layers. This comes as one step in the direction of interoperability which implies facilitated "data-exchange from annotated corpora and the reusability of annotated corpora" (Ballier and Martin 2013, 53).

4.3 Summary

In this chapter, we show that it is possible to devise a single annotation scheme to make several corpora interoperable. The state of the art presented in the first section highlights two requirements. Firstly, it shows that even though annotation is an essential part of corpora, the heterogeneity in annotation types in terms of labels and levels hinders interoperability between corpora. There is a need for a single annotation scheme which provides a common framework for corpus comparability. This advocates for a uniform approach concerning the annotation of corpora. Secondly, it highlights the fact that interoperability must depend on a common data structure which needs to be put in place in order to exchange data from all corpora within the same structure.

The second section shows how we meet the two requirements in order to determine the annotation setup which is used on three annotated corpora (one native and two learner corpora). Concerning the issue of heterogeneity in annotation, we have devised a single three-level annotation setup. For the first level, we have shown the need for a fine-grained PoS-functional annotation. We have justified why the Penn Treebank is suitable to be modified in order to introduce a functional distinction between pro-forms and determiners. The use of the modified tagset allows us to introduce i) the fine-grained functional annotation for *this* and *that* and ii) the PoS categories of all other word forms in the corpora. Our approach regarding automatic learner error detection has led us to enrich the corpus with a syntactic position annotation which gives more information on the distribution of the forms, *i.e.* oblique and nominative cases. Added to the PoS-

functional annotation, this second level helps detect errors on the *it*, *this* and *that* pro-form microsystem. As a third type of annotation, we have shown the need for a contextual layer to provide information on the endophoric/exophoric distinction between the forms. Thanks to this single multi-level annotation scheme, we have reached interoperability at annotation level since the data extracted from the three annotated corpora should be comparable.

For the second requirement regarding a common data structure, we have adopted two data structures to encapsulate all the data (text and annotation) from the three corpora. Firstly, we choose the NITE XML standard to format the corpora and allow multi-layer queries in context. Secondly, we explain why a sequenced, instance-based, data structure should help synthesise all the relevant annotations of each form in matrices of features whose complexity can be analysed automatically with machine learning tools. Overall, we have devised two data structures which provide interoperability at structure level since the data are exchangeable within each structure irrespective of their initial corpus.

These requirements need to be put in place before any analysis is carried out. The next chapter shows how the annotation scheme is implemented with several software tools depending on which annotation layer is to be created. It also focuses on the implementation of the two data structures which support corpus comparability.

Chapter 5 An interoperable structure for multi-corpus querying

This chapter explains the implementation of the annotation scheme within architectures that allow corpus comparability. Two kinds of implementations are necessary. First of all, each annotation layer needs to be applied to each corpus. To do so, automation is put in place for the PoS-functional and the positional layers thanks to two distinct software programs. In the first case, we use TreeTagger to implement a modified version of the Penn Treebank tagset. The purpose is to use the tool to PoS tag all the corpora and to apply specific tagging for the functional realisations of *it*, *this* and *that*. In the second case, we show how a specific PERL program relies on the PoS-functional layer to apply positional tags to the forms. The second kind of implementation is that of the architecture of the data. As explained in Section 4.2.3, we need two types of structures: an XML-based structure to support multi-layer queries with contextualised results and a sequenced data model based on feature matrices to support multifactorial analysis. The problem is that the corpora are in a format inherited from the annotation phase with TreeTagger. Therefore automation and conversion methods need to be applied to provide fast and consistent conversion of the TreeTagger structure into the two aforementioned structure types.

There are three sections in this chapter. In Section 5.1, we show the implementation of two layers of the three-layer annotation scheme. We show how the Penn Treebank tagset is modified to apply the PoS-functional layer to the corpora. We present the PERL implementation of an algorithm which is used to apply positional tags to the forms. In each case, we report tagging accuracy results. In Section 5.2, we show how the NITE XML structure is applied to the three

corpora. We present multi-layer queries that can yield contextualised results. In Section 5.3, we show how the corpus data is transformed into tables of sequences —or instances—of linguistic features to support multifactorial analysis carried out with software tools.

5.1 Implementing annotation layers

In this section, we successively present the implementation of the PoS-functional annotation and the positional annotation layers.

5.1.1 Implementing the PoS and functional annotation of *it*, *this* and *that*

In this section, the focus is placed on the procedure used for the PoS and functional annotation of the three corpora described in Section 4.2.1. The objective is to allow the distinction between the different functional realisations of *it, this* and *that*. To do so, we introduce finer-grained functional tags for the three forms and we PoS-tag the corpora. It is necessary to follow a specific procedure that involves i) modifying the tags applied to the forms in the Penn Treebank corpus, ii) training TreeTagger on the newly tagged version of the Penn Treebank corpus and iii) tagging the learner corpora to be compared. Section 5.1.1.1 covers the tools used for the operations. Section 5.1.1.2 describes the aforementioned first step and Section 5.1.1.3 focuses on the last two steps of the procedure.

5.1.1.1 The tools

For the PoS-functional annotation process, we use the TreeTagger[22] tool (Schmid 1994) as mentioned in Section 4.2.2.1. TreeTagger is a command line program that can be launched by simply typing its name followed by the files to train on or the files to tag. There is also a GUI interface developed by Ciarán Ó Duibhín which is supported on Windows Operating Systems[23]. Like most taggers, a training phase is

[22] Downloadable at www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ (last accessed on 7/10/2015)

[23] Downloadable at www.smo.uhi.ac.uk/~oduibhin/oideasra/interfaces/winttinterface.htm (last accessed on 7/10/2015)

carried out on already tagged texts. Once trained, TreeTagger has constructed an internal grammar. So if it has been trained on Penn Treebank compliant token-PoS pairs, it can then assign Penn Treebank PoS tags to a new text as Figure 18 shows. In this figure, the learner's words can be read vertically and, for each word, a PoS tag has been assigned by TreeTagger. Each column is separated by a tabulation.

```
DID0020-S001
<S>	SYM
<B>	SYM
okay	JJ
I	PRP
'm	VBP
I	PRP
've	VBP
been	VBN
to	TO
:	:
USA	NNP
which	WDT
was	VBD
-LRB-	-LRB-
er	UH
-RRB-	-RRB-
very	RB
impressive	JJ
because	IN
everything	NN
is	VBZ
tall	JJ
and	CC
huge	JJ
and	CC
I	PRP
arrived	VBD
in	IN
Los	NNP
Angeles	NNP
</B>	SYM
```

Figure 18: Sample of a Diderot-LONGDALE transcript tagged with TreeTagger

As will be shown in section 5.1.1.2, the fact that TreeTagger's training phase can be conducted on any tagged text to learn the grammar makes it possible for the researcher to modify some of the tags. This is obviously a very important feature of the software, as we intend to introduce the determiner/pro-form distinction. Even though other taggers might be used for the task, the fact that TreeTagger is an easily available open source application, also supported on a Linux operating

system, makes it the perfect candidate for our project as no licensing costs are incurred.

Prior to the actual training phase with TreeTagger, during which the training module of TreeTagger is run on the modified version of the Penn Treebank WSJ subset, it is necessary to modify some tags. To achieve this, Stanford University's Tregex is used (Levy and Andrew 2006). It is a tool that allows the reading of text parsed with the Penn Treebank parsing scheme. In each corpus file, the text follows an indented structure with bracketed elements. Tregex is a utility that converts the bracketed-text structure into visual trees. Figure 19 shows a representation of the sentence: “For six years, T. Marshall Hahn Jr. has made corporate acquisitions in the George Bush mode: kind and gentle” (wsj_0100.mrg-1). To function, corpus bracketed files are imported into the Graphical User Interface by way of regular expressions. Regex patterns can be created to query the corpus.

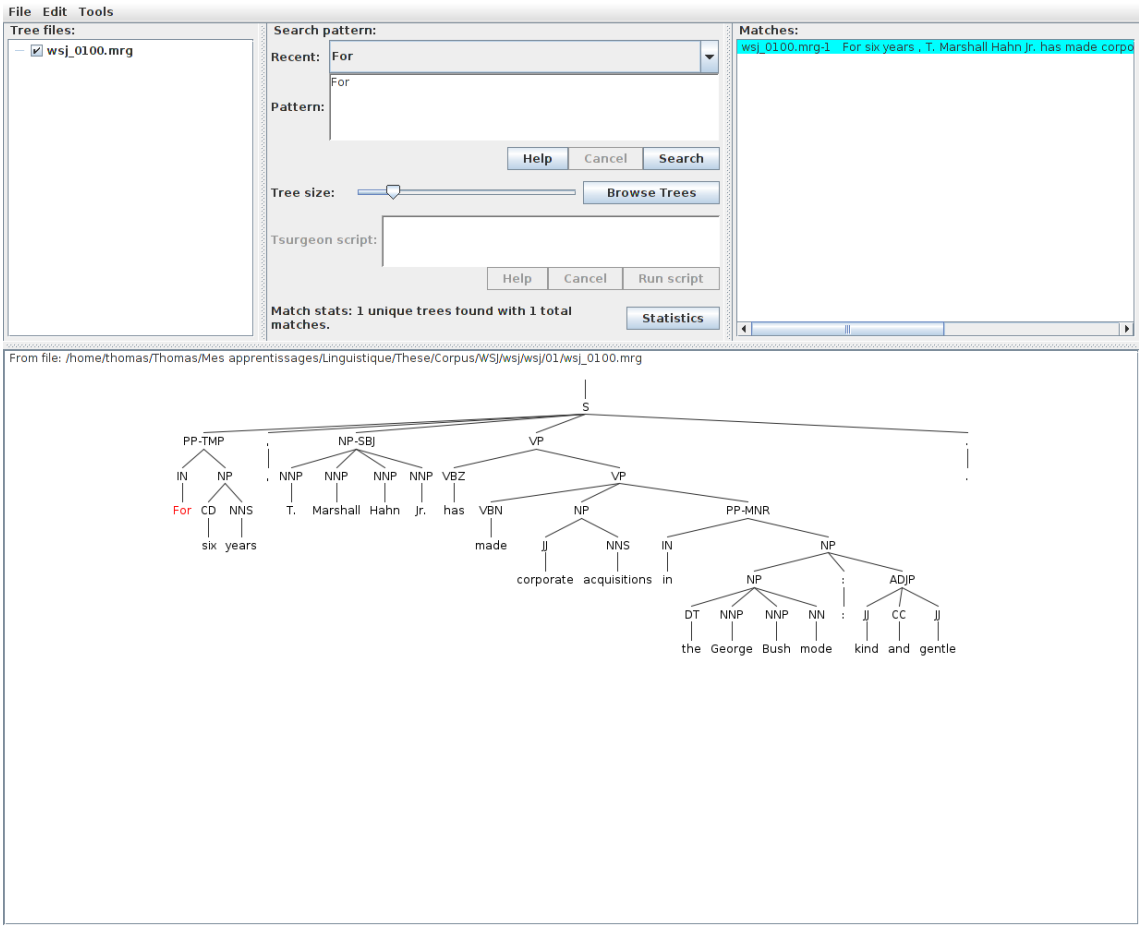

Figure 19: An example of a tree structure in the Penn Treebank corpus

The sentences that match the pattern are returned and highlighted. In Figure 19, the basic pattern *For* (note the capital F) was entered and the case-sensitive system returned the one sentence in our sample which includes the three letters *f, o, r* in a row. Tregex includes a utility called Tsurgeon (Levy and Andrew 2006) that allows the editing of trees. This utility draws on the pattern identification functionality and adds the possibility of changing the tree elements that correspond to the pattern. In the previously mentioned sentence, it would be possible to get all initial *For* forms to be changed to another label more appropriate for prepositions heading prepositional phrases, in order to refine a POS tagset to distinguish this prepositional use from other clausal uses of *for* (where *for* can be replaced by *because*). This principle would work on PoS tags provided that a Tregex query would identify them.

The power of the tool lies in its ability to query the trees vertically and horizontally. Vertically, it is possible to find specific constituents dominated by other specific constituents. For instance, still with the same sentence, it is possible to query all NPs dominated by a VP. This pattern only returns the *corporate acquisitions* NP as it is the only NP part of the VP. Horizontally, it is possible to look for a succession of elements be they words, PoS or constituents of the trees. For instance, it is possible to define patterns matching an adjective (JJ) immediately followed by a noun in the singular or plural form NN(S). This returns a chain of PoS-tagged items that includes *corporate acquisitions* again. These two types of vertical and horizontal queries show that it is possible to target the same query results but in different manners. This is particularly useful as shown in Section 5.1.1.2 which concerns the modification of PoS tags.

All in all, Tsurgeon and TreeTagger are used in conjunction to apply PoS-functional annotation to the two learner corpora. The native corpus, which is the WSJ subset of the Penn Treebank, is somewhat simpler as tagging is already performed, and only tag modification is necessary. As a result of the PoS functional annotation phase, all three corpora files are formatted and converted to a one-line-per-word

format, as exemplified in Figure 18 (the resulting tags are different from those shown in the figure).

A final step, and not the fastest one, is undertaken, namely manual checking of the two learner corpora. Indeed, a fully-automated annotation process yields a level of errors (see Section 5.1.1.3) that cannot be accepted for subsequent statistical analysis. Consequently, all forms of *it*, *this* and *that* are manually checked in context in order to verify the accuracy of the process. Errors are corrected to ensure the most robust statistics.

5.1.1.2 Modifying the Penn Treebank tagset

To introduce functional tags in the PoS tagset of our annotation scheme, we use the WSJ corpus which has been PoS-tagged and parsed with the Penn Treebank scheme. The purpose is to use the parsed structure to modify the PoS tags of the corpus and to create a new version finally annotated with PoS-functional information. This newly annotated version of the corpus can then be used for the training phase of TreeTagger. While training, the tool learns the PoS-functional scheme that has been introduced.

The principle is to open the WSJ corpus files with an editing tool in order to search for and modify the forms that correspond to previously defined patterns. In the case of *this* and *that*, the task is to introduce the determiner/pro-form distinction. In other terms, the legacy DT tag, as employed in the Penn Treebank, must be eliminated. In turn, new tags must be introduced. As explained in Section 4.2.2.1, the DT and TPRON labels are created to mark the determiner/pro-form distinction. However, the fact that *this* and *that* also have other functions in sentences needs to be taken into account to make sure that each form receives a clear label. Incidentally, the two forms can be adverbials (see second occurrence of *that* in example 46) and thus require a specific label when realised in this function. If we only take *that* into consideration, it appears that *that* can also have two extra

hypotactic functions. It can either be a complementiser (see first occurrence of *that* in Example 46) or a relative pronoun (Example 47).

46) but it's interesting why not because I:I think *that* (er) well I like to: cook for myself so and I don't want to walk *that* far just to buy some food so DID0014.S001

47) and we listen the same kind of weird music *that* no one else listen to (er) apart from us DID0038-S001

As well as the functional realisations of *this* and *that,* their interactions with each other and *it* need to be taken into account (see Section 3.2.2.2.1). The fact that the two forms are used with confusion by learners makes it paramount to also identify *it* pronouns in the corpus. The problem is that *it* can also have several realisations in the same manner as *this* and *that*—for studies on the detection of *it* according to functional realisations, see (Boyd *et al.* 2005) and (Paice and Husk 1987). There is the referential realisation of the pronoun and there are others that are not referential. So this distinction also needs to be considered in order to only retrieve the pronoun realisation which can be, in turn, compared with pro-form realisations of *this* and *that*. As the objective is to train TreeTagger with the highest level of accuracy, it is crucial to make sure that each functional realisation of either of the forms is accurately identifiable in the Penn Treebank corpus so that it can be used as a gold standard for further language processing. The next subsections give details on how each form is identified unequivocally in the Penn Treebank corpus.

5.1.1.2.1 Identifying and modifying *this* tags

The first thing we do is determine a way to identify each type of *this* in the Penn Treebank WSJ subset. In order to do so, Stanford University's Tregex is used to load the corpus and query it vertically and horizontally. We proceed by establishing an inventory of all *this* forms in the 24 subsets of the Penn Treebank WSJ corpus. The first step consists in searching all forms according to the PoS tag (see PoS part in Table 18). Apart from the DT tag, 4 forms of *this* receive the noun singular NN tag

or the noun proper NNP tag, which appears to be inconsistent with the rest of the tagged forms. A closer look at the hapax shows that the NN-tagged *this* is a pro-form, and thus needs to be tagged as such (see Example 48). The NNP-tagged *this* forms, however, reveal that the form is part of two proper nouns referring to two TV programs (see Examples 49, 50 and 51). The adverb-RB form however appears to be correctly tagged (see Example 52).

48) Denise McDonald, a spokeswoman for MADD, says, “It's scary, because anybody could do *this*.” wsj_1625.mrg-70

49) Sears , Roebuck & Co. signed a contract with Bob Vila , the former host of the popular public television program '*This* Old House', 0 *T*-1 to star in a half-hour home improvement show sponsored * by the giant retailer. wsj_0962.mrg-1

50) With Mr. Vila as host, '*This* Old House' became one of the Public Broadcasting Service's top 10 programs, *-1 airing weekly on about 300 of the network 's stations and seen * by an average of 12 million viewers. wsj_0962.mrg-6

51) “One of the things that *T*-1 continues *-3 to worry me is this monetary warfare between the Treasury Department and the Federal Reserve Board,” said *T*-2 Lawrence Kudlow, a Bear , Stearns & Co. economist, on ABC's '*This* Week'. wsj_2384.mrg-14

52) He says the big questions—“Do you really need *this* much money *-2 to put up these investments?” […] wsj_0629.mrg-52

This first search only allows us to detect tagging errors. The second step is the search for the pro-form and determiner distinction (see the functional part in Table 18). It is mandatory to use tree-based queries, as only the position in the syntactic structure can provide information on the distinction. To isolate the determiner forms, we use a Tregex pattern that seeks all *this* DT forms which are dominated by an NP and which are not the last child of an NP. To capture pro-form forms, the query searches for all *this* DT forms which are dominated by an NP and which are the last child of an NP. Once the results for these two queries are added, it appears that some occurrences are missing (see Table 19). Further research with patterns

shows that some DT tagged forms appear under the ADVP and ADJP tags, which in fact corresponds to an adverbial function. Another hapax is found with the PRN tag used to mark parenthetical information in the sentence (Bies *et al.* 1995, 50) (see Table 20). In this case, *this* appears to be a determiner. As a result of this inventory, the same number of occurrences is found with POS-based and function-based queries, and a series of patterns is identified that matches either a determiner or a pro-form form. Therefore, these patterns can be used with Tregex's module Tsurgeon to automate the modification of DT tags that correspond to queries that retrieve pro-forms. Table

This (form count)	/^[T|t]his$/
POS	**PoS-based queries**
This NNP	/^[T|t]his$/ > /^NNP$/
This NN	/^[T|t]his$/ > /^NN$/
This RB	/^[T|t]his$/> /^RB$/
This DT	/^[T|t]his$/> /^DT$/

Functions	**Tree-based queries**
This determiner and pro-form	DT </^[T|t]his$/ & > /^NP.*$/
This/determiner	DT </^[T|t]his$/ & [> /^NP.*$/ & !>: /^NP.*$/]
This/pro-form	DT </^[T|t]his$/ & >: /^NP.*$/
This/adverbial tagged DT in Verb Phrase	DT </^[T|t]his$/ & > /^ADVP.*$/
This/adverbial tagged DT in Adjective Phrase	DT </^[T|t]his$/ & > /^ADJP.*$/
This/determiner tagged DT in Parenthetical marking	DT </^[T|t]his$/ & > /^PRN.*$/

Table 18: Queries used to detect *this* forms in the Penn Treebank WSJ corpus according to their PoS and functional distinctions

This (total count of the form)		**2853**
POS		
This NNP	3	
This NN	1	
This RB	1	
This DT	2848	
Total		**2853**
Functions		
This determiner and pro-form		2839
This/determiner	2197	
This/pro-form	642	
This/adverbial tagged DT in Verb Phrase		4
This/adverbial tagged DT in Adjective Phrase		4
This/determiner tagged DT in Parenthetical marking		1
Sub-total DT		2848
This NNP		3
This NN		1
This RB		1
Total		**2853**

Table 19: *This* counts for POS-based and function-based queries

This/adverbial tagged DT in Verb Phrase	DT </^[T|t]his$/ & > /^ADVP.*$/	4
This/adverbial tagged DT in Adjective Phrase	DT </^[T|t]his$/ & > /^ADJP.*$/	4
This/determiner tagged DT in Parenthetical marking	DT </^[T|t]his$/ & > /^PRN.*$/	1
This NNP	/^[T|t]his$/ > /^NNP$/	3
This NN	/^[T|t]his$/ > /^NN$/	1
This RB	/^[T|t]his$/> /^RB$/	1

Table 20: Summary of problematic cases for *this* counts, corresponding queries and counts

Each of the patterns is passed on to Tsurgeon to modify the *this* DT tags according to their actual function. Three tags are used in the modification process. We keep DT and RB to indicate the determiner and adverbial functions. We introduce a new tag TPRON for the pro-form function. The second *functions* part of the table shows the distribution of the 2,848 *this* DT accounted for in the first part.

5.1.1.2.2 Identifying and modifying *it* tags

As pointed out in Section 3.2.2.2.1, *this*, *that* and *it* interact within a microsystem of reference. For learners, they appear to be competitor forms. Our purpose is to compare all the referential forms and therefore we need to make them retrievable. The issue concerning the *it* form is that it is not only a referential form as a pronoun, but it is also a non-referential form. The literature on this issue essentially assigns three or four categories to the non-referential forms. For Biber *et al.* (1999, 332) there are 3 functions (First two examples provided by Biber *et al.*.

1. Empty subject/object. It occurs where there are no participants to fill the subject/object slot, *i.e.* talking about the weather, time and distance. For example: “it's nice today”.
2. Anticipatory subject/object. It is inserted as subject where a clause has been extraposed, *i.e.* “I was thinking, it'd be nice to go there”.
3. Subject in cleft constructions placing focus on a particular element in the clause, *i.e.* “It is this difference that underlay the comment made to me by one interpreter” (Extracted from the COCA[24] corpus).

For (Huddleston and Pullum 2002) there are five uses of *it* “that are not anaphoric and do not refer directly to salient entities”. The authors distinguish several functions:

1. Extraposition as in “it's ridiculous that they've given the job to Pat”. In this case, the pronoun *it* takes the subject position and the subordinate clause is placed in the second part of the utterance with an extraposed subject position. The authors specify that “the subject properties of the extraposed element are transferred from the subordinate clause to *it*”. The clause can replace *it*.

2. Impersonal *it* as in “It seemed that/as if things would never get any better”. In this case, “the subject is semantically empty, so that the

[24] http://corpus.byu.edu/coca/ (Last access March 31, 2016)

content clause represents the sole argument of the matrix clause. Replacement of the *it* clause is possible with a paraphrase containing an adverb" (Huddleston and Pullum 2002, 960). The authors add that the subordinate clause cannot replace *it*.

3. *It*-cleft construction as in "It was your father who was driving – No it wasn't, it was me". In this case, there is no meaning in itself for *it* and no part of the construction can be regarded as its antecedent.

4. Weather, time, place, condition as in "It's only two weeks since she left." For them, there is no meaning for *it*. "It has the purely syntactic function of filling obligatory subject position."

5. *It* as subject with other predicative NPs, as in "It's a wonderful view" in which *it* can be replaced by *this.* In this case, the *it* does refer to an exophoric entity which, in Huddleston and Pullum's view, places it outside of anaphora.

When we cross both Biber *et al.* and Huddleston and Pullum's views, there are four common syntactic features to isolate non-referential/non-anaphoric uses of *it*:

1. Impersonal use

2. Extrapositional use

3. Cleft use

4. Weather/time/distance

Now that we have seen how non-referential *it* is distinguished, it is relevant to look at it with the perspective of the Penn Treebank. Table 21 provides an overview of the Tregex queries used to identify the various constructions in the Penn Treebank.

Type of *it*	Position	Tregex query	Example
Extrapositional	subject	PRP < /^[I|i][T|t]$/ > (NP > /^NP-SBJ.*/)	wsj_0037.mrg-34 *It* *EXP*-1 's a shame 0 their meeting never took place .
	object	PRP < /^[I|i][T|t]$/ > (NP > (NP [$++ /^S.*/ > VP]))	wsj_1849.mrg-22 From the history of capitalism we can take *it* *EXP*-1 as a sound bet that if it *EXP*-2 takes only 43 cents * to buy a dollar 's worth of a firm 's capital stock , an alert entrepreneur wo n't look the other way .
Cleft		PRP < /^[I|i][T|t]$/ >> /^.*CLF.*$/	wsj_0126.mrg-28 `` *It* is the very building of it *ICH*-4 that *T*-2 is important , not how much of it *T*-3 is used *-1 or its economics . "
Impersonal		PRP [< /^[I|i][T|t]$/ & > (/NP-SBJ.*/ $+ (VP [< (/^VB.*$/ < /\bseem.*\b|\blook.*\b|\bappear.*\b|\bhappen.*\b|\bturn.*\b|\btake.*\b/) & [< /^SBAR.*$/ \| < (/^ADJ.*$/ < RB)]]))]	wsj_1469.mrg-37 For a short time after June 4 , *it* appeared that the trade picture would remain fairly bright .
Time		PRP [< /^[I|i][T|t]$/ & > (/NP-SBJ.*/ $+ (VP [< (/^VB.*$/ </\b's\b|\bis\b|\bwas\b/) & < (/^NP.*$/ < (/^NP*$/<< time))]))]	wsj_0088.mrg-29 That response annoyed Rep. Markey , House aides said 0 *T*-1 , and the congressman snapped back that there had been enough studies of the issue and that *it* was time for action on the matter .

Table 21: Tregex queries to identify non-referential *it* in the Penn Treebank WSJ corpus

For extrapositional *it*, a PoS-based type of exploration provides details on the sisters that follow specific PRP *it* depending on the construction. A tree-based type of exploration provides details on the constituents that dominate the forms, *i.e.* the parents. PoS-based queries can rely on the EXP tag used for *expletives*. Taylor explains the case: “In order that certain kinds of constructions can be found reliably within the corpus, we have adopted special marking of some special constructions. For example, extraposed sentences which leave behind a semantically null *it* are parsed as follows, using the *EXP* tag” (Taylor, Marcus, and Santorini 2003, 14) see also (Bies *et al.* 1995, 25)

Another way of identifying extra-positional *it* may be achieved via tree-based queries. The following specification indicates the tree-based sequence: “Clauses

that are extraposed from subject position are labelled S or SBAR. The extraposed clause is attached at VP level and adjoined to the *it* with *EXP*- attach. The NP containing *it* and *EXP* is tagged -SBJ."(Bies *et al.* 1995, 25) In other terms, *it* is dominated by the PRP tag, which is dominated by an NP which is dominated by an NP-SBJ.

By cross-referencing queries based on the EXP tag and queries based on tree branches, we see several differences in the results. Firstly, the EXP-based query does not capture *it* followed by the ICH tag. Secondly, the tree-based query captures a limited number of referential *it*. They appear in contexts such as

53) Coda, an oil and gas concern, said 0 *it* and its partners received $ 7 million *U* in cash and $ 10 million *U* in five-year notes for the Kansas intrastate pipeline. (wsj_1083.mrg-2)

When looking at the occurrences, they fall into two classes: when *it* is one of two elements of an NP in a subject position and when *all* is found right after *it*. Consequently, the tree-based query is favoured (see Table 21) because it also captures all *it* forms followed by the ICH tag. The few referential *it* forms however have to be eliminated manually.

In the Penn Treebank, cleft constructions are found under the S-CLF, for declarative utterances, and under the label SQ-CLF, for questions, and under the label SINV-CLF, for inversion in clefts (Bies *et al.* 1995). So the query in the table identifies forms of *it* PRP that are indirectly dominated by a CLF.

Impersonal *it* is characterised by its being followed by verbs like appear and seem as in "it seems that..." (Huddleston and Pullum 2002) The query in Table 21 matches all *it* PRP forms that are dominated by an NP-SBJ which is a sister of a VP that includes a clause or an adjectival phrase introduced by a particular verb such as *appear*.

Time-related structures are part of the it-cleft construction, *i.e.* “It wasn't so long ago that a radio network funded * by the U.S. Congress […] was accused ...” (wsj_2406.mrg-1). However, there are cases related to time that are not classified as cleft in the Penn Treebank. These correspond to a configuration in which the noun *time* does not appear as being modified by a relative clause. Instead, the NP is a sister of a VP which is introduced by *be* and includes the word *time*. Other time-, distance- or weather-related queries are tested but no systemic query can be identified to capture occurrences that are not already part of the above queries. Meteorological uses of *it* are more likely to be retrieved on a lexical basis. It is thus judged preferable not to include them since training of the TreeTagger needs to be done on true occurrences.

Once all *it* constructions are matched to their queries, it is possible to proceed to the modification of the PoS tags to reflect the referential/non-referential distinction. For referential *it,* we keep the PRP tag and, for non-referential *it*, we introduce a new PNR tag for Pronoun Non-Referential. Tsurgeon scripts are used to systematically apply the modification across the whole corpus.

5.1.1.2.3 Identifying and modifying *that* tags

The identification procedure for *that* forms is similar to the one used for *this*. After an initial count of the occurrences in relation to their PoS tag, a detailed inventory of all forms for each syntactic configuration is carried out. It reveals a number of inconsistencies in which, for instance, actual complementiser occurrences (nominal and verbal alike) are in fact tagged as DT in the initial Penn Treebank. Further exploration is carried out to reveal many more inconsistencies and a synopsis is offered in Table 22.

Assigned categories	Initial Penn Treebank – Before corrections	Modified Penn Treebank – After correction
Total *that* complementiser	5992	5696
Total *that* relative pronoun	2483	2899
Total *that* determiner and pro-form	1918	1790

Total *that* adverbial	26	36
NN tag	1	0
VBP tag	1	0
TOTAL	10421	10421

Table 22: Recap of tagging inconsistencies in the Penn Treebank WSJ corpus

Table 23 gives details regarding the confusions in tagging forms of *that*. For instance, the table shows that 804 *that* pro-forms were tagged as DT, which is in line with the tagset's definition as it originally include pro-forms. Conversely, there are 177 occurrences of *that* which were erroneously tagged as 'subordinator' IN in the corpus and which happen to be pro-forms.

	Assigned DT	Assigned IN	Assigned WDT	Assigned NN
Determiner	813	35	0	1
Pro-form	804	71	66	0

Table 23: Confusion matrix of tagged *that* forms in the Penn Treebank and their true pro-form or determiner function

By rematching patterns, which correspond to syntactic configurations, to their actual functional tag, we build a comprehensive inventory of the queries that can be used to modify the tags. Annex F shows the Tregex syntactic queries used to re-assign PoS tags according to their true class. Green-coloured patterns in the annex indicate cases that were initially tagged otherwise in the Penn Treebank (tagging errors). For the sake of convenience, and because the purpose of our study is to explore the determiner and pro-form functions of the forms, Table 24 is an excerpt that provides details on the patterns.

DT	/^DT$/ < /^[T\|t]hat$/
Assigned DT	/^DT$/ < /^[T\|t]hat$/ > /^NP.*/
Pro-form	/^DT$/ < /^[T\|t]hat$/ >- /^NP.*/
Determiner	/^DT$/ < /^[T\|t]hat$/ > /^NP.*/ !>- /^NP.*/
Pro-form	**/^DT$/ < /^[T\|t]hat$/ > /^INTJ.*/**
Assigned IN (complementiser)	**/^IN.*/ < /^[T|t]hat$/ > /^NP.*/**
Pro-form	**/^IN.*/ < /^[T|t]hat$/ >- /^NP.*/**
Determiner	**/^IN.*/ < /^[T|t]hat$/ > /^NP.*/ !>-/^NP.*/**

Assigned WDT (relatives)	**/^WDT$/ < /^[T\|t]hat$/ > /^NP.*/**
Pro-form	**/^WDT$/ < /^[T\|t]hat$/ >- /^NP.*/**
Determiner	**/^WDT.*/ < /^[T|t]hat$/ > /^NP.*/ !>-/^NP.*/**
Assigned NN	
Pro-form	**/^NN.*/ < /^[T|t]hat$/**
	Total *that* determiner and pro-form

Table 24: Syntactic patterns that match determiner and pro-form uses of *that* in the Penn Treebank WSJ corpus

Table 25 sums up the changes of labels and rephrases the query patterns in terms of positional properties. It gives the resulting frequency counts. Each line corresponds to a query that matches the tag assigned in the corpus or a function (indented). For instance, there are 1,614 occurrences of *that* tagged DT in the corpus of which 801 are pro-forms and 813 others are determiners. Bold lines indicate patterns that do actually identify determiners and pro-forms but which were initially tagged otherwise. For example, there are 71 occurrences of *that* tagged IN which are actually pro-forms. If the Penn Treebank annotation had been fully consistent, these occurrences should have been tagged DT. The aforementioned annex provides a comprehensive understanding of how form counts balance off. While making this point, our purpose is to show that a detailed inventory based on the syntactic trees of the Penn Treebank lays the ground for the completion of a gold standard in which all occurrences of the form do actually correspond to the tags that are to be assigned.

DT	DT tagged *that* in the Penn Treebank	1918
Assigned DT	DT tagged *that* part of an NP	1614
Pro-form	DT tagged *that* last child of NP	801
Determiner	DT tagged *that* not last child of NP	813
Pro-form	**DT tagged *that* dominated by interjection**	**3**
Assigned IN (complementiser)	**IN tagged *that* part of an NP**	**106**
Pro-form	**IN tagged *that* last child of NP**	**71**
Determiner	**IN tagged *that* not last child of NP**	**35**
Assigned WDT (relatives)	**WDT tagged *that* part of an NP**	**66**
Pro-form	**WDT tagged *that* last child of NP**	**66**
Determiner	**WDT tagged *that* not last child of NP**	**0**
Assigned NN		**1**
Pro-form	**NN tagged *that***	**1**
	Total *that* determiner and pro-form	**1790**

Table 25: Counts of occurrences of *that* according to their position in the Penn Treebank

The modification process with Tsurgeon relies on the patterns that match the functions. In the case of *that* a series of new tags is introduced. The complementiser tag for *that* is changed to TCOM (*That* COMplementiser) in order to distinguish it from other word categories that are also tagged IN in the Penn Treebank, *i.e.* prepositions. The relative pronoun is also assigned a new tag—TREL (*That* RELative)—as the Penn classifies it as a *wh-* determiner, which is not without creating confusion as regards the position in the sentence. RB is kept for adverbial forms of *that*. Finally, DT is used for determiner forms and TPRON is used for pro-forms, allowing for easy retrieval and comparisons of both *this* and *that*. The efficiency of our queries has resulted in spotting 903 mistagged tokens (an 8.67% error rate).

5.1.1.3 Training TreeTagger with a new tagset and tagging

In this section, we cover the tagging accuracy after the modification of the Penn Treebank. Training is carried out on subsets 02 to 22 which make up the 950,000 tokens of the training sample (the other subsets are traditionally kept for development and final testing). With the modified training sample of the corpus, it is possible to start the training phase of TreeTagger. Training on the new version of

the annotated corpus lets the tagger learn all the tags and thus the new tags devised for *it*, *this* and *that*. A PERL script is used to extract the pairs of tokens (words or punctuation) and tags from the parsed structure of the corpus. We obtain a two-column table in which tokens come first (see Figure 20). This table corresponds to TreeTagger's format and its training module is run on the WSJ training sample.

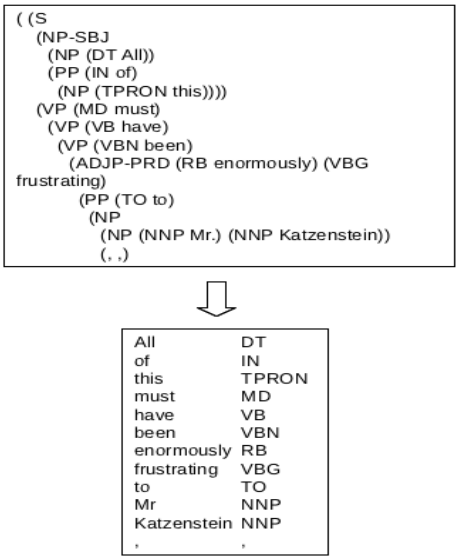

Figure 20: Extracting token-tag pairs from the parsed Penn Treebank corpus

Subset 24 (selected randomly) also undergoes tag modification for the purpose of preparing a gold standard test file in order to measure the accuracy of the tagging process. Two instances of this file are used. In the test file, the tags are stripped from the tokens to create a one-column file ready to be tagged. The second instance is the reference file that has resulted from the tag-modifying process based on the rules established in 5.1.1.2. As a result, several counts can be computed (see Annex G for the PERL script):

- The number of correctly assigned tags to a form (Nc)

- The number of tags automatically assigned to a specific form (Na)

- The actual number of correct tags to be found for a form (Nr)

Precision and recall correspond to the following formulae:

- Precision = Nc/Na
- Recall = Nc/Nr

The results for all types of occurrences of *this* and *that* are presented. As far as the test sample of the Penn Treebank is concerned, the global accuracy (overall number of correctly tagged form out of the total number of forms) is 96.03%. Out of the 32,853 tags, 31,550 were correctly assigned to the tokens. This figure is equivalent to the 96.34% accuracy reported in Schmid's experiment (Schmid 1994, Section 6). Modifying the tags does not seem to have any major impact on the global tagging accuracy of TreeTagger. When taking all forms of *this*, without distinguishing the tags, the global F-score for the form is 98.68%. *That* forms are correctly tagged in 89.81% of cases. The drop in tagging performance may be due to the greater variety of tags used for *that*, which introduces more uncertainty in TreeTagger's tag assignment process. The F-score for *it* is 90.08%, which may seem odd considering it may only be assigned two types of tags. Global observations give details on the tagging accuracy of all the tokens in the corpus. To find out more on the underlying trends of the tagging process, it is necessary to study the forms in detail and analyse the way tags are assigned to *it*, *this,* and *that.* Let us first consider *this* and *that* in relation to the DT and TPRON tags as they introduce a new distinction in the Penn Treebank tagset (see Table 26).

	Recall %	Precision %	F-Score %	Number of true occurrences
This DT	100	98.38	99.18	61
This TPRON	83.33	100	96.55	15
That DT	75	78.94	76.92	20
That TPRON	85.18	95.83	90.19	27

Table 26: Results after tagging *this* and *that* as determiners and pro-forms in the Penn Treebank

The first comment that can be made is about the contrasting results per tag. The DT tag fares better than the TPRON tag for occurrences of *this* whilst it is the opposite for *that*. As DT is mainly a tag characterised by its position in trigrams, it seems logical that it is better handled than TPRON which is more semantic in its characterisation. The fact that *that* DT is not tagged as accurately as *that* TPRON raises the question of possible conflicting tags. Examining the confusion matrix for the form in Table 28 reveals that the DT tag conflicts with the TCOM tag, which is not surprising as the complementiser[25] may frequently be followed by a noun just like DT. This trigram similarity introduces confusions for TreeTagger. All the details regarding tagging errors for each tag are shown in the two confusion matrices (see Tables 27 and 28).

This	DT	TPRON	RB
Tagged DT	61	1	0
Tagged TPRON	0	14	0
Tagged RB	0	0	0

Table 27: Confusion matrix for *this* in the Penn Treebank

That	DT	TPRON	TCOM	TREL	RB
Tagged DT	15	0	4	0	0
Tagged TPRON	0	23	0	1	0
Tagged TCOM	5	3	153	12	0
Tagged TREL	0	1	2	56	0
Tagged RB		0	0	0	0

Table 28: Confusion matrix for *that* in the Penn Treebank

As far as *it* is concerned, the high global accuracy level hides an underlying trend. The confusion matrix (see Table 29) shows that *it* PRP is handled very efficiently by TreeTagger. However, the PNR tag shows serious shortcomings as hardly any tag is correctly assigned. This may be due to several factors. First of all, in the training sample, the number of PNR tags is low in relation to the PRP tags which in turn introduces a bias in the probabilistic algorithm of the application. Secondly, PNR tags can be assigned when looking at broad contexts, which conflicts with TreeTagger's probabilistic approach that relies on trigrams.

[25] For instance, in the sentence “Put down *that* phone” (wsj_2425), *that* is tagged as TCOM instead of DT.

It	PRP	PNR
Tagged PRP	167	17
Tagged PNR	1	1

Table 29: Confusion matrix for *it* in the Penn Treebank

It appears that automated tagging of *it* is problematic. Conversely, the tagging process of *this* and *that* on the Penn Treebank tagset provides results that are satisfactory from the point of view of the linguist who seeks help in the annotating process. These mitigated results show that annotation can be first performed automatically before manual verification and correction of the tags that are assigned. Overall, this defines a semi-automatic method that ensures a good compromise between cost of labour and tagging accuracy. It also provides the consistency afforded by the algorithm.

Having tested the method on a native corpus, it is also important to test it on a learner corpus in order to determine the level of accuracy provided by TreeTagger, as re-trained with our categories. To do so, a test sample of the Diderot-LONGDALE corpus is taken. It is made up of 20,095 tokens (markup symbols and words). Similarly to the previous test, we have two instances of the test file. The first instance includes tokens with their tags that result from a manual annotation process. The second instance just includes the tokens before being tagged by TreeTagger. The objective is to see how the tags for the forms are handled. The following confusion matrices provide the details.

It	PRP	PNR
Tagged PRP	534	44
Tagged PNR	36	4

Table 30: Confusion matrix for *it* in the Diderot LONGDALE corpus

This	DT	TPRON	RB
Tagged DT	70	11	0
Tagged TPRON	0	15	0
Tagged RB	1	1	0

Table 31: Confusion matrix for *this* in the Diderot LONGDALE corpus

That	DT	TPRON	TCOM	TREL	RB
Tagged DT	5	2	2	0	1
Tagged TPRON	1	40	9	1	0
Tagged TCOM	2	31	70	21	2
Tagged TREL	0	18	5	8	0
Tagged RB	0	2	0	3	0

Table 32: Confusion matrix for *that* in the Diderot LONGDALE corpus

The annotation process shows more tagging errors than on the Penn Treebank but with various degrees depending on the form. *It* still appears very problematic as far as the new PNR tag is concerned. Only four out of 48 occurrences of *it* PNR are tagged correctly. Conversely, the *it* PRP tag shows more accuracy even though a substantial number of errors is reported. Concerning *this*, the DT tag is handled efficiently but the TPRON tag yields many errors (55.55% recall). The results for *that* also show tagging errors. The DT tag shows a low recall (62.5%) but the ratio is calculated on a low number of DT forms (8 occurrences), which induces broad variations. TPRON is also problematic (43.01% recall) and confusions occur mostly with the TCOM tag, which shows that the form is used by learners compared with natives in confusing configurations. As the tag assignment depends on trigram sequences, it shows that mode- or speaker-related factors impact on the syntactic construction of segments including *that*. As far as TCOM is concerned, recall is higher than precision (55.55%), which corroborates the previous statement. The tagger gives a lot of the correct forms but, when assigning tags, it gives the tag to non-corresponding forms. Errors are also reported on the TREL tag which is largely confused with the TCOM tag. This also shows that syntactic configurations are impacted by the mode or the type of speaker which results in tagging confusions. Regarding adverbial uses, it appears that they are non-existent in our data and, as such, their null frequency may characterise a learner criterial feature (Hawkins and Buttery 2010; Hawkins and Filipović 2012). All in all, these results show that a semi-automatic approach to functional tagging is appropriate as it yields a faster annotating process whilst ensuring tag verification. These results also show that comparing learner corpora with reference corpora, such as the WSJ, is subject to difficulties due to a drop in tagging accuracy for learner corpora.

5.1.2 Implementing the positional annotation layer

Positional annotation is carried out automatically by implementing a deterministic algorithm that uses TreeTagger's PoS-functional annotation to detect nominative and oblique cases for each occurrence of the three forms. The most striking examples involve subject and non-subject positions for the opposition of *it* and *that* in anaphoric chains. For the sake of readability, we reproduce such an example already commented in Chapter 3: “He discovered that she had slept with several other boyfriends before him. *That* shocked him a good deal, and they had a quarrel about *it*” (Stirling and Huddleston 2002, 1507). We mostly exploit this opposition in the guise of a marked tagging, where subject position is assigned the 'unmarked' nominative, as it is the generic case, and the rest assigned a 'marked' oblique feature, due to the cognitive complexity it implies: “As a theoretical construct, markedness presupposes the notion of formal complexity, whereby the marked is structurally more complex and the unmarked more simple." (Givón 1995, 25). We define oblique tokens as those in which the form is not a nominative case, which includes the object position of a noun phrase and adverbial subordinated clauses such as “I went to Morocco *this summer*”. The operationalisation of the positional annotation algorithm that distinguishes subject position from alternate positions is presented underneath:

- For any token, if the token is a *this*, *that* or *it* form, initiate the positional value of the form to Not Applicable (NA).
    - If the form is assigned the TPRON or PRP tag
        - check if the form is part of an interrogative structure. Check if there is a modal prior to the form or if the form belongs to a BE/DO + form + (adverb) + adjective or verb. In these cases assign NOMI, otherwise OBLI.
        - check if the next two tags correspond to verbs or modals. In this case assign the NOMI positional value to the form, otherwise assign OBLI.
    - If the form is assigned the DT tag as part of a noun phrase, *i.e.* a DT followed by an noun or an adjective and a noun,
        - check if the following forms of *that* NP are verbs or modals. In these cases, the attribute value is NOMI, otherwise it is equal to OBLI.

- check if the form is part of an interrogative structure. Check if there is a modal prior to the noun phrase or if the form belongs to a BE/DO + noun phrase + (adverb) + adjective or verb structure. In these cases assign NOMI, otherwise OBLI.

The PERL script in Table 33 is the program dedicated to the elucidation of nominative cases, *i.e.* cases in which the forms are subjects of verbs. In accordance with our typology of positions, oblique cases correspond to all other cases where the forms are not subjects. The program unfolds in three conditional stages. Firstly, any of the three tokens (singular and plural forms) is searched in a corpus file. When matched, the syntactic position is initialised as *NA* for "Not Applicable". This is to cover all cases of the forms including those which are neither determiners nor pro-forms. The second conditional stage focuses on matching the pronominal forms of the matched forms. These are checked with their right context and if they are followed by either a verb or a modal, including an adverb in between, then the syntactic position is designated as subject (NOMI line 16). In any other case, the position assignment is oblique (OBLI line 19). The third conditional stage deals with the determiner forms of the demonstratives. If they are matched as such, that is, if they are followed by a noun or an adjective and then a noun, the grammatical category of the subsequent form is checked for verb or modal. Be it the case, the *this* or *that* determiner is assigned a subject position. Of course, it is not in itself a subject but it is part of a noun phrase that functions as such and it may be relevant for analytical purposes to have this level of distinction in which subject and object NPs can be distinguished. In case there is neither verb nor modal after the NP, the oblique tag (OBLI line 38) is assigned to the form.

It must be mentioned that in oral corpora, repetitions are quite frequent especially for pronouns and pro-forms as they are used to initiate sentences. As exemplified in Figure 30 page 245, the first *it* is repeated. In order to capture the first *it* as a nominative case, the program includes a fall-back solution where the following two elements can also be one of the same grammar category and then a verb or modal. The program allows the first *it* to be assigned the subject tag even though, at first glance, it is followed by another pronoun. Similarly, another fall-back solution is

implemented to correct PoS tagging errors on other forms than *it*, *this* and *that* (which have been manually checked as mentioned in Section 5.1.1.1). Exploration via a concordancer revealed that some *it* forms may be followed by forms erroneously tagged as POS for possessive case or NNS for plural noun. For instance, in the following utterance: “it's really hard actually but it it'*s* ok” (DID0014-S001), the second copula is tagged POS even though it is not a genitive case. The algorithm makes room for these errors that correspond to actual verb forms (see lines 12 and 31). The algorithm presented in Table 33 is inserted into a loop which ensures its application to each occurrence of either of the forms.

```
1 #if token is this, that or it and if it is a pro-form then
2 if (($tokens[$i] =~/^[T|t]his$/ ) or ($tokens[$i] =~/^[T|t]hat$/) or
  ($tokens[$i] =~/^[I|i]t$/) or ($tokens[$i] =~/^[T|t]hese$/) or /^[T|
  t]hose$/) {
3             my $position = "NA";
4             # check if it is followed by a modal or a verb or by an
  adverb or repeated pro-form and a verb or modal.
5           if ($tags[$i] =~ /TPRON|PRP/) {
6                     if ( #Question construction
7                               ($tags[$i-1] =~ /MD/)
8                               or (($tags[$i-1] =~ /^do$|^does$|^am$|
  ^are$|^is$|^was$|^were$/) and (
9                                                       ($tags[$i+1] =~ /^JJ$|
  ^VB$|^VBG$/)
10                                                                  or
  (($tags[$i+1] =~ /^RB$/) and ($tags[$i+2] =~ /^|^VB$|^VBG$/)))
11                                        or ($tags[$i+1] =~ /V.*|MD|NNS|
  POS/) #Affirmative or negative construction - NNS and POS is added to
  correct Treetagger's tagging errors on 's and third person singular
  verbs.
12                                        or (($tags[$i+1] =~ /RB|TPRON|
  PRP/) and ($tags[$i+2] =~ /V.*|MD|NNS|POS/) #TPRON and PRP are added
  to encompass repetitions in the case of oral expression
13                                                      )
14                                        )
15                              ) {
16                              $position = "NOMI";
17                              }
18                              else {
19                              $position="OBLI";
20                              }
21                    }
22       #Check the form is a DT
23                    elsif (($tags[$i] =~ /DT/) and (($tags[$i+1] =~
  /NN.*/)
24
           or (($tags[$i+1] =~ /JJ|DT/) and ($tags[$i+2] =~ /NN.*/))))
  {
```

```
25                              if (#Question construction
26                                      ($tags[$i-1] =~ /MD/)
27                                      or (($tags[$i-1] =~ /^do$|
  ^does$|^am$|^are$|^is$|^was$|^were$/) and (
28                                                      ($tags[$i+3]
  =~ /^JJ$|^VB$|^VBG$/)
29                                                      or
  (($tags[$i+3] =~ /^RB$/) and ($tags[$i+4] =~ /^|^VB$|^VBG$/)))
30                                      or ($tags[$i+2] =~ /V.*|MD|NNS|
  POS/) #Affirmative or negative construction
31                                      or (($tags[$i+2] =~ /RB|TPRON|
  PRP/) and ($tags[$i+3] =~ /V.*|MD|NNS|POS/)
32                                                      )
33                                      )
34                              ) {
35                              $position = "NOMI";
36                              }
37                              else {
38                              $position="OBLI";
39                              }
40                      }
```

Table 33: Extract of sequencing PERL program dedicated to the deterministic identification of nominative cases of *it*, *this* and *that*

Automation is, of course, not without flaws and accuracy measurements of the algorithm are presented in the next few lines. To compute precision and recall, two random samples are extracted from the three corpora (these samples are also used in the distributional analysis presented in Chapter 6). The first one includes 108 occurrences of *this* and *that* as determiners or pro-forms and the second one includes occurrences of *it*, *this* and *that* as pro-forms. Each sample comprises 36 occurrences of each corpus (NOCE, Diderot-LONGDALE and Penn Treebank WSJ). The contingency tables are presented underneath together with accuracy metrics.

	Actual POSITION		
Predicted position	NOMI	OBLI	Totals
NA	3	2	5
NOMI	40	6	46
OBLI	2	55	57
Totals	45	63	108

Table 34: Contingency table for positional tag assignment in sample 1

	NOMI	OBLI
Precision	0,8695652174	0,9649122807
Recall	0,8888888889	0,873015873

Table 35: Precision and recall for *NOMI* and *OBLI* tags in sample 1

	Actual POSITION		
Predicted POSITION	NOMI	OBLI	Totals
NOMI	73	5	78
OBLI	3	27	30
Totals	76	32	108

Table 36: Contingency table for positional tag assignment in sample 2

	NOMI	OBLI
Precision	0,9358974359	0,9
Recall	0,9605263158	0,84375

Table 37: Precision and recall for *NOMI* and *OBLI* tags in sample 2

The results show a rather high accuracy in the tagging process. Due to specific shortcomings in the determiner part of the algorithm (figures not included) or unexpected text configurations such as pauses or transcribers' comments, some forms are not detected as determiners and are assigned the NA tag (see Table 34). Overall, the algorithm fares well on both samples, which indicates that the pro-forms and determiner forms are detected correctly. Nevertheless, due to the number of errors, the automation requires a manual verification. For each sample used in the studies presented in Chapters 6 and 7, manual corrections are introduced so as to ensure the highest degree of accuracy. The automation greatly accelerates the manual process.

5.2 A multi-layer XML structure

In this section, the construction of the XML structure of each corpus is covered. Structuring the annotations into a set of layers separated into distinct files linked to each other via internal XML links allows a clear distinction of the annotation types. In Section 5.2.1, the NITE XML structure is reviewed. In section 5.2.2, the focus is placed on how the annotated corpora are encapsulated in their XML structure and Section 5.2.3 explains how queries can be applied within such a structure.

5.2.1 A multi-level file structure

As described in Section 4.2.2, the objective is to create several layers of annotation for the three corpora described in 4.2.1. It is relevant to say that this work is part of a larger framework for the CLILLAC-ARP research team. Members of the team at the University of Paris-Diderot aim to develop phonological layers of annotation for the corpus. Research is being carried out in the directions of suprasegmental and phonetic features of learner language. The annotation proposed in our work is a partial contribution to the overall project. The fact is that the annotated corpus will result in many annotation layers which are not all developed at the same time and by the same people, not to mention other annotations not yet imagined for this corpus. Considering this heterogeneity in the annotating process, it seems difficult to try to have all annotations placed in a fixed annotation scheme. A fixed scheme would include all annotation layers for each recording and further addition of annotation layers might turn out to be difficult. What is therefore necessary, is to find an architectural design that allows any annotation layer to be added to an existing consistent structure without incurring the cost of restructuring the entire corpus.

The multi-level approach proposed by Carletta *et al*. appears as a solution to the problem of adding many kinds of annotation to the same basic data and relating them together (Carletta *et al.* 2003, 1). The NITE object model—implemented in the NITE XML Toolkit (NXT hereafter) (Carletta *et al.* 2006)—is specifically designed to allow any number of annotation layers to be added to an existing corpus by aligning it with either the time or the text units that are its basic components. In actual fact, the design corresponds to a set of XML files, each including one annotation layer and some information on how it relates to other files. Each file is organised in a tree-like manner in which data elements are the parents of other data elements. Each data element possesses attributes to either characterise the element or to link it up with another element. The NITE NXT software package comes with a set of examples. One of them corresponds to the processing of a one-sentence “corpus”. Reading the corresponding files helps reveal the functioning of the NITE model. There are several files for this corpus as each

file corresponds to one annotation layer called a coding. For the sake of simplification in explaining the functioning principles, Table 38 shows two of the files that compose the "corpus" structure: The syntax file, named *o1.syntax.xml,* provides a view of syntactic trees of the corpus whilst the word file, *o1.words.xml,* gives details on the word units. In the syntax file, the syntactic trees are structured with XML tags whose names are those extracted from the parsed Penn Treebank. Each tag includes a set of attributes such as *nite:id* or *hlem*. For instance, the line `<np nite:id="np_1" hlem="man">` describes an NP whose unique identification number is *np_1* and whose lemma is *man*. The elements that are dominated by the NP correspond to the word units. The principle of the NITE object model is to avoid mixing layers. As such, the word units do not appear in the syntax file. What appears though, is links to the other file, namely *o1.words.xml*.

Before detailing the word file, it is important to focus on the method used to create the links between the NP information placed in one file and the word unit information placed in another one. As can be seen in *o1.syntax.xml* (Table 38), the *np* tag dominates two other *nite:child* identical tags with different *href* attribute values. This attribute is what allows the link with the linguistic information on words that are stored in *o1.words.xml*. This file includes details on each word unit of the corpus and its PoS. The *word* tag is identified with its *nite:id* attribute-value pair and includes an *orth* attribute-value pair for the orthography. It also gives the PoS of the word. To summarise, syntactic parsing annotation and PoS annotation are split into two files, which are linked to each other via the special *nite:child.* This linking system ensures annotation layer interoperability whilst conserving the initial hierarchy of the annotated corpus in which one layer draws children from the previous one. In this case, word units are children of the syntactic constituents.

File: o1.syntax.xml	**File: o1.words.xml**

`<?xml version="1.0" encoding="ISO-8859-1" standalone="yes"?>`
`<nite:stream nite:id="syntax_str_1"`
`nite:content="set"`
`xmlns:nite="http://nite.sourceforge.net/"`
`>`
	`<?xml version="1.0"`
	`encoding="ISO-8859-1"`
	`standalone="yes"?>`
	`<nite:stream`
	`nite:id="word_str_1"`
	`xmlns:nite="http://nite.source`

```
   <s nite:id="s_1">
      <np nite:id="np_1" hlem="man">
         <nite:child
href="o1.words.xml#id('w_1')"/>
         <nite:child
href="o1.words.xml#id('w_2')"/>
      </np>
      <vp nite:id="vp_1" hlem="buy">
         <nite:child
href="o1.words.xml#id('w_3')"/>
         <np nite:id="np_2" hlem="toy">
            <nite:child
href="o1.words.xml#id('w_4')"/>
            <nite:child
href="o1.words.xml#id('w_5')"/>
         </np>
         <pp nite:id="pp_1" prep="for"
hlem="child">
            <nite:child
href="o1.words.xml#id('w_6')"/>
            <np nite:id="np_3"
hlem="child">
               <nite:child
href="o1.words.xml#id('w_7')"/>
               <nite:child
href="o1.words.xml#id('w_8')"/>
            </np>
         </pp>
      </vp>
   </s>
</nite:stream>
```

```
forge.net/">
   <word nite:id="w_1"
nite:start="0.0"
nite:end="0.2" orth="the"
pos="DT"/>
   <word nite:id="w_2"
nite:start="0.2"
nite:end="0.7" orth="man"
pos="NN"/>
   <word nite:id="w_3"
nite:start="0.7"
nite:end="0.9" orth="bought"
pos="VBD"/>
   <word nite:id="w_4"
nite:start="1.0"
nite:end="1.4" orth="these"
pos="CD"/>
   <word nite:id="w_5"
nite:start="1.4"
nite:end="1.7" orth="toys"
pos="NNS"/>
   <word nite:id="w_6"
nite:start="1.9"
nite:end="2.1" orth="for"
pos="IN"/>
   <word nite:id="w_7"
nite:start="2.1"
nite:end="2.3" orth="his"
pos="PP$">
      <nite:pointer
role="ANTECEDENT"
href="o1.syntax.xml#id('np_1'
)"/>
   </word>
   <word nite:id="w_8"
nite:start="2.3"
nite:end="3.0" orth="children"
pos="NNS"/>
</nite:stream>
```

Table 38: File structure for the NITE object model

There is another type of link between the two files and it provides coreferential information. As opposed to the previous type of link that focuses on parent-child relationships, this link connects the word units to their textual antecedents. To do so, the *nite:pointer* tag is used in o1.words.xml. The tag is embedded within the word tag in the following simplified manner:

```
<word>
<nite:pointer>
</word>
```

This micro-structure ensures both word and coreferential annotations for the same element as shown in the following excerpt from o1.words.xml in Table 38.

```
<word nite:id="w_7" nite:start="2.1" nite:end="2.3" orth="his" pos="PP$">
<nite:pointer role="ANTECEDENT" href="o1.syntax.xml#id('np_1')"/>
</word>
```

Let us look at the exact syntax used to code coreference. The coreferential information is coded within the *nite:pointer* element which links the word file to the syntax file where the coreferring entity is located. The *nite:pointer* tag includes two types of information: one attribute-value pair gives the role of the link, *i.e.* antecedent, and the other one gives the path to the file and the identification of the data element which is the textual antecedent. This is different from the previously mentioned parent-child link in which the word is pointed at from the syntax file with the *nite:child* tag.

The corpus is composed of many annotation layers—called 'codings' in NITE's terminology—of which syntax and words are only two instances. Other files include prosody and gesture annotation. Consequently, there are as many files as there are annotation layers and it is essential to have a procedure that ensures the coordination of the files so that they can all be loaded for queries. For this purpose, a metadata file, which describes the data set, is also created (Carletta *et al.* 2006, 11). The code in Table 39 is an extract of the code used in the metadata file for the aforementioned one-sentence “corpus” example. Its description provides an understanding of its function. In fact, it acts as an index for the NITE XML Toolkit software applications. The Toolkit is made up of several applications devoted to corpus annotation and search. Searching a corpus with the toolkit is carried out with the NXT Search application. In order to load the data set spread across several files, the metadata file is opened, which triggers the loading of all the other files it points to within its code. In short, the metadata file lists the dependencies of the file structure of a corpus.

The file named *single-sentence-metadata.xml* includes a coding part that documents the different annotation layers used in the corpus (for a complete view of the file

see Annex H). All the codings are placed between the *codings* tag which is a parent of several *coding-file* tags that define each coding or annotation layer. For instance, the syntactic parsing annotation layer is tagged as a *coding-file* with a *name= "syntax"* attribute-value pair and is the parent of three types of children tagged as either *structural-layer, time-aligned-layer or featural-layer.* The example only shows the first two types. Lines 14 to 17 provide the characterisation of the syntax layer made up of one code that takes *s* as a name. This layer has a lower dependency called the 'phrase layer' also defined in a *structural-layer* tag (Lines 18 to 29). This layer has a set of codes that are listed with attribute-value pairs such as `<code name="vp">`. These codes are characterised by the attribute tag that gives further possible values, such as lemma, in `<attribute name="hlem" value-type="string"/>`

The second annotation layer or coding corresponds to word units and is characterised from line 31 to 43 in Table 39. This layer is time aligned as its tag suggests. Its possible attributes such as orthography and PoS are listed and specified comprehensively so that the possible values of these attributes are all listed in the file, *e.g.* CC and CD.

File: single-sentence-metadata.xml

```
1. <?xml version="1.0" encoding="UTF-8"?>
2. <!-- <!DOCTYPE corpus SYSTEM "meta-standoff.dtd"> -->
3.
4. <!-- NXT METADATA EXAMPLE FOR A STANDOFF CORPUS JEAN CARLETTA AND
   JONATHAN KILGOUR
5.      ADAPTED FOR STEFAN EVERT'S EXAMPLE CORPUS 3/9/2
6. -->
7. <corpus description="Test Corpus" id="single-sentence"
   links="ltxml1" type="standoff">
8.     <!-- GENERIC CORPUS INFORMATION -->
9.  ...
10.             <!-- CODINGS -->
11.             <codings path="../xml/SingleSentence">
12.                 <interaction-codings>
13.                  ...
14.                     <coding-file name="syntax">
15.                         <structural-layer name="syntax-layer"
   draws-children-from="phrase-layer">
16.                             <code name="s"/>
17.                         </structural-layer>
```

```
18.                         <structural-layer name="phrase-layer"
   recursive-draws-children-from="words-layer">
19.                             <code name="vp">
20.                                 <attribute name="hlem" value-
   type="string"/>
21.                             </code>
22.                             <code name="np">
23.                                 <attribute name="hlem" value-
   type="string"/>
24.                             </code>
25.                             <code name="pp">
26.                                 <attribute name="hlem" value-
   type="string"/>
27.                                 <attribute name="prep" value-
   type="string"/>
28.                             </code>
29.                         </structural-layer>
30.                     </coding-file>
31.                     <coding-file name="words">
32.                         <time-aligned-layer name="words-layer">
33.                             <code name="word">
34.                                 <attribute name="orth" value-
   type="string"/>
35.                                 <attribute name="pos" value-
   type="enumerated">
36.                                     <value>CC</value>
37.                                     <value>CD</value>
38.                                    ...
39.                                 </attribute>
40.                                 <pointer number="1"
   role="ANTECEDENT" target="phrase-layer"/>
41.                             </code>
42.                         </time-aligned-layer>
43.                     </coding-file>
44.                 </interaction-codings>
45.             </codings>
46.
47.         ...
48.
49.             <observations>
50.                 <observation name="o1"/>
51.             </observations>
52.         </corpus>
```

Table 39: Extract of XML code used to specify the structure of annotations in the NITE object model

In short, each annotation layer corresponds to a coding whose attributes and values are all described in the metadata file. Opening this file gives NXT all the necessary information to process all the other corpus files since all the various data elements are defined here. There are several benefits to this multi-level structure for corpora. It allows "the representation of both time-aligned and hierarchically annotated

data" and parents "can inherit time information from their children" (Heid et al. 2004, 1457). Another important advantage is its support for intersecting hierarchies. NXT's structural relationships are very flexible thanks to the use of independent non-congruent structural annotations (Calhoun et al. 2010, 389). Finally, the separation of annotation into multiple files allows different people to create unrelated annotations with a view to merging their output at a later stage. This latter argument fits the profile of the work carried out on the Diderot-LONGDALE corpus by team members and even its broader extension which is the LONGDALE corpus.

5.2.2 Implementations of the NITE object model on three corpora

In this section, we cover the implementation of the NITE object model on three corpora. First of all, we show how two NITE XML files are created from the files resulting from the annotation phase. We then describe the structure of the metadata file that needs to be created in order to read the files from the NXT application software.

The file structure that mirrors the corpus depends on the annotations which are planned for implementation. Section 4.2.2 describes the types of annotation which are necessary for the analysis of *this* and *that*. For the file implementation, we decide to focus on the PoS-functional, the positional and the contextual annotations. The first two are conducted semi-automatically and for the third one, the endophoric and exophoric tags are added manually. The NITE object model deals with annotation layers under the concept of coding. Therefore, the PoS-functional and the positional annotation layers are to be placed within one file and the contextual annotation layer corresponds to another file. All together, the two files include the following information: orthography, PoS, position and context-related information. This information is encapsulated in two codings. The *words* coding includes orthography, PoS-function and position annotations, and the *context* coding only includes endophoric and exophoric details. For each coding a

file is created. Both files are governed by a metadata file that documents the structure.

The construction of the coding files is carried out automatically from the TreeTagger-formatted corpus output files. At this stage in the construction of the setup, it is important to mention that the data take the form of three-column format files (see Figure 21). These files result from a two-fold annotation phase centred on PoS and context annotations. Firstly, TreeTagger has been used to PoS tag the corpus files (including the functional distinctions for *it*, *this* and *that*) on the basis of one token tag pair per line. Secondly, for each *it* pronoun (PRP tag) and *this/that* pro-form or determiner (TPRON and DT tags), a context-related tag has been added. In accordance with the possible retrieval contexts (see Section 2.3.3), endophoric and exophoric annotation (ENDO/EXO tags) is added to the lines which include an *it* pronoun or a *this* or *that* pro-form. The output files come in a three-column format.

<A>	SYM	
okay	JJ	
and	CC	
what	WP	
do	VBP	
you	PRP	
think	VBP	
about	RB	
that	TPRON	ENDO
</A>	SYM	
<B>	SYM	
about	RB	
that	TPRON	ENDO
I	PRP	
think	VBP	
well	RB	
-LRB-	-LRB-	
er	UH	
-RRB-	-RRB-	
I	PRP	
think	VBP	
it	PRP	ENDO
's	VBZ	
-LRB-	-LRB-	
er	UH	

Figure 21: PoS and semantic annotation of a corpus in a three-column format

Two PERL scripts (see Annexes I and J for full details, only extracts of code are commented hereinafter) are used to convert the set of three-column files into two files (*o1.words.xml* and *o1.context.xml*) by merging them and appending the right XML coding for each of the tokens according to its column. One script deals with tokens, PoS and position and places them in a NITE model compliant XML structure whilst the other script constructs a set of files dedicated to the contextual annotation (ENDO and EXO tags). In each of the programs, a common part is dedicated to placing the data into what are called arrays in PERL. The principle is that of a table consisting of three columns just like in Figure 21 but which includes indexes. These indexes allow the programmer to determine which position in the table to focus on and, thus, what to do with the values.

The following is the PERL extract used to handle the data in arrays. It is inspired from (Tanguy and Hathout 2007)

```
while (my $line = <ENTREE>) {
chomp $line;
        my @array = split ( /\t/, $line );
          push ( @tokens, $array[0] );
          push ( @tags, $array[1] );
          push ( @context, $array[2] );
}
```

Figure 22: PERL extract on how to split TreeTagger-type files into PERL arrays/tables

What it essentially does is that for each line of the TreeTagger-type files shown in Figure 21, the line is split into three and the elements are placed in a table-like manner in an array called *array*. Then each column of the array is pushed into three different arrays. The first array deals with tokens. The second one stores all the PoS tags and the last one the context tags. The result is that for any given index of the three arrays, the retrieved elements correspond to the actual line in the file. That way, it is possible to process each line of each file of each corpus.

In the PERL script dedicated to the PoS and positional annotation layers (see an extract in Figure 23), the tokens and PoS are placed in a tree-based structure where divisions between speakers appear as branches. Words can be seen as their leaves. It is thus possible to extract words from one or other speaker. After much thinking, pauses are not used to split potential utterances since they do not necessarily lead to a new utterance and rather mark hesitations within the same utterance. A conservative approach is thus preferred, *i.e.* only tags for clear indisputable speech divisions are used and switches in speakers appear as the only safe markers in an oral corpus. As well as PoS, a positional attribute is also implemented in this script. The positional algorithm described in Section 5.1.2 is introduced and checks the close context of each form to determine whether it is located in a nominative or oblique case. Figure 23 shows an extract of the script. This program is inspired from (Tanguy and Hathout 2007, 345). A few comments are required to explain how the data are manipulated. First, in lines 1 to 3, the XML::Writer module is used since it ensures the production of well-formed XML documents. This module

allows the creation of an object which is stored in the *$scripteur* variable. This object embarks the methods to write XML code with correct tag elements. Start tags are created for the *nite:stream* and *txt* elements. The *$scripteur* object is used all throughout the program whenever an XML line needs to be produced after certain conditions are fulfilled. This program also includes the algorithm (partially presented from lines 4 to 8) dedicated to the positional values of the forms.

After the positional value has been computed (see Section 5.1.2 for the implementation script), the $scripteur object is called on to produce the desired XML code which include attributes whose values derive from PERL arrays in which all words and PoS tags have been previously placed in the program. If the form is matched, the position attribute is assigned a value and the $position variable is one of the arguments of $scripteur (see line 9). If the form is not *it*, *this* or *that*, no positional attribute is created in the XML word tag (see line 14). The XML module also handles the creation of the *nite:child* tag by creating a *href* attribute whose value is the address of the file in which the context attribute can be found (Calhoun *et al.* 2010, 391-394). The file path, lines 10 and 15, includes a first, hard-coded part as it is a constant: `o1.context.xml#`. The second part of the name is variable as it includes a specific identification: `id(".$r."-ctxt-".$i.")`. The `$r` element refers to the file name in which the contextual annotation can be found. The `$i` variable corresponds to the counter in the loop defined line 4, and its counts coincide with those used for the contextual annotation, hereby ensuring that the word and PoS annotation layer is aligned with the contextual annotation layer.

```
1 my $scripteur = new XML::Writer ( DATA_MODE=>1, DATA_INDENT=>1,
  OUTPUT=>\*SORTIE);
2 #$scripteur->startTag( "nite:stream", "nite:id"=>"word_str_1",
  "xmlns:nite"=>"http://nite.sourceforge.net/");
3 $scripteur->startTag( "txt", "nite:id"=>$1);
4 #position algorithm immplementation
5 for (my $i=0; $i <= $#tokens; $i++){
6 #capture any token except speakers' tags.
7 if (( $tokens[$i]=~ /^(.*)$/ ) and ( $tokens[$i] !~/<A>|<B>|<\/A>|
  <\/B>/ )) {
8 [...]
9           $scripteur->startTag( "word", "nite:id"=>$r."-".$i ,
  "orth"=>$tokens[$i], "pos"=>$tags[$i], "position"=>$position);
```

```
10                    $scripteur->dataElement( "nite:child", "",
  "href"=>"o1.context.xml#id(".$r."-ctxt-".$i.")");
11                    $scripteur->endTag("word");
12          }
13          else {
14                    $scripteur->startTag( "word", "nite:id"=>$r."-".$i
  , "orth"=>$tokens[$i], "pos"=>$tags[$i]);
15                    $scripteur->dataElement( "nite:child", "",
  "href"=>"o1.context.xml#id(".$r."-ctxt-".$i.")");
16                    $scripteur->endTag("word");
17                    }
18 }
19 if ($tokens[$i] =~/<A>|<B>/) {
20                              $ut_id++;
21                    $scripteur ->startTag ("speaker",
  "nite:id"=>$r."ut".$ut_id, "agent"=>$tokens[$i]);
22                    }
23 if ( $tokens[$i] =~/<\/A>|<\/B>/ ){
24                    $scripteur->endTag( "speaker" );
25                    }
26          }
1 $scripteur->endTag("txt" );
```

Figure 23: Extract of PERL script to convert the PoS-functional and positional annotations from Treetagger-type to XML files

The second script (see Figure 24) places the ENDO and EXO tags in a NITE model compliant structure. The extract of the PERL program which determines how this is done requires some comments. Firstly, a loop is created to move down the PERL array in which the Treetagger-formatted data has been placed. It mirrors the data structure shown in Figure 21. Secondly, the XML::Writer module is called via *$scripteur* for the production of an XML compliant data structure. At this stage, the context tag is created with a *nite:id* and an identification number that corresponds to the position in the array. Incidentally, this position is the same as the one assigned to the tokens and PoS tags in the other script. This ensures the linking of each contextual annotation tag with its token.

```
[…]
1 for (my $i=0; $i <= $#tokens; $i++){
2
3 if ( $context[$i]=~ /^(.*)$/ ){
4 $scripteur->startTag( "context", "nite:id"=>$r."-ctxt-".$i,
  "context"=>$context[$i] );
5
6 $scripteur->endTag( "context");
7           }
8 }
9 $scripteur->endTag( "txt" );
[…]
```

Figure 24: Extract of PERL script to convert contextual annotation from TreeTagger-type to XML files

The resulting two files are interconnected. Table 40 provides the details of both files. For context, a simple structure includes the context tag with a *context* attribute-value pair. For the *words* xml file, the *word* tag dominates the *nite:child* tag which provides the access path to the context located in the *o1.context.xml* file. For instance, the word tag number 34, whose *orth* attribute is *it*, and whose PoS attribute is PRP, in *o1.words.xml* (in bold characters) has a nite:child that refers to element 34 in 01.context.xml. This element 34 includes the context attribute which takes the value ENDO.

o1.words.xml	**o1.context.xml**
<nite:stream nite:id="word_str_1" xmlns:nite="http://nite.sourceforge.net/"> […] <speaker nite:id="DID0014-S001ut3" agent="<A>"> <word nite:id="DID0014-S001-28" orth="can" pos="MD"> <nite:child href="o1.context.xml#id(DID0014-S001-ctxt-28)"></nite:child> </word> <word nite:id="DID0014-S001-29" orth="you" pos="PRP"> <nite:child href="o1.context.xml#id(DID0014-S001-ctxt-29)"></nite:child> </word> <word nite:id="DID0014-S001-30" orth="tell" pos="VB"> <nite:child href="o1.context.xml#id(DID0014-	<?xml version="1.0" encoding="UTF-8"?> <nite:stream nite:id="context_str_1" xmlns:nite="http://nite.sourceforge.net/"> <txt nite:id="DID0014-S001"> […] <context nite:id="DID0014-S001-ctxt-32" context=""></context> <context nite:id="DID0014-S001-ctxt-33" context=""></context> **<context nite:id="DID0014-S001-ctxt-34" context="ENDO"></context>** […]

S001-ctxt-30)"></nite:child> </word> <word nite:id="DID0014-S001-31" orth="me" pos="PRP"> <nite:child href="o1.context.xml#id(DID0014-S001-ctxt-31)"></nite:child> </word> <word nite:id="DID0014-S001-32" orth="something" pos="NN"> <nite:child href="o1.context.xml#id(DID0014-S001-ctxt-32)"></nite:child> </word> <word nite:id="DID0014-S001-33" orth="about" pos="IN"> <nite:child href="o1.context.xml#id(DID0014-S001-ctxt-33)"></nite:child> </word> **<word nite:id="DID0014-S001-34" orth="it" pos="PRP" position="OBLI"> <nite:child href="o1.context.xml#id(DID0014-S001-ctxt-34)"></nite:child> </word>** </speaker> […]	

Table 40: Extract of annotation files of the Diderot-LONGDALE corpus formatted according to the NITE Model

The metadata file (see Table 41) is written manually, for it is a file, limited in size, that describes the XML structure used in the annotation files. For each of them, a coding is created with all the necessary details and it describes the hierarchy of the tags. In the context file, the structural-layer element is used in order to show dominances between tags, *i.e.* <*txt*> dominates <*context*> and the latter is composed of attributes and a pointer element whose target value is the *words* coding. In other words, the *nite:pointer* tag ensures the relationship between codings as it defines the reference to the *words* coding. The *words* coding is also defined in the form of a hierarchy with three elements in the following manner:

```
<txt>
     <speaker>
          <word>
```

This hierarchical structure means that each word belongs to a specific speaker and is located in a particular text. Attribute-value pairs are defined and all PoS attribute values are listed. This is important because when NXT is launched and the application loads the various codings thanks to the description of the structure, all possible attribute values must be defined in order to be interpreted later.

o1.metadata.xml

```
1   <!-- CODINGS -->
2     <codings path="">
3         <interaction-codings>
4
5    <coding-file name="context">
6                 <structural-layer name="text-layer" draws-children-
  from="context-layer">
7           <code name="txt"/>
8               </structural-layer>
9           <structural-layer name="context-layer">
10                      <code name="context">
11                          <attribute name="target" value-
  type="string"/>
12                          <attribute name="context" value-
  type="enumerated">
13                              <value>ENDO</value>
14                              <value>EXO</value>
15                                  </attribute>
16                             <pointer number="1" role="TYPE"
  target="words"/>
17                      </code>
18                  </structural-layer>
19             </coding-file>
20
21
22             <coding-file name="words">
23                <structural-layer name="text-layer" draws-children-
  from="speaker-layer">
24                    <code name="txt"/>
25                </structural-layer>
26                 <structural-layer name="speaker-layer" draws-children-
  from="words-layer">
27                    <code name="speaker">
28                             <attribute name="agent" value-
  type="string"/>
29                    </code>
30                </structural-layer>
31
32                  <featural-layer name="words-layer" draws-children-
  from="context-layer">
33                      <code name="word">
34                          <attribute name="orth" value-type="string"/>
35                          <attribute name="pos" value-
  type="enumerated">
```

```
36                                <value>CC</value>
37                                <value>CD</value>
38                                <value>DT</value>
39 ...
40                                <value>VBN</value>
41                                <value>VBP</value>
42                                <value>VBZ</value>
43                                <value>WDT</value>
44                                <value>WP</value>
45                                <value>WP$</value>
46                                <value>WRB</value>
47                            </attribute>
48                        </code>
49                    </featural-layer>
50                </coding-file>
51            </interaction-codings>
52        </codings>
53
54     <observations>
55            <observation name="o1"/>
56        </observations>
57 </corpus>
```

Table 41: Extract of specific coding used for the annotation of a learner corpus compliant with the NITE model

One last aspect must also be clarified, that is how file names are constructed. From Table 41, we can see that two *coding-file* elements are defined and their *name* attributes are used in the actual file names they point to. NXT automatically adds the observation name, given further down in the metadata, in first position and the XML extension to recompose each file name. For instance, the *<coding-file>* element name is appended to the *<observation>* tag name to form the *o1.words* name followed by the XML extension. When loading, NXT uses this to find the *o1.words.xml* file. The result is a three-file structure controlled by the metadata file and annotation layers are encoded in two independent files. Queries can be applied across both interoperable layers thanks to the relationships that are defined within the XML structure. This process is applied to the Diderot-LONGDALE corpus and, since it is mostly automated, it can be deployed on any other TreeTagger-tagged corpora. Our contribution to the research team's efforts comes in the form of an XML NITE compliant annotated version of the Diderot-LONGDALE corpus.

We apologise for the level of technicality in these pages but one of the objectives of our research is to compare learners of different L1s with each other and with natives. That is the reason why several corpora are used in our experimental setup. This means that many files are created and exploring them requires the explanation of their naming structure including their *coding-file* elements. After their annotation, these corpora must be processed to convert their multi-column format into a NITE-compliant XML structure. The process is very similar to that used for the Diderot-LONGDALE corpus except for the way speech utterances are split. In the spoken Diderot-LONGDALE corpus, the transcripts are split when speakers change. In the NOCE and the Penn Treebank, which are written corpora, students' texts are split at the two strong punctuation marks which are the full stop and the question mark.

Table 42 shows the sentence: “After all, this isn't old money […]”. Each token has been placed within a *word* XML element. As well as its unique NITE identification, two other attribute-value pairs are used for the words and for the part of speech including the functional PoS tag assigned to *this*. All *word* elements are placed within an *utterance* element. The move from one utterance element to another is operated via the punctuation. When a full stop or a question mark is encountered, the utterance tag is closed and a new one is reopened.

Each word element embeds a *nite:child* element whose *href* attribute-value pair provides the exact location of the context-related annotation for that word. In the case of *this*, the contextual annotation is to be found in the *wsj.context.xml* file in which the wsj-ctxt-1344 context element can be found. Table 43 shows the excerpt of the context file and the previously identified element appears within a context XML element. This element contains a context attribute with a tag indicating the endophoric value of the form referred to.

wsj.words.xml

```
<utterance nite:id="wsjut57">
  <word nite:id="wsj-1341" orth="After" pos="IN">
   <nite:child href="wsj.context.xml#id(wsj-ctxt-1341)"></nite:child>
  </word>
  <word nite:id="wsj-1342" orth="all" pos="DT">
   <nite:child href="wsj.context.xml#id(wsj-ctxt-1342)"></nite:child>
  </word>
  <word nite:id="wsj-1343" orth="," pos=",">
   <nite:child href="wsj.context.xml#id(wsj-ctxt-1343)"></nite:child>
  </word>
  <word nite:id="wsj-1344" orth="this" pos="TPRON">
   <nite:child href="wsj.context.xml#id(wsj-ctxt-1344)"></nite:child>
  </word>
  <word nite:id="wsj-1345" orth="is" pos="VBZ">
   <nite:child href="wsj.context.xml#id(wsj-ctxt-1345)"></nite:child>
  </word>
  <word nite:id="wsj-1346" orth="n't" pos="RB">
   <nite:child href="wsj.context.xml#id(wsj-ctxt-1346)"></nite:child>
  </word>
  <word nite:id="wsj-1347" orth="old" pos="JJ">
   <nite:child href="wsj.context.xml#id(wsj-ctxt-1347)"></nite:child>
  </word>
  <word nite:id="wsj-1348" orth="money" pos="NN">
   <nite:child href="wsj.context.xml#id(wsj-ctxt-1348)"></nite:child>
  </word>
[...]
  <word nite:id="wsj-1361" orth="." pos=".">
   <nite:child href="wsj.context.xml#id(wsj-ctxt-1361)"></nite:child>
  </word>
 </utterance>
```

Table 42: Extract of WSJ subset of Penn Treebank converted into a word layer of NITE XML format

wsj.context.xml

```
<txt nite:id="wsj">
[…]
<context nite:id="wsj-ctxt-1344" context="ENDO"></context>
[…]
</txt>
```

Table 43: Extract of WSJ subset of Penn Treebank converted into a context annotation layer of NITE XML format

As a result of the conversion processes, all three corpora are formatted according to the NITE model. Each corpus is split into two XML files in which all annotation information is encapsulated. These files can then be loaded into NXT Search via a metadata file which acts as an index of the structure. Queries can be written to retrieve any number of any combination of elements.

5.2.3 Querying occurrences with NXT Search

As already mentioned, the reason for structuring the data into a set of layers separated into distinct files and yet linked to each other via internal XML links is that it allows a clear distinction of the data types. Yet, it is possible to write cross-level queries that search all layers and therefore display results that combine several types of constraints regarding annotation. With the use of NXT's search module it is possible to explore any NITE-compliant corpus by extracting contexts that contain the forms corresponding to a combination of criteria. The interesting aspect is that these criteria can span several annotation layers as opposed to some concordancer software applications that can only provide results related to word tokens. NXT Search is an application that can do just that thanks to its query language and the XML structure applied to corpora. The query language, called NQL, relies on a syntax that uses the XML elements as variables making it possible to retrieve elements of any annotation layer and thus to combine them all in a single query. An export of the query results is also possible. Thanks to this export functionality, results can be saved in a spreadsheet for subsequent manipulation and statistical analysis.

Figure 25 is an example of a query that retrieves all occurrences of nominative *this* pronouns that are tagged as exophoric forms. The query syntax works in a two-fold manner as explained in (Carletta *et al.* 2006, 10). There is a first line of variable declarations separated from the subsequent lines with a colon. The lines determine the constraints that apply to the various XML elements which are present in the various files that encode each annotation layer. Constraints in lines two and three specify the dominance between the elements. In other terms, the *speaker* element dominates the *word* element which, in turn, dominates the *context* element. In the next five lines, attribute-value pairs are set with either regular expression patterns or simple text. The last line indicates another query[26] which is applied to the results of the first query. This query ensures the capture of a speaker's full utterance for each matched form of the first query. It allows the user to see the form in context.

[26] This query is a contribution from Jonathan Kilgour from the University of Edinburgh.

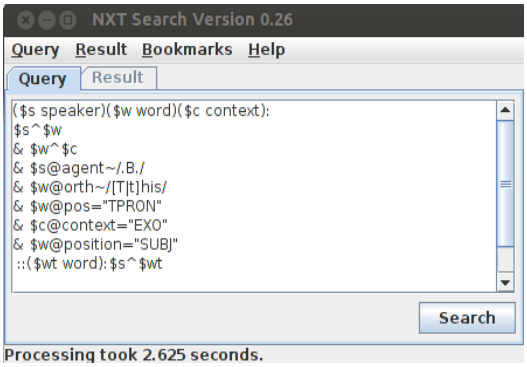

Figure 25: NXT query for the retrieval of all forms of *this* pronouns which are tagged as exophoric and produced by speaker B in a subject position

After a query has been written in the query tab window of NXT Search, its results are read in the result tab window (see Figure 26). This window is divided into two parts. The upper one provides the list of matches and the lower one gives the details of the matches. In the list of matches it is possible to expand them individually. Details are then given regarding the exact reference of each constraint that has been matched in the query. For instance, the line:

```
<nite:pointer role = “$c” xlink:href=“o1.contex.xml#id(DID0034-S003-ctxt-686)”/>
```

corresponds to the context variable that takes the EXO value in the DID0034-S003 recording and whose line identification is 686. Equivalent information is provided in the lower part of the window in a table view. However, more contextual details appear, as it is easy for the user to see the form and its matching constraints within its close context. The *@agent* column indicates the speaker, *B* being the learner. The *@orth* column gives the actual transcript extract of the conversation. *@pos* informs the user on the PoS-function tag assigned to the matching form (TPRON in this case). It also gives the PoS of the transcript's sequence. *@position* gives the information on the nominative (SUBJ) or the oblique (OBJ) case of each *this*, *that* and *it* forms. Finally, the type of context, *i.e.* endophoric or exophoric appears in the *@context* column.

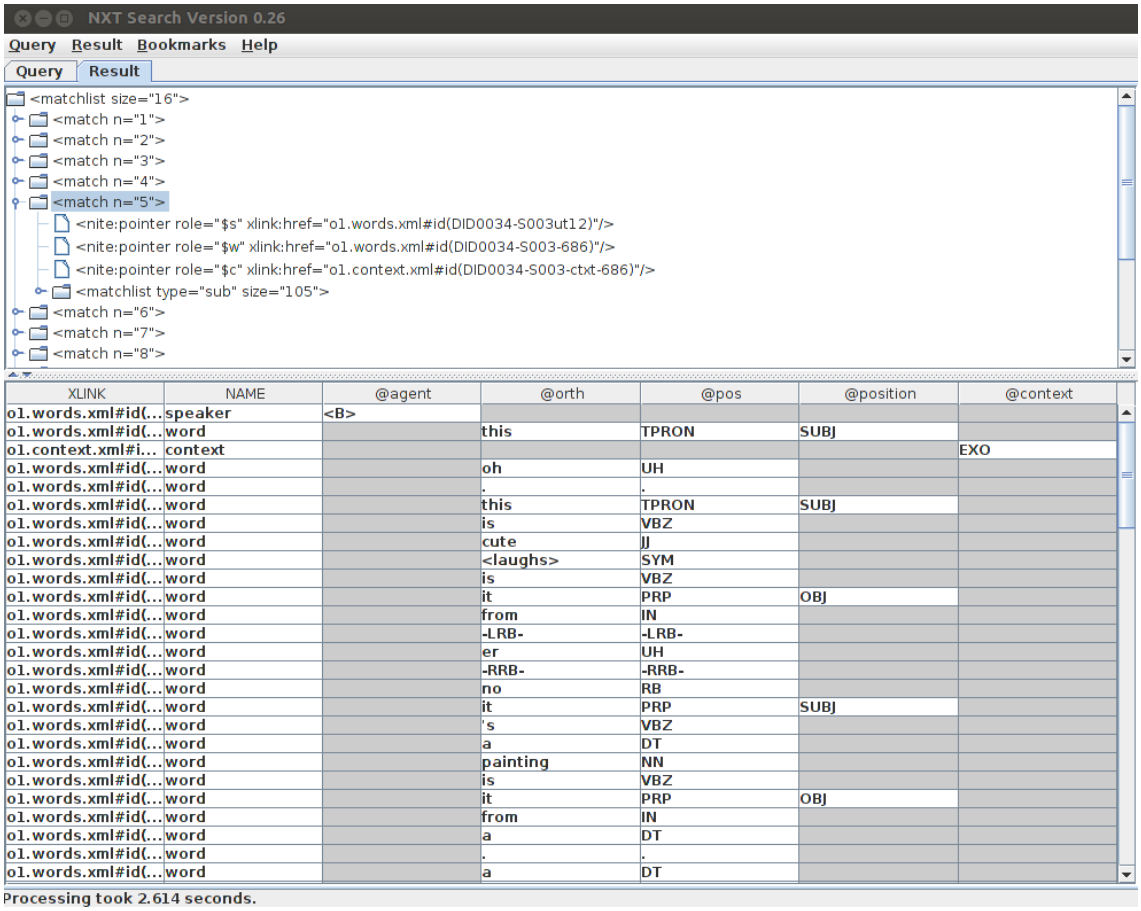

Figure 26: NXT result window after querying for exophoric *this* pronouns

Therefore, NXT Search is a tool that allows the user to query several layers of annotation at the same time. The query language's strength is its ability to combine all annotation layers in simple syntax. This relies on the NITE model that finds its roots in distinct but interrelated annotation files. The architecture allows for complex in-depth structures with cross-references between different layers and NXT search provides an easy access to that. One limit, though, is that it does not contain conditionals and loops and thus prevents the writing of queries in which conditional precedence is examined. Nominative and oblique cases would need such options to be queried if they were not encoded as annotation.

The structure also allows multi-corpus queries. NXT Search scans what is called an 'observation' which is made up of several files, one for each annotation layer. By adding each annotation layer of each corpus to its corresponding file, the observation includes all the corpora, and query results will include elements whose

unique identification number indicates the source corpus. For instance, in the upper part of the window of NXT search result tab of Figure 26, match number 5 includes a pointer that embarks the exact source files, hence referring to the Diderot-LONGDALE corpus whose files all start with the three letters DID (see Section 4.2.1 for a description of the file names).

5.3 Sequencing data for distributional analysis

The previous two sections have helped understand how the three annotated corpora are placed into identical data structures which support queries of the same form. These queries help retrieve specific utterances from the corpora but they do not support sampling methods used in machine learning approaches. In this section, we describe what we call 'data sequencing'. We seek to determine tables of features that characterise contexts in which the forms can be found. These sets of features are organised in tables used in regression analysis for example. In Chapters 6 and 7, two approaches will rely on data sequences to provide results in the domain of distributional analysis and learner error detection. The following paragraphs are devoted to the way data is processed prior to its use for distributional analysis.

5.3.1 Feature sequences for memory-based classification or prediction of the forms

In this section, we explain the principle of sequencing corpora. Firstly, we state the specifications of the features which are sequenced. Secondly, we explain the working principle of feature sequencing. Finally, we present how the sequencing principle is put into operation.

5.3.1.1 Specifications and explanations of the features

The choice of features plays an important role in the sequencing process and it is important to detail the reasons for their selection. We first discuss general features

that are common to native and non-native English. Secondly, we cover interlanguage-related features as they are thought to indicate degrees of grammaticality.

5.3.1.1.1 Common features for the forms

The purpose is to convert every occurrence of *it*, *this* or *that* in a line of features so that the resulting matrix can be used for any analysis of these forms in context. First, the PoS category of the form is extracted and placed in the line as a first feature. It is followed by a set of features that are added is the 3-gram environment of each form. In other terms, the 3 PoS tags and tokens to the left and the right of each form are extracted and displayed as features. Another feature is the syntactic position as it is a basic grammatical distinction concerning pro-forms and determiner forms part of an NP (see Section 2.5.1). This distinction is endorsed as an extra feature which is added with a specific algorithm presented in Section 5.1.2. This algorithm is used in several dedicated programs including the NITE XML formatting program presented in Section 5.2.2. The positional feature can therefore take two values: either it is OBLI for oblique cases of the forms or it is NOMI for nominative cases of the forms. One other group of syntagmatic features created in the sequencing process is related to the use of the forms within constructions that include prepositions. Prepositions can be found as introductory forms of *this*, *that* (as determiners or pro-forms) and *it*. They can also be found right after one of these pro-forms or the NP in which they are as determiners. Prepositions are part of the syntactic process of utterance construction in which the position of the form is decided. This constitutes a feature whose role may depend on the form. Two features with pairs of values are created: PREPINT or 0 for the introductory position and PREPPOST or 0 for the post position of the preposition.

Another kind of feature which is introduced in the sequencing process is related to the semantic field, in the sense that the features intend to reflect semantic aspects of the forms in their context. As explained in Sections 2.2.2 and 3.3.1, the

endophoric and exophoric distinction in the use of *this* and *that* seems to be significant (but debatable as to its level of significance) in their selection. It is therefore paramount to introduce it as a feature for its analysis. The feature may take the endophoric, exophoric or 0 values. 0 corresponds to these forms that are neither determiner nor pro-forms, *e.g.* relative pronoun *that,* complementiser *that,* non-referential *it*. To do so, the PERL array that includes all the context tags is searched in parallel with the forms. When a form *this*, *that* or *it* is matched, its ENDO or EXO tag is collected and printed on the line of features that corresponds to that form.

There are also discourse related features which denote semantic choices made in the act of utterance. In Section 2.5.2, it was shown that the use of *this* and *that* produces meaning effects on utterances. It was also noted that their use is closely linked to that of *it*, including in referential chains (see Section 2.5.1). As much as it matters in native English, it also plays a role in learners. However, learners' interlanguage may be impacted by these meaning effects as they may misuse the forms. To prepare a study of the phenomenon, it is relevant to capture text and PoS features which are traces of such discourse issues and which may indicate different degrees of grammaticality.

5.3.1.1.2 Interlanguage-related features

Learner use of language is characterised by non-native-like uses of some markers. These markers are surrounded by speech features which can co-occur with them and indicate possible errors, for example. Learners' errors and other aspects of interlanguage can be characterised by indicators such as the frequencies of co-occurrences of these various speech features. These indicators can be used as potential predictors of use of a given marker and they give indications on the likelihood of how the combination of several speech elements may impact the actual use of a marker. This approach draws from the idea of a probabilistic grammar in the wake of Bresnan with the prediction of dative alternation in natives

(see Bresnan and Nikitina 2003, Bresnan *et al.* 2007) and that of Gries with his work on particle placement (Gries 2003). Similar work on learner language was also carried out by Tono by linking PoS information and learner levels (Tono 2013). By extracting relevant phenomena in the form of speech elements, we intend to account for learner productions and their use of *it, this* and *that*. In this section, we focus on the description of the features. What follows exemplifies some of our linguistic variables in our analysis. More often than not, our variables correspond to a binary representation: presence or absence of a given linguistic marker. We give some examples of our features.

One of the meaning effects in the use of the forms is detachment and it can be achieved via the act of temporal rupture with the moment of utterance as in:

54) anyhow (er) . Greenwich Village was very nice (er) I remember spending a lot of time at (er) the . this bookstore called Barnes and Nobles . and . *those* were . like it was like huge in the middle of Greenwich Village (DID0155-S001)

In the utterance the learner uses *those* with the past form of *be.* The pro-form comes after the use of *this* as a determiner and before *it*. This chain shows some kind of hesitation in the selection and the past tense might be playing a role in the cognitive process. Therefore, it appears important to have a feature that indicates the past tense of verbs used in the surrounding of the forms. In the sequencing process, the past simple feature is created with a pair of two possible values ('-' for null or ED) coding the presence or absence of the tense. Conversely, the present tense and, more specifically the BE+ING aspect, reflects the synchronisation between the moment of utterance and the predication referred to. There is identification between the moment of utterance in which the speaker is positioned and the moment of the predication referred to. The predication is located within the speaker's sphere (see Section 2.5.2). The following learner example shows that the use of the form with the ING form is not straightforward:

55) I always liked history so (er) it was really interesting but I really didn't like the linguistic parts and the grammar I'm not really good at this because I I do it like I feel it and if you start to say adjectives adverbs (er) I'm kind of lost when I do it I know I can do it but I I'm not really good in technique . so *that*'s getting worse in the in the second year (mm) (DID0146-S002)

In this example, the learner chooses *that* which might be considered as odd partly due to the use of the present continuous tense. Should a present tense verb appear in the close context (BE+ING or present simple), the VBZ PoS tag is used as a feature value, 0 being the other possible value. This constitutes what could be called 'the present tense feature'.

Another utterance-related feature that has semantic meaning is the one related to negation. The notion of speaker's sphere can also be realised by the use of negation in speech as in the following native example:

56) Learner: but it's nothing we: like . the year before where there there were Deep Purple and (er) . and Manu Chao and (er) there were was so much crowd I couldn't go I couldn't even seen the the big screen. Native: oh *that*'s not so fun (DID0162-S001)

In this example, the native denies the notion of fun and the selection of *that* endorses this rejection. A learner may have issues with this. In other terms, the existence of *not* in the surrounding of the form suggests a possible rejection of an entity by the speaker. As an illustration, a search on *not* as a collocate of *this* and *that* on the ICE-GB corpus shows differences. By looking at the 3-gram context of each form, an association measure is computed with AntConc[27]. The Mutual Information (MI) score gives the probability of occurrence of each form near *not* relative to how many times they occur in ICE-GB. There are more complex metrics (Stefanowitsch and Gries 2003) for association measures, but even a simple one as MI (mutual information) shows this. The score obtained for a search on all the ICE-GB text files gives the following results: the *that-not* association (3.16453) is higher than for the *this-not* association (3.01552). This indicates that in native English

[27] In AntConc, there are two metrics for association measures, we have used the z-score. AntConc is a concordancer available at http://www.laurenceanthony.net/software.html (Last access March 31, 2016)

there is a stronger association of *that* with *not* than *this* with *not*. Interlanguage might show different patterns and to capture this phenomenon, we create a feature whose values are NOT or null.

The semantic value of the forms can also be traced back into the actual discourse structure employed by the speaker. The choice of using a form at the start of an utterance or statement may suggest a refocusing process by which the speaker might place the entity referred to in the foreground of his/her discourse. This phenomenon may occur in after a coordinating conjunction as exemplified in the following:

57) he's walking in the street and there's is (em) super music (em) in the (em) back in back and (em) there's You Make my Dream Come True something like Hall and Oates and *that*'s a movie of the eighties and it's very (em) joyful. (DID0121-S001)

In the example, the learner uses *that* as a support for a statement on the previously mentioned movie and the learner's intent may be to place the entity in the foreground. To capture such a semantic value, three features are collected within the annotated corpus. The existence of any strong punctuation, any coordinating conjunction or simply a capital letter on the form itself, may all point to the start of an utterance or statement. Three features encapsulate this with pairs of values: CC or 0, PUNC or 0, CAP or 0.

Finally, the notion of endophoric anaphora (see Sections 2.3.3 and 3.3.1) also plays a semantic role and it is important to reflect it within the choice of features. In such referential processes, entities are referred to by way of noun phrases, proper nouns, previously used personal pronouns or pro-forms and *wh* pronouns. All these categories indicate a reference to a discourse entity which may or may not be the one referred to by *it, this* or *that*. Nevertheless, the existence of an entity has an influence on the predicational context and the choice of the anaphor (see Section 2.4.1). In the following example the pronoun *I* is used just after the use of *this*.

58) [...] I write it I write it down on the paper and then . during the night or another day I take this paper and I and I look and I say well it's a good idea so I just (er) . go (er) further in my (er) idea . *this* is how *I* do (DID0161-S002)

In this case, it might be argued that the choice of *this* is partly compelled by the personal pronoun. Learners may demonstrate a specific trend and that is the reason why it is relevant to capture this kind of element. Consequently, it is important to create specific features that indicate reference. Three such features are created in order to include the distinction between the nature of the anaphors. Nouns, pronouns and *wh* pronouns lead to three features with pair value attributes: REFNN or 0 for the noun feature, REFPRON or 0 for the pronoun feature and REFWH or 0 for the *wh* pronoun feature.

The aforementioned features are summarised in the following table.

Feature description	Feature values (PoS tags[28])
Verb in the preterite form in order to mark temporal distantiation within the context	Null or ED*
The present tense and, more specifically the continuous aspect, reflects the synchronisation between the moment of utterance and the predication referred to.	Null or VBZ
Negation in the surrounding context to suggest a possible rejection of an entity by the speaker.	Null or NOT*
Strong punctuation in order to mark possible changes of focus, reference or topic	Null or PUNC*
Coordinating conjunction in order to mark possible changes of focus, reference or topic	Null or CC
First letter in upper case of a word in order to mark possible changes of focus, reference or topic	Null or CAP*
Personal pronouns in order to mark reference to a discourse entity, e.g. the speakers' existence	Null or REFPRON*
Nouns in order to mark reference to a discourse entity, e.g. the possible coreference of the demonstrative and the noun	Null or REFNN*
Wh- pronouns in order to mark reference to a discourse entity	Null or REWH

Table 44: Features related to interlanguage

To sum up, we have retained some linguistic properties that we believe to be relevant in our analysis. Chapter 7 will rank these features to show which are the

[28] Penn Treebank scheme except when there is an asterisk.

most important ones. For this classification technique, we need a dataset where all the features are present for each token of our subsystem of markers (this is called a 'vector' of features). The preceding examples were based on the presence or absence of a given tag. The next section describes the making of this dataset, based on the automatic retrieval of the relevant contextual tags and specifically, the procedure used to exploit the PoS tag sequences corresponding to the context of the markers.

5.3.1.2 The principle of feature collection

The objective of the process is to convert actual annotated corpora into arbitrarily ordered sets of features that correspond to specific characteristics of contexts in which specific forms can be found. In other terms, after identifying a particular occurrence of a form, its textual context—what we call n-grams of tokens—is isolated and placed as a set of features such as one token to the left of the form, two tokens to the right of the form and so on. Following this principle, 3-grams to the left and the right of each form are used as features in the sequence. As well as tokens, functions of the forms can also be placed in the feature sequence. For instance, if the form is a determiner or a pro-form, DT and TPRON are relevant features to extract and to place into the sequence. PoS elements are also relevant features to collect in the same manner as token n-grams are sequenced. Finally, the context domain via which reference is carried out is also a feature worth collecting. The context feature can take two values in the sequence: endophoric and exophoric. There are two kinds of sequences to prepare, depending on their use. They can be used with memory-based learning algorithms for classification purposes (our program for Chapter 7) or they can be used for statistical analysis of the forms' distributions (Chapter 6).

The principle is that all *it, this* and *that* forms are processed no matter their PoS-function tag. For each occurrence, the specific features are taken from the annotated corpus texts to be placed in lines of features that are assigned one of the

three possible classes: *this*, *that* or *it*. These are used to feed a memory-learning algorithm that will compute statistics with the purpose of classification in relation to the features. Specific features are looked for in each annotated text in order to be placed in the sequences. The idea is to isolate specific text and annotation units from all other corpus elements and to place them in the same line as a set of features (the 'vector' of features'). The numerous lines form sequences used to test the forms in terms of classification performance. A PERL script scans the data and collects specific elements depending on constraints set in the program. For instance, if there is a verb form in the past simple tense in the preceding context of a *that* form, then it is collected as a feature. Similarly, other features are looked for in corpus-annotated data.

Globally, the process is that of a text conversion. In other terms, the input data is the three-column format that results from the annotation stage (see Figure 21 page 219) and the output is a set of lines, each including a number of features whose values are null or not. The PERL program includes a number of parts. The first part is identical to the one described in Section 5.2.1 in which we describe how the data is placed in a three-column table (see Figure 22 for the program extract). This table, also called an array, can be accessed to retrieve any word, PoS tag or context tag positioned anywhere in its structure. As a result, it is possible to search for specific textual elements that surround each *this, that* or *it* form in the annotated files. Each match is displayed in one line. So each line corresponds to all the searched elements that surround an *it, this* or *that* displayed as the last element of the line. Figure 27 shows the first four lines of the WSJ training subset of the Penn Treebank after such a sequencing process.

PRP	Cars	NNPS	Inc.	NNP	said	VBD	expects
	VBZ	its	PRP$	U.S.	NNP	ENDO	NOMI
	VBZ	ED	-	-	-	-	REFNN
	-	-	-	-	it		
DT	rule	NN	changes	NNS	proposed	VBD	past
	JJ	summer	NN	that	TREL	EXO	OBLI
	-	ED	-	-	-	-	REFNN
	-	REFWH	-	-	this		
TREL	this	DT	past	JJ	summer	NN	,
	,	among	IN	other	JJ	-	-
	-	-	-	-	-	-	-
	-	-	-	-	that		
TCOM	the	DT	letters	NNS	maintain	VBP	investor
	NN	confidence	NN	has	VBZ	-	-
	-	-	-	-	-	-	-
	-	-	-	-	that		

Figure 27: Extract of WSJ training subset of Penn Treebank after the sequencing process

Each line is not meaningful in itself (this is not a concordancer view) as it only represents the elements that are searched for in the surrounding of a form. Table 45 shows the example of the conversion of a sentence into a vector of features, which is in fact the second line displayed in Figure 27.

Annotated text sample			Sequenced sample (set of features and class in bold)				
[...]							
letters	NNS						
to	TO						
the	DT						
agency	NN						
about	IN						
rule	NN		DT	rule	NN	changes	NNS
changes	NNS			proposed	VBD	past	JJ
proposed	VBD			summer	NN	that	TREL
this	DT	EXO		EXO	OBLI	-	ED
past	JJ			-	-	-	-
summer	NN			REFNN	-	REFWH	-
that	TREL			-	**this**		
,	,						
among	IN						
other	JJ						
things	NNS						
,	,						
[...]							

Table 45 One sentence of the WSJ subset of the Penn Treebank prior to and after sequencing process

In the right-hand frame of the table, each symbol on the line corresponds to a feature. The DT tag corresponds to the PoS function of the *this* form placed at the

end of the line. The *rule*, *changes*, *proposed*, *past*, *summer* and *that* tokens correspond to the left and right token n-grams of the form. The NN, NNS, VBD, JJ, NN, TREL tags correspond to the left and right PoS n-grams of the form. The EXO tag is one of the three possible values of the context feature. OBLI is one the three possible values of the positional feature (NA, OBLI, NOMI). The series of –, -ED, REFNN, REFWH tags shows whether specific linguistic features are null or not (hence the hyphen sign for null).

5.3.1.3 Operationalisation of the principle

The aforementioned procedures ensure the conversion of corpus texts from their textual representation into a sequence of lines of features. The PERL script that takes charge of this process includes a loop that ensures that it is carried out iteratively. Figure 28 shows an extract of the script that guides the insertion of features. This extract shows how the -ED feature is handled. After checking whether the PoS tag of the *this* is a TPRON or a PRP (line 2), each PoS tag in the 3-gram is checked (line 3) so as to verify if it matches the value VBD (verb in the preterite in the Penn Treebank tagset). If it does, the variable $featED is assigned the value "ED" which is then printed as part of the list of features. This type of process can be repeated automatically on any number of corpus files and the features can be modified in order to create different sequences of features that can then be tested with the memory-based learning algorithm.

```
1 #If the form is a pro-form or a pronoun
2           if ($tags[$i] =~ /TPRON|PRP/) {
3 if (($tags[$i-3] =~ /VBD/) or ($tags[$i-2] =~  /VBD/) or ($tags[$i-
  1] =~ /VBD/) or ($tags[$i+3] =~ /VBD/) or ($tags[$i+2] =~  /VBD/)
  or ($tags[$i+1] =~ /VBD/)) {
4                     $featED = "ED";
5 }
```

Figure 28: Extract of PERL program in which the ED feature search is described

After assigning the right values to all the features of an instance, the instance is printed in the file to be used by the classifier (Figure 29 shows the part of the script dedicated to printing each instance of features and their class). The full PERL script

is presented in two versions in two annexes. One version (Annex K) outputs the data for R, a statistical software application, and the other one (Annex L) is to output the data for TiMBL, a machine-learning software application. Due to specific experiments carried out with these tools, there are two versions of the program.

```
print SORTIE  $tags[$i],
          "\t", $tokens[$i-3],"\t", $tags[$i-3],
          "\t", $tokens[$i-2],"\t", $tags[$i-2],
          "\t",$tokens[$i-1],"\t", $tags[$i-1],
          "\t", $tokens[$i+1],"\t", $tags[$i+1],
          "\t", $tokens[$i+2],"\t", $tags[$i+2],
          "\t", $tokens[$i+3],"\t", $tags[$i+3],
          "\t", $context[$i],"\t",$position,
          "\t", $featVBZ,"\t", $featED,"\t", $featNOT, "\t", $featCC,
"\t", $featCAP, "\t", $featPUNC,
          "\t", $featREFNN, "\t",$featREFPRON, "\t",$featREFWH, "\t",
$featPREPINT, "\t", $featPREPPOST,
          "\t",$tokens[$i],"\n";
```

Figure 29: Extract of sequencing PERL script dedicated to printing instances of features and their class

In principle, the sequencing process distinguishes itself from a concordancer text-based search. It does not give a list of words in context but rather it displays each occurrence with a set of features which are extracted from the co-text and its corresponding annotated information. All the elements that surround a form and that are linguistically relevant are systematically preselected. Each set or line of features corresponds to a combination that matches a form and the memory-based algorithm helps deal with all the possible combinations by learning them and extrapolating from them in order to predict which forms would correspond to new combinations. The setup, based on annotation and data sequences, enables the researcher to zoom into the texts to observe how specific linguistic items behave. In Section 7.1, we report an experiment in which various features are tested for their role in the selection process of pro-forms.

5.3.2 Feature sequences for regression analysis

This second type of set of features is very similar to the first one insofar as it corresponds to lines of features, but it is different as the lines do not include

classes. Instead, the lines represent the distribution of context and PoS features that surround a specific form. This set of features is not created for class predictions but rather to analyse which features (or variables) play a significant role in the observed distribution (see Chapter 6 for a regression analysis). By sequencing corpus data as lines of features, we can apply regression analysis in order to explore the complexity of the data. By displaying the corpus data in sequences of features, we pre-select a number of features whose significance needs to be tested.

Technically, the process is nearly identical to that described in 5.3.1. The only difference is that headers are added to serve as variable names in the regression analysis and that *it, this* and *that* are not used as classes. Instead, they are displayed as one variable among others, *i.e.* TOKENS in Figure 31. The annotated corpus files (see Section 5.1) need to be transformed so as to display the data for each form in each text. The principle is to have a sequential representation of the data rather than a textual one. The text is not looked at in its entirety but instead we extract a form and look at the textual and contextual elements that surround it. On each line, we obtain sets of elements that match a particular form. The total number of lines corresponds to a sequence that is a representation of the text which is centred on the forms. Figure 31 shows an extract of the resulting transformation of the utterance “so: . It's a it's this topic to me for me . And it” (the ':' mark indicates a lengthening of the vowel sound. During tokenisation, it was erroneously interpreted as a token) from the Diderot-LONGDALE corpus presented in Figure 30.

```
DID0035-S002            NNP
[...]
so          RB
:           :
.           .
it          PRP         ENDO
's          VBZ
a           DT
it          PRP         ENDO
's          VBZ
this        DT          ENDO
topic       NN
to          TO
me          PRP
for         IN
me          PRP
.           .
and         CC
it          PRP         ENDO
```

Figure 30: Extract of the Diderot-LONGDALE corpus in its initial annotated format

```
1.  DIDID                   TOKENS      TAGS        TOKENS3BEFORE
                TAGS3BEFORE             TOKENS2BEFORE       TAGS2BEFORE
                TOKENS1BEFORE           TAGS1BEFORE         TOKENS1AFTER
                TAGS1AFTER              TOKENS2AFTER        TAGS2AFTER
                TOKENS3AFTER            TAGS3AFTER          CONTEXT     POSITION
2.  [...]
3.  DID0035-S002.seq        it          PRP         so
                RB                      :                   :
                .                       .                   's
                VBZ                     a                   DT
                it                      PRP                 ENDO        NOMI
4.  DID0035-S002.seq        it          PRP         it
                PRP                     's                  VBZ
                a                       DT                  's
                VBZ                     this                DT
                topic                   NN                  ENDO        NOMI
5.  DID0035-S002.seq        this        DT          a
                DT                      it                  PRP
                's                      VBZ                 topic
                NN                      to                  TO
                me                      PRP                 ENDO        OBLI
    [...]
```

Figure 31: Sequence of linguistic items surrounding a form

In these two figures, we can see the transformational process that is at work as it results in a feature-sequence representation of the text rather than a representation of its syntagmatic order. This change in representation allows us to create lines of features which only display the relevant annotation and text units which depend on

the forms. All other textual and annotation elements are discarded. The feature-sequence representation in Figure 31 provides a focused view of the forms of interest, *i.e. it* and *this* in this case, as their textual and contextual environment is displayed in sets of features, hence forming sequences. This process is carried out by a PERL program in which the data are formatted so that the first line corresponds to the headers. When carrying out a distributional analysis, these headers will correspond to variables that might be tested in statistical models yielding significance levels for each one of them (see Chapter 6). So the data is represented as a set of columns, one of which being the *tokens* column. In this column, three values (two on the figure) may be found: *this*, *that*, *it*.

To understand the columns let us take line 3 as an example. The first element called DID0035-S002.seq corresponds to the source transcript of a specific occurrence. The DIDID column corresponds to the Diderot-LONGDALE identity. Then, the *it* token (in the TOKENS column) is the occurrence which has been extracted. It is followed by its functional realisation as pronoun labelled PRP (in the TAGS column). This will become useful in subsequent statistical stages which will require random samples of individuals and the file names will allow the identification of a form with an individual. Next, there is a series of token-PoS pairs which are taken in the 3-gram context of the *it* form. The ENDO (in the CONTEXT column) element provides information on the contextual value of the *it* form. Finally, the NOMI element shows that *it* is in a subject position. The position is computed in relation to other surrounding forms and their grammar categories and according to the algorithm described in Section 5.1.2.

As a result of the sequencing process, we obtain a sequential table of features that surround each form. All corpora can be converted into such tables, providing a view of the data that focus on *it*, *this* and *that*. Subsequent manipulations may filter out some of the columns for specific analysis. For instance, only *this* and *that* pro-forms and pronominal *it* could be selected together with their PoS-functional and

contextual features to support an analysis of the microsystem described in Section 3.2.2.2.1.

5.4 Summary

In this chapter, we demonstrate that proprietary data structures that impede corpus comparability can be overcome thanks to the implementation of an interoperable super-structure made of several layers of annotation. We show that the implementation of a a triple-layered annotation scheme on three corpora is possible. We demonstrate that the subsequent implementation of two specific types of data architecture shapes the corpora in ways that make them interoperable.

Concerning the implementation of the annotation scheme, the need was to modify the Penn Treebank tagset to tag the corpora with fine-grained tags for *it*, *this* and *that*. By using several software applications such as Tregex and TreeTagger, the three corpora have been tagged and every single occurrence of any of the three forms has been assigned a functional tag. We also have developed a specific PERL program dedicated to the assignment of positional tags. It implements a set of deterministic rules which rely on PoS tags to assign positional labels to the forms.

Concerning the implementation of the data architectures, we provide two programs. The first one helps convert TreeTagger file formats into a NITE XML structure. We show how this structure can subsequently be used within NXT Search to open a corpus and perform queries with constraints that span the three annotation layers to retrieve occurrences of *it*, *this* and *that*. The second program we provide is dedicated to converting corpus data into tables of sequences of linguistic features. These tables organised in columns of linguistic features can subsequently be imported into application programs that are used for multifactorial analysis.

All in all, two levels of interoperability have been included in the corpora. The use of an identical annotation scheme on three corpora gives room to comparable

query results. With NITE NXT Search, occurrences of the forms in their close context can be retrieved from any of the corpora, which means that utterances related to the pro-form microsystem can be compared. The use of identical data structures allows data from the different corpora to be mirrored against each other, which helps compare data across the corpora. By placing occurrences in instance-based tables, multi-factorial analysis can be carried out in order to identify the features that are significant in the pro-form microsystem, as exemplified in the next chapter.

Chapter 6 Statistical analysis of the distribution of *this* and *that* across learner and native corpora

In this chapter, we use the corpora in their newly annotated structures to explore the pro-form microsystem modelled in Section 3.3. The purpose is to use statistical modelling to study how learners of different L1s make use of the forms compared with natives. We use modelling techniques to identify the linguistic factors that play a role in the use of *this* and *that*. More specifically, we put to the test the hypothesis of the existence of the pro-form microsystem among learners. We try to see if evidence can be found in terms of parameters that control the system. Our analytical method follows three stages. Firstly, the need for data representativeness requires the randomised selection of a sample of occurrences from the corpora. Secondly, as a preliminary step, we need to compare the use of *this* and *that* forms across all corpora, regardless of their pro-form realisation, to see if the functional realisation—and any other features such as the context and the position—appear as significant factors in the selection of a form. Finally, based on results regarding the significance of functional realisations as a factor, we address the core issue of our thesis, *i.e.* the pro-form microsystem which includes *this* and *that* but also *it*. To look for evidence of this learner-specific microsystem and due to its tripartite nature (*this*, *that* and *it*), we need to carry out a multinomial regression analysis in order to explore the significance of several factors. By comparing occurrences of different L1s, we analyse learner errors to identify specific error types.

In Section 6.1, we justify and explain the randomised method applied to extract representative samples from the corpora. In Section 6.2, we analyse the two deictic forms regardless of their pro-form function in order to verify the significance of

factors such as the L1 and the functional realisation in the selection of the demonstratives. In Section 6.3, we test our thesis on the existence of the pro-form microsystem in learner English with a regression model. All the analyses presented in the sections take all three corpora into account, which allows comparisons between L1s, including native English.

6.1 Preparation of the data samples

Prior to explaining how the data samples are created, it is necessary to justify the sampling method. We have been conservative in our analysis and have respected the constraints of our statistical modelling techniques: to preserve the independence of observations, we cannot take into account all the occurrences but we need to sample them.

Extraction of the samples is done by following a stratified strategy to make sure that we obtain the same number of occurrences from each corpus. This strategy ensures an equal representation of the population types. In his description on sampling methods used for the LOB and the Brown corpora, Biber concludes that: "[...] stratified samples are almost always more representative than non-stratified samples (and they are never less representative)" (Biber 1993, 244). Gries also advocates for stratified strategies when it comes to sampling methods: "Adequate sampling strategies are of vital importance in this context. One possibility would be to use stratified sampling on the basis of top-down distinctions. Sampling for example, different speaker groups (*e.g.* as defined by perceived proficiency), different speakers, different genres, etc." (Gries 2008, 422). Each of the corpora used in our study corresponds to a speaker group specific in its L1, genre and mode. It is crucial to make sure that all are evenly represented in order to prevent one speaker group from outweighing the others.

The preparation of the samples requires adjustments for computation in R (R Core Team 2012)—an open source statistical analysis software application. The main objective is to compare the three corpora that have all been processed in sequences

of features as explained in Section 5.3.2. Each of the corpora is part of a file which can be imported into R for the creation of data frames. In this tabular representation of the data, each sampled token of the corpus is a line (an observation) and the relevant variables are the columns. Here, we describe how the three corpora are combined into one single data frame to provide a base for corpus comparison. Importing each previously created corpus file into R creates a data frame per corpus. Before combining the three data frames, the data of each corpus undergo a number of manipulations to ensure a streamlined structure. In order to create a unique identifier for each corpus, the CORPUS category is added as a new column with three different values: *did*, *noce* and *wsj* (See last column of Table 46). Because the singular/plural distinction is not taken into account in our analysis, all *these* and *those* values are converted into *this* and *that* values respectively. Also, all forms are converted to lowercase. Combining all three data frames results in a single, long table including all occurrences of the forms with their annotation feature values attached. Because the data of all corpora are located in a single structure they are now made comparable according to their features. Table 46 shows an extract of this data frame. The first line includes the headers of each column of the table. Each line consists in an occurrence of either *this, that* or *it* and its surrounding features.

	DIDID	TOKENS	TAGS	TOKENS3BEFORE	TAGS3BEFORE	TOKENS2BEFORE	TAGS2BEFORE	TOKENS1BEFORE	TAGS1BEFORE	TOKENS1AFTER	TAGS1AFTER	TOKENS2AFTER	TAGS2AFTER	TOKENS3AFTER	TAGS3AFTER	CONTEXT	POSITION	CORPUS
1411	DID0038-S001.seq4R	this	TPRON	if	IN	I	PRP	do	VBP	as	IN	a	DT	singer	NN	ENDO	OBLI	did
									...									
336	GR-B-1-160-5.seq4R	this	DT	are	VBP	transmited	JJ	by	IN	phones	NNS	.	.	As	IN	ENDO	OBLI	noce
									...									
68	wsj.test.seq4R	this	DT	Boys	NNPS	are	VBP	big	JJ	year	NN	in	IN	Budapest	NNP	EXO	OBLI	wsj

Table 46: Extract of the sample including occurrences from the three corpora

Two samples are created from this dataset. One sample is used for the analysis of *this* and *that* regardless of their function (see Case Study 1 in Section 6.2). The other one is used for the analysis of the forms as pro-forms only (see Case Study 2 in Section 6.3). The method is the same with different parameters applied. The data frame presented in Table 46 is filtered so as to retain only the occurrences whose categories are TPRON or DT (case study 1) and TPRON and PRP (case study 2).

To apply the stratified strategy on data frames, special functions in R are used to randomly and uniquely extract the same number of observations from each corpus. For Case Study 1, we want to analyse *this* and *that* in their determiner and pro-form functions. Exactly 36 determiner or pro-form occurrences are randomly extracted from each corpus. Once combined as a sample, there are 108 randomly selected occurrences of the two forms. For Case Study 2, we want to analyse pro-forms (including *it*) according to their L1. Exactly 36 pro-form occurrences of *it*, *this* and *that* are randomly extracted from each corpus. Once combined, they form the second sample also with 108 occurrences of the three forms too.

At this point, some comments must be made about the selection of the samples that emanate from a population composed of three corpora. It is important to mention that the Diderot-LONGDALE sample is also designed to reflect the longitudinality of the LONGDALE project and so the unique transcripts correspond to 12 unique speakers recorded three times at different points in time. This is a violation of the sample collection principle since individuals are supposed to be unique. However, since a lot of time was invested in the annotation process and its manual verification, it was judged beneficial to keep the longitudinal aspect in the sample. Further work on annotating more files would of course provide a more robust base for statistical analysis.

The WSJ sample does not include the distinction between individuals. Consequently, in terms of sample collection, this introduces another violation. The

reason for such a sampling is that, as far as we know, the text files of the WSJ articles do not include the authors' distinction, which makes it impossible to populate speakers' IDs across each corpus and, thus, across data frames. The requirement of the WSJ as a gold standard is explained in Section 4.2.1 but it may be recalled that it was favoured for its robustness in terms of parsing and its uniqueness in terms of genre. This latter point may also appear as a weakness when it comes to comparisons between corpora of different genres but it may not be so in our setup in which the single genre provides a good reference point for comparison with a learner corpus (Barlow 2005, 345). Recent research provides another argument that attenuates the problem of the variety in discourse types. In a series of statistical experiments on the Switchboard (oral) and the WSJ (written) corpora, Bresnan *et al.* noted little difference in dative alternation usage. The statistical model developed for the oral corpus applied with great accuracy to the written subset of the Penn Treebank (Bresnan *et al.* 2007, 97). Therefore, one question may be raised regarding probabilistic grammar—a grammar based on statistically predicted occurrences of forms—*i.e.* is it influenced by linguistic items and their environment rather than by types of register?

6.2 Case Study 1: The distribution of *this* and *that* in learner and native corpora: impacting factors

In this section, we analyse the correlation between *this* and *that*—as determiners or pro-forms—and other factors. Firstly, a general hypothesis is operationalised in which the forms' correlation with the speakers' L1s, regardless of any other factors, is analysed. Secondly, a multi-factorial approach is favoured to see if there is a correlation between the forms and other factors such as the speakers' L1s, the context, the function, and the position of the forms. The purpose is to see if some of these factors play a significant role in the selection of the determiners or pro-forms.

6.2.1 Forms with regard to native and non-native English

This is the most simple case scenario as it only involves observing how the dependent variable TOKENS (which corresponds to *this* or *that* in this case) evolves according to one categorical independent variable, that is CORPUS. In other terms, the aim is to see whether both variables are independent from each other. This model based on two variables is a simplified version of reality as it ignores other possible factors that may intervene in the selection of either *this* or *that*. Nevertheless, it may be considered as a first interesting step because the relationship between these two components is an abstraction of the differences in use of the forms depending on the L1 of the speaker or the written or spoken mode of the corpus. Testing this relationship provides a first answer as to whether or not the two elements vary independently from each other. The principle is to see if the proportions of the forms are the same from one corpus to the other or if they differ. This statement can be formulated as a statistical hypothesis such as:

- H0: The frequencies of the levels of the dependent variable TOKENS do not vary as a function of the levels of the independent variable CORPUS.

- H1: The frequencies of the levels of the dependent variable TOKENS vary as a function of the levels of the independent variable CORPUS.

After formulating the hypotheses, and by following Gries's procedure (Gries 2010, 179) for the analysis of such variables, the sample (as described in Section 6.1) is loaded. Counts of each form for each corpus category are operated, which gives the results presented in Table 47. One effect that can be read from these frequencies and, even more so, from the bar plot in Figure 32 is the variations in use. Speakers in the DID corpus, who are of French L1, use *this* and *that* just as much when speakers in the NOCE corpus, of Spanish L1, strongly prefer *this*. The native speaker WSJ standard shows a preference for *this* too but not to the same extent as Spanish learners. There seems to be a correlation between an L1 and the

use of the forms but a statistical test must be carried out to see if these results are significant.

	WSJ	NOCE	DID	Sum
That	9	7	17	33
This	27	29	19	75
Sum	36	36	36	108

Table 47: Observed frequencies of each form in each corpus

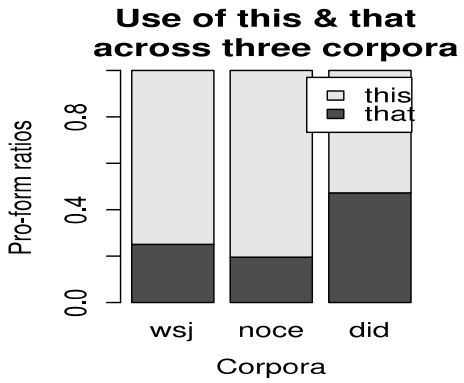

Figure 32: Proportions of the forms per corpus

The test that can be applied here is the Chi-squared test. It requires that all observations should be independent of each other (in this case all observations in the NOCE and the Diderot-LONGDALE corpora correspond to independent transcript files, and we have carefully sampled the corpora to obtain independent observations). Another assumption of this test is that 80% of all expected frequencies should be greater than 5 and that none of them should be null (Gries 2013 [2009], 179). Also, caution should be taken when the total for all cells is less than 50. Table 48 shows that the expected frequencies match the assumptions. The Chi-squared test of independence relies on the sum of the differences between the observed and expected values. Should the obtained value be different from 0 and should this difference be significant (p-value < 0.05), the statistic would indicate that TOKENS is dependent on CORPUS.

	WSJ	NOCE	DID	Sum
That	11	11	11	33
This	25	25	25	75
Sum	36	36	36	108

Table 48: Expected frequencies of each form in each corpus

The *chisq.test* function is used in R to compute the test with the following results:

```
Pearson's Chi-squared test
data: raw.totals
X-squared = 7.3309, df = 2, p-value = 0.02559
```

The degree of freedom indicator (*df*) shows that two categories (CORPUS and TOKENS) vary freely as the number of observations increases. *Df* is computed as follows:

df = (number of rows-1) * (number of columns-1) = (2-1) * (3-1) = 2

The p-value indicates a significant result. The proportions of the uses of *this* and *that* are not due to chance and they actually depend on the corpora, which is logical for at least two reasons. Firstly, the spoken and written modes introduce variability in the use of the forms that has already been documented in native English by Biber *et al.* (1999, 273, 349). Secondly, the native v. non-native distinction in the test suggests learner-dependent variability. In short, it seems that the frequencies of the forms vary not only with the native/non-native distinction but also with the L1 distinction. Spanish learners favour *this* whilst French speakers use both equally. This shows a correlation between the forms and the speakers' L1s which needs to be evaluated in terms of effect size in order to know how strong it is. Cramer's *V* is the correlation coefficient which gives this information by applying the following formula (Gries 2013 [2009], 186):

$\sqrt{\frac{\chi^2}{n(min[n\,rows, n\,columns]-1)}}$ where χ^2 is the Chi-squared value and *n* is the number of cells (without sums) in Table 47.

Still following Gries (2013 [2009]), the computation gives:

```
sqrt(test.proforms.determiners$statistic/sum(raw.totals)*(min(dim(raw.to
tals))-1))
V = 0.2605356
```

The effect is to be understood within the 0 to 1 range, 1 denoting perfect correlation. The 0.26 value shows a small enough effect so the correlation is not random, nor is it strong. (Cohen 2009) gives 0.3 as a medium value. With the same R function it is also possible to retrieve the expected values in order to compare them with the values presented in Table 47 and proceed to the computation of Pearson residuals. These quantify the magnitude of the difference between expected and observed values. Their square value quantifies contributions, *i.e.* which value in the table contributes most to the test result. Table 49 gives an insight into the combinations. How each value in the table contributes to the Chi-squared test is a matter of comparison with the value of the overall Chi-squared (χ^2). The larger the value, the larger the contribution of that value to the χ^2. Trends can be observed and *that* in the two learner corpora contributes mostly to the test which indicates learner variations compared with native speakers in the WSJ. *This* in the learner corpora also contributes to the Chi-squared, to a lesser extent than *that* though. It is interesting to note that the contributions of the forms in the native corpus are small, which could be interpreted as a low variation in use between expected and observed frequencies. This makes sense as native language is much less prone to variations than learner language is (see Section 3.1.1).

	WSJ	NOCE	DID
That	0.3636	1.4545	3.2727
This	0.1600	0.6400	1.4400

Table 49: Contributions to the Chi-squared test per form and per corpus

Table 50 on the Chi-squared test's residuals informs the reader about one extra indication regarding the effect of each intersecting value of the contributions: “The Pearson residuals […] reveal the direction of effect for each cell: negative and positive values mean that observed values are smaller and larger than the expected values respectively.” (Gries 2013 [2009], 187). So the variable combination that contributes most positively to the significant result of the test is *that* in the Diderot-LONGDALE sample. Conversely, the combination that contributes most negatively is *that* in the NOCE sample. This difference indicates two different trends in the use of the form depending on the L1 of the speakers. French learners tend to overuse

that, which increases the contribution of the form to the test result whilst Spanish learners tend to do the opposite, which downsizes the effect of the form. This inverted trend could be the sign of L1 specific positive transfer processes. The French L1 spoken corpus favours *that*, which is a trend also found in spoken native English by Biber (1999, 274, 349). French L1 learners of English would positively endorse this trend, reflecting a difference with the written mode of the WSJ sample. Conversely, Spanish L1 learners favour *this* to the point of using it abusively in comparison with the native corpus. This may be due to the written mode of the corpus and its propensity to give more room to *this* in English (Biber *et al.* 1999, 274, 349) So it seems that the spoken and written modes influence the choice of the form and that the L1 may lead to some form of overuse.

	WSJ	NOCE	DID
That	-0.6030	-1.2060	1.8090
This	0.4000	0.8000	-1.2000

Table 50: Pearson residuals per form and per corpus

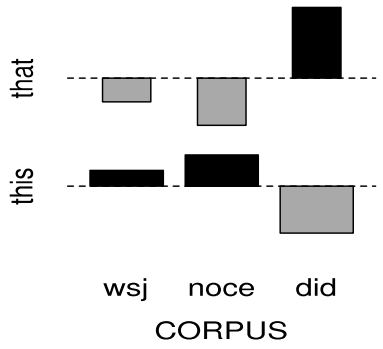

Figure 33: Effects of forms per corpus

A graphical representation can be given as a conclusion of the analysis of the use of the forms in relation to the three corpora. A plot (see Figure 33) can be drawn with R's *assocplot* function. The surfaces of the boxes are proportional to the differences between observed and expected values. Observed values greater than expected ones are shown in black above the dotted lines. Conversely, grey boxes indicate expected values greater than observed ones. The heights are indexed on Pearson

residuals and the widths on expected frequencies. *That* in French L1 is the least expected use that has the strongest effect even though the effect overall is not that strong. The model clearly shows difficulties for learners to handle the two forms and more exploration is needed with the inclusion of more variables, which is the purpose of Section 6.2.2.

6.2.2 Predictors of the forms – an exploratory analysis of the impact of several factors on the selection of *this* or *that*

This section is a multifactorial analysis of the distribution of the forms as determiners and pro-forms. The purpose is to explore various factors influencing their selection. As Gries puts it: "Essentially, all linguistic phenomena are multifactorial in nature: there is always more than one cause for any given effect and often we need to take moderator and confounding variables into consideration." (Gries 2013 [2009], 300). The situation in this case is not different. The study of *this* and *that* across several corpora does not just involve the forms and corpora as variables. Other elements related to the forms such as different types of annotated information may be part of the equation. In Section 6.1, an extract of the sample is shown in Table 46 and it shows the many different variables that can be taken into account. They are all categorical in nature. In such a case, two approaches are appropriate. Firstly, a logistic regression model helps explore how the binary categorical TOKENS variable evolves in relation to other predictor variables such as CORPUS, TAGS, POSITION or CONTEXT. Secondly, a visual interpretation of the statistical results is provided by a decision tree in which the selection process of the forms is represented in branches determined by significant variables.

6.2.2.1 Binary logistic regression model

The first approach is a binary logistic regression model and relies on hypotheses that operationalise the type of correlation between the forms and other factors. The hypotheses stem from the idea that the selection of *this* and *that* is linked to the

presence of specific PoS, textual or contextual elements used to characterise the form and its close environment. So the formulation might be as follows:

- H0: There is no correlation between TOKENS and other predictors (independent categorical variables and their interactions[29]).

- H1: There is a correlation between TOKENS and other predictors (independent categorical variables and their interactions).

The data, as described in Section 6.1, are loaded into R. The headers of the table are the independent predictor variables. It is important to mention that reordering of the values of the CORPUS variable of the data frame is operated in order to set WSJ as the first level of the variable. The reason is that the results of the generalised linear model include an intercept indicator which represents the predicted value of *that* when all predictor variables are set to their first level. Consequently, since we are interested in observing how learner corpora differ from the native corpus, reordering the native corpus as the first level of the CORPUS predictor places it within the intercept's scope, and the other indicators of the CORPUS variable are given in relation to the first one. In other words, statistical tendencies will be analysed in reference to the native corpus.

The purpose of the analysis is to test H0 and H1 with a model. A significant p-value for the model would appear as evidence of a correlation between the forms and significant predictors that are yet to be identified. However, before interpreting results, it is necessary to select the most significant model. For this, the selection procedure is twofold. Firstly, we compute several comprehensive models in terms of variables and select the most significant one on the basis of an indicator called AIC (Akaike Information Criterion). The model with the lowest AIC indicator provides information on the goodness of the model following a lower-the-better logic (Cornillon *et al.* 2010, 179). The AIC is a "measure that relates the quality of a

[29] Interactions between two independent variables occur when their joint effect on the dependent variable (TOKENS here) is not predictable from their individual effect.

model to the number of predictors it contains (and thus operationalizes Occam's razor)" (Gries 2013 [2009], 261). Secondly, once the model is selected, we operate internally on the variables so as to eliminate the least significant variables. This helps distinguish the relevant variables with regards to the predicting power of the model. In other terms, the trimmed model informs us on the variables that are statistically significant in the selection of the forms.

Selecting the most comprehensive model starts with formulating a model that includes all possible variables. The first model's formula can be written as follows:

TOKENS~CORPUS+TAGS+CONTEXT+POSITION+TAGS3AFTER+TAGS2AFTER
+TAGS1AFTER+TAGS3BEFORE+TAGS2BEFORE+TAGS1BEFORE+TOKENS1AFT
ER+TOKENS2AFTER+TOKENS3AFTER+TOKENS1BEFORE+
TOKENS2BEFORE+TOKENS3BEFORE

– in which the TOKENS variable (its levels are *this* and *that*) is a function of

- the corpora described with CORPUS
- the PoS function of the forms determined by TAGS
- the nominative or oblique cases found in POSITION
- the endophoric or exophoric context found in CONTEXT
- the PoS tags of the forms that surround each selected occurrence in its 3-gram context (TAGxBEFORE/AFTER).
- The tokens that surround each selected occurrence in its 3-gram context (TOKENSxBEFORE/AFTER)

When running this first model a warning is issued in R related to complete and quasi-complete separation in logistic regression. The warning says that "fitted probabilities 0 or 1 occurred" and that "the model did not converge". The reason seems to be due to the many levels present in each of these predictor variables. It appears that some of their rarely occurring levels happen to divide the dataset perfectly[30]. In other terms, for a given level, it is always the same value which is set for the response variable. Consequently, this raises the predicting probability to 1. This is a problem because the value of the dependent variable (*this* or *that* in this case) is always predicted by a specific level of the TAGxBEFORE/AFTER variable. This variable, albeit rare, is a perfect predictor of a *this* or a *tha*t and, if it is introduced in the model, its predicting power unbalances it. In this case study, the problem is likely to be linked to the low number of occurrences of forms of a particular level in the sample. The following details show information on the structure of the data frame that represents the sample in R:

```
'data.frame':   108 obs. of  18 variables:
 $ DIDID: Factor w/ 83 levels "DID0014-S001.seq4R",..: 14 13 6 33 18 12 15 35 34
26 ...
 $ TOKENS: Factor w/ 2 levels "that","this": 1 1 2 2 2 2 2 1 1 1 ...
 $ TAGS: Factor w/ 2 levels "DT","TPRON": 2 2 1 1 1 1 1 1 1 2 ...
 $ TOKENS3BEFORE: Factor w/ 688 levels "=",":",".","a",..: 23 58 184 204 2 384 219...
 $ TAGS3BEFORE: Factor w/ 45 levels ":",".","CC","CD",..: 14 12 19 22 1 28 11 18 40...
 $ TOKENS2BEFORE: Factor w/ 769 levels "=",":",".","...",..: 3 482 462 408 419 426...
 $ TAGS2BEFORE: Factor w/ 44 levels ":",".","CC","CD",..: 2 19 12 5 35 27 30 14 19...
 $ TOKENS1BEFORE: Factor w/ 612 levels "<?>","=",":",..: 323 305 343 82 350 98 298...
 $ TAGS1BEFORE: Factor w/ 39 levels ":",".","CC","CD",..: 18 31 27 10 23 27 21 3 31...
 $ TOKENS1AFTER: Factor w/ 490 levels ":",".","a","about",..: 251 24 166 274 166
236...
 $ TAGS1AFTER: Factor w/ 39 levels ":",".","CC","CD",..: 31 25 13 13 13 16 13 13 13...
 $ TOKENS2AFTER: Factor w/ 586 levels "=",":",".","a",..: 64 6 105 182 153 31 3 216...
 $ TAGS2AFTER: Factor w/ 44 levels "",":",".","CC",..: 10 26 9 12 37 26 3 9 9 6 ...
 $ TOKENS3AFTER: Factor w/ 749 levels " "," =",":",..: 41 7 290 138 429 8 491 434...
 $ TAGS3AFTER: Factor w/ 47 levels "",":",".","CC",..: 9 28 22 33 6 28 15 15 18...
 $ CONTEXT: Factor w/ 2 levels "ENDO","EXO": 1 1 2 2 2 1 1 1 1 1 ...
 $ POSITION: Factor w/ 2 levels "NOMI","OBLI": 1 2 2 2 1 2 2 2 2 1 ...
 $ CORPUS: Factor w/ 3 levels "wsj","noce","did": 3 3 3 3 3 3 3 3 3 3 ...
```

[30] See IDRE-UCLA's FAQ page on statistical analysis for full details on the way the dependent variable separates a predictor variable perfectly. http://www.ats.ucla.edu/stat/mult_pkg/faq/general/complete_separation_logit_models.htm (Last access March 31, 2016)

The number of levels for the TOKENSxBEFORE/AFTER and TAGSxBEFORE/AFTER predictors is very high, which makes each level rare enough to generate complete or quasi-complete separation.

Even though IDRE-UCLA's FAQ recommends applying strategies which involve merging certain values of the incriminated variables in order to decrease the number of levels, this solution appears rather difficult, if not counter-productive in terms of linguistic analysis, because of the loss of the functional information carried by the levels of the tag-type predictors. Merging the tokens-type predictors would yield no meaningful information about that new predictor. Therefore, a new model is tested without the 3-gram token-type and tag-type variables.

The four predictor variables are entered in the model with all the interactions of CORPUS in order to reflect the fact that some of the variables may evolve specifically according to each corpus. Specific patterns may appear for each of these variables depending on the corpus. The following model is set to include this nesting specificity: TOKENS ~ CORPUS + TAGS + POSITION + CONTEXT + CORPUS:TAGS + CORPUS:POSITION + CORPUS:CONTEXT.

The model includes variables and also interactions between variables. Interactions between variables yield specific effects depending on the variables' levels[31]. By introducing interactions into the model, each level of TOKENS is examined in relation to the combined effects of two variables. For instance, the CORPUS:TAGS interaction combines the effects of the variables CORPUS and TAGS on the TOKENS dependent variable.

When running the model, the same warning is issued again, which is probably due to the small number of data points and thus to the fact that, for some points, the same level of a predictor appears, which gives a 100% predicting probability. Therefore, we decide to test three other models including one interaction each:

[31] The levels are the different categories that a nominal variable can have in a model. For instance, in the case of CORPUS, the levels are *did*, *wsj* and *noce*.

```
Call:
glm(formula = TOKENS ~ CORPUS + TAGS + CONTEXT + POSITION + CORPUS:TAGS,
    family = binomial, data = wsj.noce.did.proforms.determiners)
         AIC: 101.39
Call:
glm(formula = TOKENS ~ CORPUS + TAGS + CONTEXT + POSITION +
CORPUS:CONTEXT,
    family = binomial, data = wsj.noce.did.proforms.determiners)  AIC:
114.88

Call:
glm(formula = TOKENS ~ CORPUS + TAGS + CONTEXT + POSITION +
CORPUS:POSITION,
    family = binomial, data = wsj.noce.did.proforms.determiners)  AIC:
108.34
```

As it has the lowest AIC (AIC = 101.39), we retain the model which includes the CORPUS:TAGS interaction. A summary of this model, called *model.01*, is presented below:

```
> summary (model.01)

Call:
glm(formula = TOKENS ~ CORPUS + TAGS + CORPUS:TAGS + POSITION +
    CONTEXT, family = binomial, data =
wsj.noce.did.proforms.determiners)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6643  -0.2117   0.4281   0.7596   1.2899

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)            1.0921     0.6249   1.748  0.08054 .
CORPUSnoce             1.0559     0.7904   1.336  0.18158
CORPUSdid              1.4652     0.9655   1.518  0.12913
TAGSTPRON             -0.3001     0.9309  -0.322  0.74721
POSITIONOBLI          -1.0527     0.6265  -1.680  0.09292 .
CONTEXTEXO             3.4806     1.3172   2.642  0.00823 **
CORPUSnoce:TAGSTPRON   0.1821     1.2667   0.144  0.88568
CORPUSdid:TAGSTPRON   -4.9912     1.7016  -2.933  0.00335 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 132.948  on 107  degrees of freedom
Residual deviance:  85.387  on 100  degrees of freedom
AIC: 101.39

Number of Fisher Scoring iterations: 6
```

The significance of the model can be measured by computing a Chi-squared p-value —based on the differences between i) the deviance indicators and ii) the degrees of freedom (Gries 2013 [2009], 298). The computation carried out with R returns the following result: $\chi^2 = 47.56$, p-value = 4.338661e-08, df = 7, $p < 0.001$. The model as a whole is very significant. Consequently, H0 can be rejected and a correlation between TOKENS and the predictors exists. However, their significance is yet to be established.

The second fold of the selection procedure is to operate internally on the variables of the model. Reasoning by elimination is the underlying principle of the variable selection procedure. By applying the principle of Ockham's razor, insignificant variables are eliminated. We want to test which predictor variables can be eliminated from the model without impacting it significantly. The procedure can be operated either manually or automatically and the choice of variable elimination may depend on various indicators. In this case study, a manual procedure is applied. We first compute the significance of each variable in the model and then eliminate non-significant ones. With R's *drop1* function, several predictors are tested in a stepwise manner so as to know which variable least impacts the overall model. In other words, with this function, the model is computed with and without each of its variables and the results are as shown below:

```
> drop1(model.01, test="LR")
Single term deletions

Model:
TOKENS ~ CORPUS + TAGS + CORPUS:TAGS + POSITION + CONTEXT
            Df Deviance    AIC     LRT  Pr(> Chi)
<none>           85.387 101.39
POSITION     1   88.466 102.47  3.0794 0.0792884 .
CONTEXT      1   98.257 112.26 12.8697 0.0003339 ***
CORPUS:TAGS  2  101.261 113.26 15.8739 0.0003573 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Three predictor variables are displayed, two of which are highly significant (CONTEXT and the CORPUS:TAGS interaction). POSITION is not significant as it shows a p-value which is greater than the 5% threshold. In other terms, its

elimination would not alter the results of the test significantly and the model would not be significantly worse. Following the principle of Ockham's Razor, a new model, called *model.02*, is computed by omitting the POSITION variable.

```
> model.02 <- update(model.01, ~. -POSITION)
> summary(model.02)

Call:
glm(formula = TOKENS ~ CORPUS + TAGS + CONTEXT + CORPUS:TAGS, family =
binomial, data = wsj.noce.did.proforms.determiners)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8500  -0.3189   0.5969   0.6681   0.9695

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          0.56635    0.51156   1.107   0.2682
CORPUSnoce           0.82233    0.75725   1.086   0.2775
CORPUSdid            1.06844    0.92362   1.157   0.2474
TAGSTPRON           -0.05552    0.89164  -0.062   0.9503
CONTEXTEXO           3.47752    1.38402   2.513   0.0120 *
CORPUSnoce:TAGSTPRON 0.05313    1.23428   0.043   0.9657
CORPUSdid:TAGSTPRON -4.53255    1.60721  -2.820   0.0048 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 132.948  on 107  degrees of freedom
Residual deviance:  88.466  on 101  degrees of freedom
AIC: 102.47
Number of Fisher Scoring iterations: 6
```

A Chi-squared p-value can be computed with R: p-value = 5.933085e-08 df = 7, $p < 0.001$. *Model.02* is more significant than *model.01* in terms of p-value. It is interesting to note that *model.01*'s AIC is nevertheless lower (101.39) but not so different from *model.02*'s (102.47). Due to the fact that *model.02* includes one less variable in its function, it is retained for further analysis. All in all, *model.02*'s computation results show that H0 can be rejected and that TOKENS is correlated to CORPUS, CONTEXT, TAGS and the CORPUS:TAGS interaction.

Now that the model has been selected as a whole, we look at the variables which appear as significant. The summary of the model shows that CONTEXTEXO, and

CORPUSdid:TAGSPRON are significant (p-value < 0.05) and very significant (p-value < 0.01) respectively. The readings of the estimates show that the CORPUS:TAGS interaction tends to decrease the effect on the dependent variable whilst the CONTEXT variable increases the effect. To explore the meaning of the coefficients, the values that the model predicts are computed with R (see Annex M) and a contingency table is returned (see Table 51). It indicates the details of classification by cross-tabulating predicted values with the observed values extracted from the sample.

	Predicted values	
	that	this
that	15	18
this	0	75

Table 51: Contingency table for the classification of forms with a binary logistic regression model

These results convert into a rate of misclassifieds (MC):

```
 MC <- sum(prediction.label!
=TOKENS)/nrow(wsj.noce.did.proforms.determiners); MC
[1] 0.1666667
```

So *model.02* predicts more than 80% of occurrences, which is above the 0.5 chance threshold. If we consider the proportions of observed values, *this* accounts for 69% of all data. Therefore, with 84% accuracy the model provides and improvement of 15%. The predictions can also be analysed in more detail so as to determine how each predictor impacts the model. Let us analyse the main effect predictor with R's *effect* function (see Annex M for the code). By default, R predicts the second level of the dependent variable (TOKENS), which happens to be *this* in our case. Regarding *that,* since the model is binary (two levels for the dependent variable), low prediction for *this* can be interpreted as a preference for *that*. The results for *this* are presented in Tables 52 and 53 and include confidence intervals.

To start with, the CONTEXT predictor is significant and its estimate shows that the switch from the endophoric level to the exophoric level increases the probability attached to the second level of the independent variable, which is *this*. R's *effect* function gives access to the predicted probabilities of occurrence of the form. Table

52 shows that the model predicts a high enough probability of *this* when the position is endophoric and a very high probability of the form when it is exophoric. The preference for *this* becomes stronger in the exophoric context.

CONTEXT	PREDICTIONS	LOWER	UPPER
ENDO	0.6509	0.5112	0.7687
EXO	0.9837	0.8186	0.9987

Table 52: Predicted probabilities of *this* in relation to the endophoric or exophoric context

The two predicted probabilities are higher than 0.5, which shows that *this* is preferred to *that* in both endophoric and exophoric contexts. However, there is a clear-cut preference for *this* when the context turns exophoric. The predicted probabilities are close to 1 and the lower bracket of the confidence interval still shows a high value (0.8186). This strong preference shows that within the exophoric context, the use of *this* is globally more likely than *that*.

The second significant predictor variable to explore is the CORPUS:TAGS interaction. The *effect* function returns the following values with confidence intervals (see Table 53). The model predicts rather high probabilities of *this* for all the corpus tags combinations except for the Diderot-LONGDALE and TPRON combination (0.0958). Conversely, this case means that the probability becomes very high for the use of *that* in the pro-form function (1-0.0958 = 0.9042). Said differently, the Diderot-LONGDALE speakers show a strong preference for *that* in the pro-form function when all other uses in other corpora and in the determiner function show a preference for *this*.

CORPUS	TAGS	PREDICTIONS	LOWER	UPPER
WSJ	DT	0.7816	0.5499	0.9129
NOCE	DT	0.8906	0.7054	0.9651
DID	DT	0.9124	0.6803	0.9808
WSJ	TPRON	0.7719	0.4219	0.9401
NOCE	TPRON	0.8904	0.6713	0.9700
DID	TPRON	0.0958	0.0158	0.4107

Table 53: Predicted probabilities of *this* in relation to corpora and tags

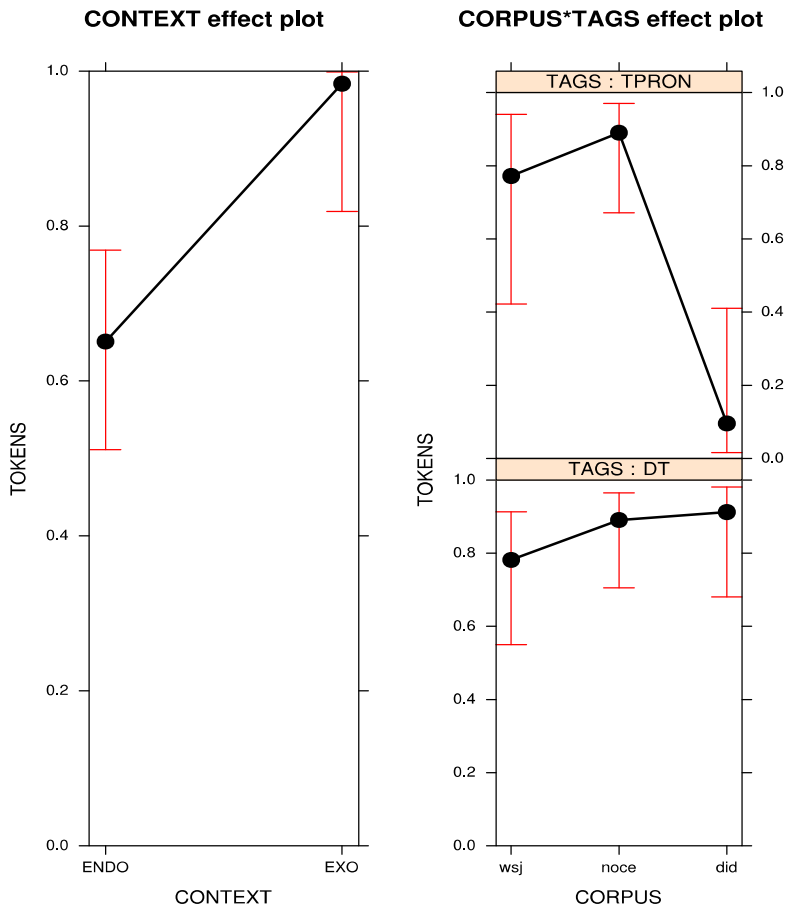

Figure 34: Effects on *this* of predictor variables in the binary logistic regression model

These predictions show that the choice of *this* or *that* is correlated to the predictors highlighted by the model. They are visualised in Figure 34 in which the increase for CONTEXT shows the preference for exophoric *this*. The right-hand diagram illustrates the strong preference for pro-form *that* in the French L1 learner corpus, as *this* TPRON is strongly rejected. Conversely, *this* DT is preferred. This is also true for the two other corpora. In fact, it seems that this trend increases with learner corpora, indicating some overuse of the form in its determiner function. In short, the binary logistic regression model approach leads to the identification of the CORPUS and TAG predictor variables that have a specific impact when interacting together. The model shows that specific settings of the variables help predict the choice of *this* or *that* in the sample. In Section 6.2.2.3, we try to interpret these results.

With the model selection and its interpretation completed, its diagnostic is necessary to test if the assumptions of the logistic regression model are met (Gries 2013 [2009], 295). This implies that a series of conditions be respected in order to ensure that the dataset matches the model's internal rules and that interpretation is meaningful (Paolillo 2002, 17-18). There are a number of diagnostic tests which can be carried out to verify the assumptions—the technicalities and the justifications of these tests are outside the scope of this thesis. Firstly, the data points must be independent. The sampling method takes this issue into account as explained in Section 6.1. Secondly, the distribution of the residuals—the differences for each data point between its observed and estimated values—must be assessed. The dispersion of residuals should be constant and can be plotted as shown in the two left-hand graphs of Figure 35. The predicted values (x-axis) are plotted v. two kinds of residuals (y-axis). In the top left-hand graph, the values should be scattered evenly around the 0 line, which may be the case. However, caution needs to be taken as there is a low number of values on the graph and it also shows that data point #148 leads to higher dispersion. In the bottom right-hand graph, the dispersion of values should show neither an increase nor a decrease when reading from left to right, which is hard to tell considering the low number of values. The logistic regression model also assumes that the residuals are distributed normally. The graph in the top right-hand corner shows the distribution of the data points. A normal distribution should show the points closely following a line, which is not quite the case here as there are sparse points on either side of the dashed line. A number of outliers appear, *e.g.* #148. Detailed observation of this point reveals that it corresponds to a *that* determiner in the WSJ corpus. It seems that because it is used in an exophoric context, this makes it stand out in the sample in relation to other uses of *that* determiner. One suggestion in such a case is to operate adjustments on the sample data with the elimination of outliers. The bottom-right hand plot confirms that some points have a large enough leverage on the model. Point number 148 and three other points on the right-hand corner of the plot clearly deviate from the cloud.

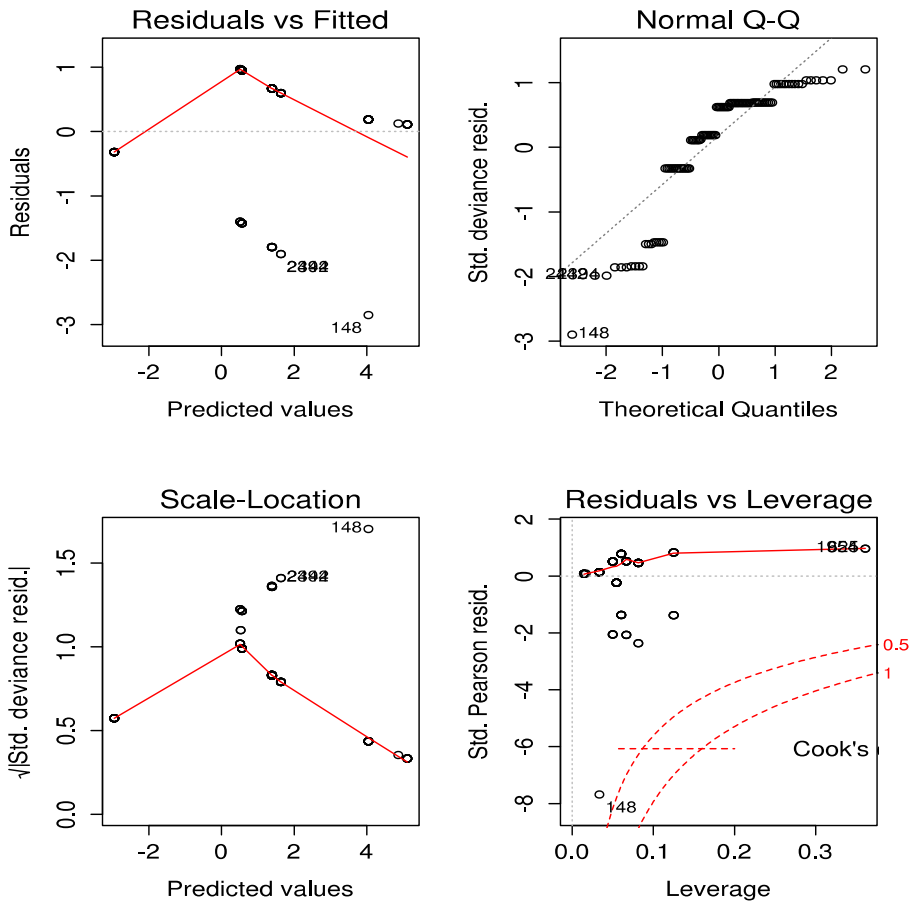

Figure 35: Model diagnostics for model.02

Thirdly, regarding data dispersion, the requirement is to have a homogeneous distribution to ensure that data variability remains within the model's boundaries. Checking overdispersion can be done via the ratio of the model's residual deviance and its residuals' degrees of freedom (*dfs)*. The ratio for *model.02* is 88.466/101 which is equal to 0.8759, greatly inferior to the reference level of 1 (Gries 2013 [2009], 315). Another assumption is about the absolute values of the standardised residuals of the model. No more than 5% should be outside of the [-2, 2] space. The computation is done via R:

```
> prop.table(table(abs(rstandard(model.02))>2))
      FALSE        TRUE
0.990740741 0.009259259
```

It shows that 0.9% of residuals are contained in the aforementioned space, which validates the assumption. Finally, one more assumption relates to what is called

dfbetas. This indicator shows "how much a regression coefficient changes when each case [data point] is removed from the data" (Gries 2013 [2009], 315). The values are expected to be lower than 0.1. R returns the results presented in Table 54 (see Annex M).

(Intercept)	CORPUSnoce	CORPUSdid	TAGSTPRON	CONTEXT EXO	CORPUSnoce: TAGSTPRON	CORPUSdid:T AGSTPRON
Min. : 0.000e+00	Min. : 0.0000000	Min. : 0.0000000	Min. : 0.0000000	Min. : 0.000000	Min. : 0.0005672	Min. : 0.0000000
1st Qu.: 6.637e-05	1st Qu.: 0.0005672	1st Qu.: 0.0000348	1st Qu.: 0.0000664	1st Qu.: 0.001274	1st Qu.: 0.0027760	1st Qu.: 0.0005416
Median : 8.349e-04	Median : 0.0027760	Median : 0.0015247	Median : 0.0018956	Median : 0.016021	Median : 0.0875876	Median : 0.0495172
Mean : 2.414e-02	Mean : 0.0457989	Mean : 0.0454573	Mean : 0.0488622	Mean : 0.046003	Mean : 0.0926788	Mean : 0.0915015
3rd Qu.: 4.063e-03	3rd Qu.: 0.0875876	3rd Qu.: 0.1257253	3rd Qu.: 0.0140090	3rd Qu.: 0.048416	3rd Qu.: 0.1268109	3rd Qu.: 0.1449908
Max. : 1.908e-01	Max. : 0.2356194	Max. : 0.4549019	Max. : 0.4132930	Max. : 0.841411	Max. : 0.4132930	Max. : 0.4404277

Table 54: Summary of dfbetas of model.02

The results show that some absolute *dfbeta* values are greater than 0.1, contrary to what they should be (Gries 2013 [2009], 316). This indicates that some coefficients are highly dependent on specific data points. Overall, the model violates a number of assumptions which weakens its robustness. Several alternative modelling strategies could be considered such as fitting a Poisson model in order to see if the data fit the distribution better. Also the size of the sample could be increased, which might give a more accurate view of the residuals' dispersion. Bearing in mind these words of caution on the assumptions of tests, this case study retains *model.02* as an entry level logistic regression model.

6.2.2.2 Decision tree model

The second modelling approach for this multifactorial analysis on the selection of *this* or *that*—in their pro-form and determiner functions—is the construction of a decision tree. Decision trees cater for improved readability in comparison with regression models. They give a simple interpretation of the statistical results due to the way the data are represented in the form of a tree. The idea is that each node leads to several branches and ultimately several leaves. In this view, each branch

can be assimilated to a type of decision for the dependent variable. Decision trees essentially help create a representation of the choices made for the selection of a form and they can be applied on qualitative variables, as is the case in this study.

The underlying concept behind decision trees is the fact that a sequential algorithm builds classes of observations. “Classes are generated thanks to binary rules that rely on independent variables. The purpose is that same-class observations are as homogeneous as possible with regards to the dependent variable.” (Cornillon *et al.* 2010, 182, my translation). In other terms, the algorithm classifies occurrences of the dependent variable so as to obtain groups of *this* and *that* forms. In this case study, R's package *rpart* is implemented to build the trees and Cornillon's methodology is followed (Cornillon *et al.* 2010, Section 6.15).

The same data sample is used as presented in Section 6.1. The model's formula is identical to that of the generalised linear model presented in Section 6.2.2.1 but without the CORPUS:TAGS interaction since *rpart* cannot take interactions. A first model, called *model.tree*, is computed and gives the following results:

```
> model.tree <-rpart(TOKENS~CORPUS+TAGS+CONTEXT+POSITION,
data=wsj.noce.did.proforms.determiners, minsplit=1, xval=108);
model.tree
n= 108

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 108 33 this (0.3055556 0.6944444)
  2) TAGS=TPRON 40 19 that (0.5250000 0.4750000)
    4) CORPUS=did 17  2 that (0.8823529 0.1176471)
      8) CONTEXT=ENDO 15  0 that (1.0000000 0.0000000) *
      9) CONTEXT=EXO 2  0 this (0.0000000 1.0000000) *
    5) CORPUS=wsj,noce 23  6 this (0.2608696 0.7391304) *
  3) TAGS=DT 68 12 this (0.1764706 0.8235294) *
```

The tree is shown in the lower part of the frame. It shows 9 nodes indicated by the numbers followed with closing brackets. The lines in bold give indications on the type of information provided for each node. The asterisk signs indicate leaves of the tree. In other terms, they show the final decision after the algorithm has stopped splitting. For instance, node *1)*, can be read as follows: the *split* rule is applied on

the *n* observations (n = 108) of the sample including a number of misclassified occurrences (*loss* = 33). The *yval* indicator shows the predicted values for the node (yval = *this*) and *yprob* gives the probability for each class. The first value indicates the probability of *that* forms (0.305556) and the second value, that of *this* forms (0.6944444).

This first model may not be very stable for predictions. The less leaves the tree has, the more stable it will be in its future predictions (Cornillon *et al.* 2010, 186). So it is necessary to prune the tree by using the *plotcp* function. To do the pruning, *rpart*'s cross-validation process is applied with the leave-one-out option by setting *xval* to the number of observations of the sample. The results are analysed with the *plotcp* function. This function shows the cross-validation error in relation to the cost-complexity parameters for trees of different sizes. Optimal pruning can be achieved with the lowest Complexity Parameter (CP) value. Figure 36 shows that with a CP value of 0.025, the tree can be pruned to 4 remaining leaves.

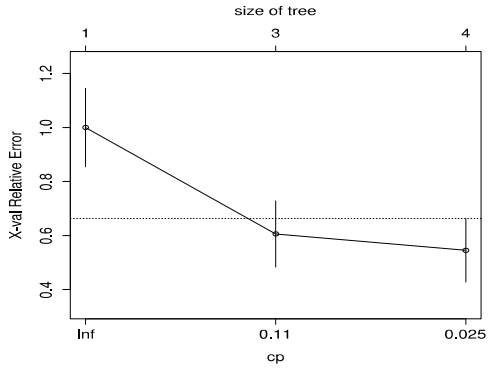

Figure 36: Complexity of the tree in relation to its size

The identification of the lowest possible CP value can be reintroduced in a new model of the tree, *i.e. model.tree.2*, and R gives the following results:

```
> model.tree.2 <-rpart(model.tree,
data=wsj.noce.did.proforms.determiners,cp=0.025);
proforms.or.determiners.tree2
n= 108

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 108 33 this (0.3055556 0.6944444)
  2) TAGS=TPRON 40 19 that (0.5250000 0.4750000)
    4) CORPUS=did 17  2 that (0.8823529 0.1176471) *
    5) CORPUS=wsj,noce 23  6 this (0.2608696 0.7391304) *
  3) TAGS=DT 68 12 this (0.1764706 0.8235294) *
```

In this second tree, there are still 108 observations (which corresponds to the size of the sample) and the probability of getting a *this* form is still 69.44%. It can be noted that *model.tree.2* is as stable as *model.tree.1* but it is simplified in terms of branches. This first node is split with the TAGS variable. The observations are split into two groups. The first group, which corresponds to node 2), TAGS=TPRON, includes 40 observations of which 52.50% correspond to *that*. There are 19 misclassified observations. The second group (node 3 TAGS=DT) is made of 68 observations of which 82.35% are *this* forms. The decision-making process can be read from the leaves. The following rules are applied:

- If the tag of the form is TPRON and the corpus is the Diderot-LONGDALE, then the algorithm selects *that*.

- If the tag of the form is TPRON and the corpus is either the WSJ or the NOCE, then the algorithm selects *this*.

- If the tag of the form is DT, then the algorithm selects *this*.

A graphical representation of the tree is produced with the *prp* function in R's *rpart.plot* library.

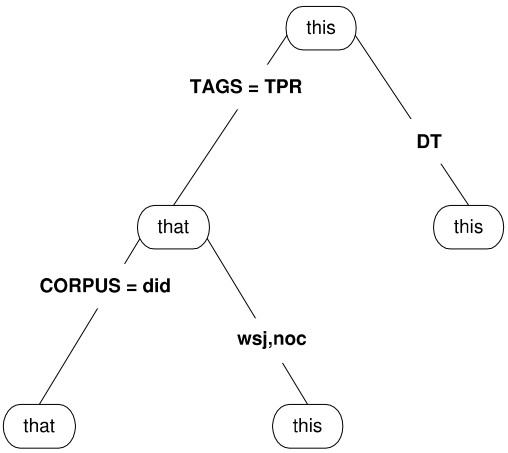

Figure 37: A conditional inference tree gives the corresponding scores for *this* and *that*

It appears that the CONTEXT predictor variable is not kept in the final decision tree while it is in the logistic regression model. The reason is that after pruning, the algorithm is simplified by eliminating an extra split of the *CORPUS=did* branch. This split (nodes 8) and 9) in the *model.tree*) relies on the CONTEXT predictor but the predictive stability of the tree is not altered with its elimination. This may be due to the high probability of *that* forms in endophoric contexts which downplays the significance of the *this* forms that appear in exophoric contexts. In other terms, the CONTEXT predictor can still be considered as a significant factor, as shown in the regression model, but the tree does not require this criterion to improve its predictions. The summary of *model.tree.2* shows that the tree classifies the forms in relation to TAGS primarily (40 occurrences for TPRON and 68 for DT). Of the 40 occurrences tagged as TPRON, the split is done according to CORPUS. The classification performance can also be computed in terms of misclassified occurrences (2 forms misclassified as *this,* 6 and 12 forms misclassified as *that*) yielding a ratio of 18.52%.

Overall, the tree presents results that are similar to the logistic regression model, insofar as the CORPUS and TAGS predictors are concerned. In the regression model the CORPUS:TAGS interaction is a key factor for the selection of the forms. The decision tree model approach cannot take interactions into account.

6.2.2.3 Discussion

Case Study 1 provides evidence on the selection of the forms in relation to other factors. A mono-factorial approach shows the correlation that exists between the corpora and the forms. The model supports the idea that, by varying corpora, the way *this* and *that* are selected is different. The spoken or oral modes are likely to play a role. However, the fact that significant differences appear between corpora of the same mode suggests that L1-specific factors are at work. By developing a multi-factorial approach, it is possible to explore other factors. A binary logistic regression model helps make CONTEXT and CORPUS:TAGS (interaction) emerge as significant variables for the prediction of the forms. Conversely, POSITION does not seem to be significant. Regarding CONTEXT, both endophoric and exophoric contexts attract mostly *this* rather than *that*. Exophora attracts *this* more than endophora does. The significance of the CORPUS:TAGS interaction suggests that some specific values, resulting from cross-referencing the levels of CORPUS and TAGS, have a strong predicting power. *This* as a pro-form is likely to appear in the written mode with a stronger effect for Spanish L1 than native English. Conversely, *that* as a pro-form is strongly linked to the oral mode and the French L1. The likeliness of the determiner function is consistent across all corpora. *This* determiner is more likely to be selected than *that*, especially with learner corpora, which suggests an L1-specific effect of overuse. The decision tree model reinforces the idea that CORPUS and TAGS play a major role by showing an order of variables for selection. It appears that the TAGS, *i.e.* the function of the form, is the first criterion. The second criterion is CORPUS, *i.e.* the corpus characterised by the L1 or the mode.

At this point in the section, it is interesting to discuss each variable. To do so, predicted values can be compared with observed values in order to seek linguistic explanations for the results and see whether the statistics cohere with the linguistic observations. The first predictor studied is the effect of the exophoric contexts on the selection of *this.* One point to discuss is whether observed values reflect this estimate. The left-hand plot of the observed values in Figure 38 seems to confirm this. Furthermore, the right-hand side of the figure shows that exophora occurs mostly in the native and the French L1 corpora as opposed to nearly not for the Spanish L1 corpus. This difference is strange at first sight as it does not reflect the differences between modes nor between L1s. However, under the hypothesis that the L1s were to be the source of the difference, the Spanish language would be logically considered as poor in exophoric reference. This is not the case, suffice to mention the tri-partite demonstrative system with *éste, ése aquél* described in (Gerboin and Leroy 1991, 49). The other hypothesis is that the high exophoric use linked to the WSJ and the Diderot-LONGDALE is due to specifics of each corpus such as the genre. In the journalistic genre, natives would have a specific use of exophora whilst French learners would have a use of it linked to the spoken mode and, more specifically, the conversation genre.

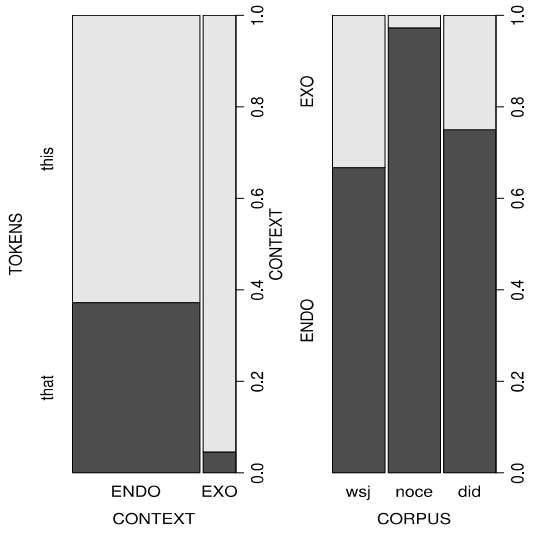

Figure 38: Observed values for the forms and corpus in relation to the type of context

In written modes, most of the referential processes are endophoric as the writers need to construct their discourse entirely for their readers. Sometimes though, which may be the case for the journalistic genre, they use exophora but in specific conditions such as referring to time periods. Exploration of the annotated transcripts with NXT Search may help in this matter as it allows the researcher to retrieve exophoric occurrences in the WSJ. It appears that most of the uses are done for time period referrals such as in example 59.

59) They make the argument in letters to the agency about rule changes proposed *this* past summer that, among other things, would exempt many middle-management executives from reporting trades in their own companies' shares. (WSJ-0203.mrg)

An extraction which matches exophoric occurrences and right 1-gram tokens from the sample with R gives the following list: *week, year, century, phone, article, spring, week, decade, fall, morning, week, morning*. In addition, most of the forms are *this* in its determiner function. It appears that in the observed values of the WSJ, *this* is largely found in temporal contexts in exophoric realisations. These observations suggest that journalists make use of *this* in specific temporal contexts. The high predicted probability of exophoric *this* seems to mirror this situation.

French learners of English also have a specific use of exophoric *this*. An R extraction which matches exophoric occurrences and right 1-gram tokens from the sample returns the following list: *one, year, one, is, year, summer, it, one, summer*. This list can be explained by the fact that the recordings are based on questions related to i) several pictures shown to the students, ii) their future plans. Regarding the first type of question, words like *one* reflect answers including a reference to a visible entity in the situation of utterance (see Example 60). Consequently, many referential processes in the corpus are exophoric, as the learners, involved in a conversational process, keep referring to the physically present pictures without necessarily naming them.

60) I don't know anything about art but this one reminds me of the book the Awakening that I read the last semester... (DID0014-S003)

Regarding the second type of question, some words in the list refer to periods of time, which indicates that French learners also need to place discourse events in time. It seems logical that, just like journalists of the WSJ, learners use *this* in temporal reference as in Example 61:

61) I think I will do it this summer (DID0014-S002)

Observed values of exophoric *this* in each corpus, crossed with observed values in each context, confirm that, whenever the *this-exophoric* combination appears, it is highly predictable. Said differently, *that* is nearly absent in exophoric contexts in our data, which leaves little uncertainty for *this*.

The second predictor which is significant in the model is the interaction between CORPUS and TAGS. Determiners seem to impact the use of *this* greatly, and even more so with learner corpora. The pro-form function shows a somewhat different scenario. The probability of *this* greatly drops when French L1 learners use the pro-form. This, conversely, shows that they have a strong preference for *that* in the pro-form function. The L1 French learners is a distinct group from the other two in terms of L1 and mode as it is spoken English. Consequently, there are two possible explanations for the difference. Either it is due to mode specificities or else it is due to transfers from French.

Exploration of the Diderot-LONGDALE sample is necessary to understand better what happens exactly. The way the form is introduced may cast some light on the source of the distinction between the two groups, *i.e.* Spanish L1 and native speakers on one side and French L1 speakers on the other. The three pareto charts (Figures 39, 40 and 41) display information regarding the distribution of the form and its left 1-gram context. They help compare native with learner data. These charts show the most frequent 1-gram tag before *that*. For example, the *IN* tag, which corresponds to prepositions, is the most present of all the tags in all three

corpora. In other terms, *that* is introduced by a preposition in most cases whatever the L1, which likely corresponds to its pro-form realisation. For natives, 50% of *that* pro-forms are introduced by *IN*, *CC* (coordinating conjunction) and comma. If we add *VB* (verbs) and *RB* (adverbs) it represents 75% (see dashed curve on chart). For Spanish L1 learners, 50% of *that* pro-forms are preceded by *IN* only. And, with CC, it represents 75% of the categories preceding the marker.

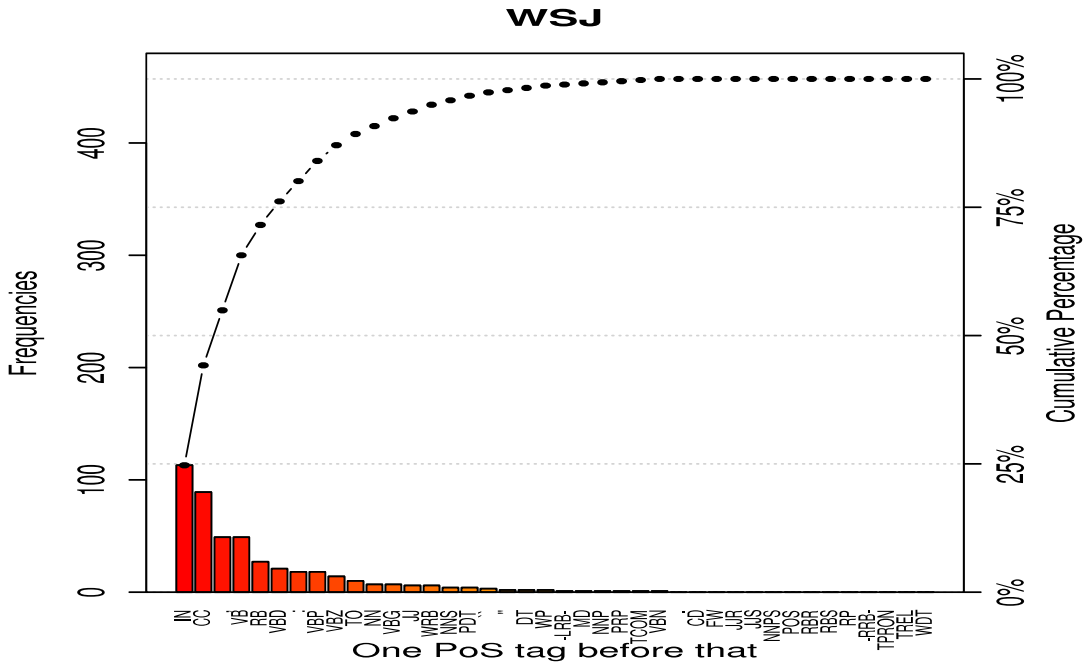

Figure 39: Distribution of left PoS tag before *that* in the WSJ

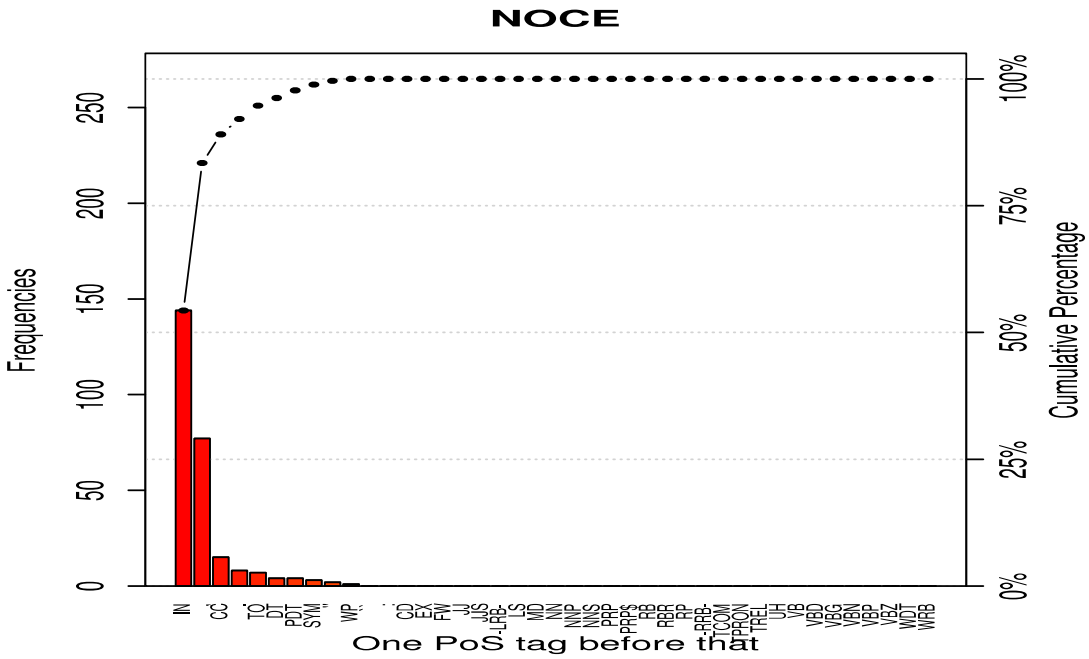

Figure 40: Distribution of left PoS tag before *that* in the NOCE

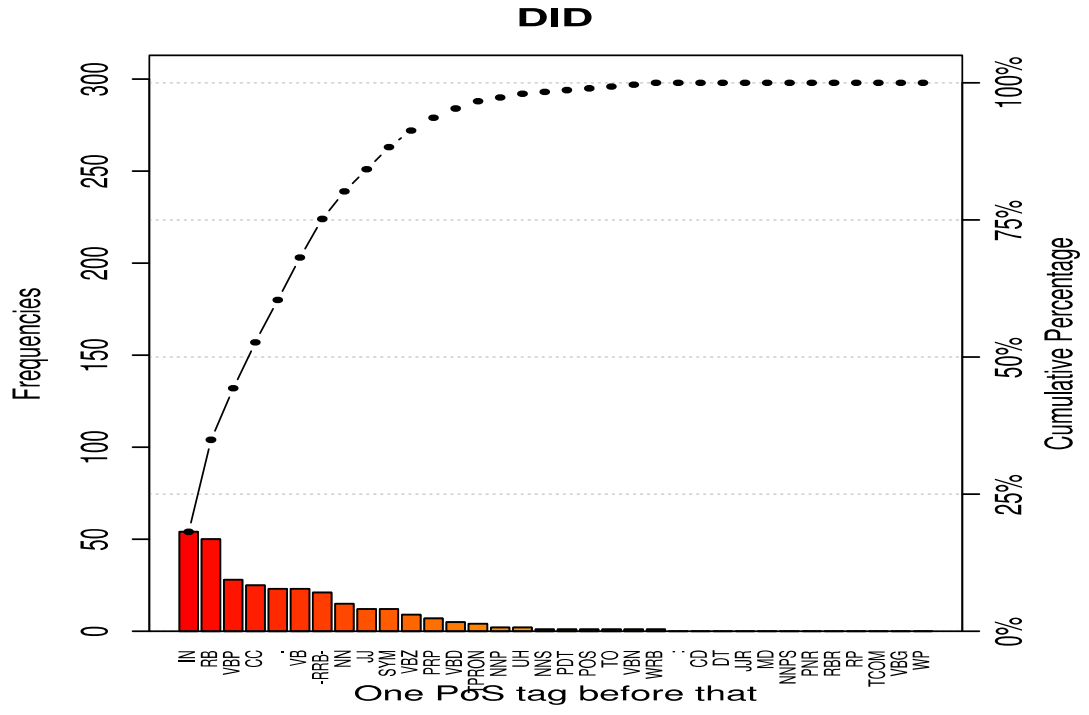

Figure 41: Distribution of left PoS tag before *that* in the Diderot-LONGDALE

For French L1 learners, 50% of *that* pro-forms are introduced by *IN*, *RB*, *VBP* and *CC*. With pauses, *VB*, and *-RRB-* (parentheses indicating transcription comments

such as laughs), it represents 75%. All this is synthesised in Table 55, which sums up the immediate preceding tags that account for 50% and 75% of the contexts of occurrence of our pro-forms It must be noted that the DID chart is drawn from the sample of data used in this case study but the two others are drawn from the actual full corpora in order to have enough forms.

Quantiles	First 50% of preceding PoS	First 75% of preceding PoS
WSJ	*IN, CC*	*IN, CC, VB, RB*
NOCE	*IN*	*IN, CC*
DID	*IN, RB, VBP, CC*	*IN, RB, VBP, CC, p*auses, *VB, RRB*

Table 55: Synthesis of mostly used PoS tags before *that* pro-forms

Some comments can be made on these figures. First, it should be noted that we find ourselves justified in distinguishing oblique positions with this overwhelming number of prepositional constructions. French L1 learners slightly underuse prepositions compared with natives (less than 25%) whilst Spanish L1 learners overuse them (55%). Of all the types of forms that precede *that* pro-forms, it seems that adverbs play a more important role among French L1 learners. They account for about 20% of all forms. Subsetting these occurrences from the sample gives a long list of concordances (see Annex N). A close analysis of the data shows that the pro-form is mostly preceded by *so* as an adverb as in: “wow and so that was good”. This example illustrates a prototypical structure used by French L1 learners. It corresponds to so + *that* + be. This might be evidence of a 'conspiracy' of sequences with *and so* being used as a discourse marker, as in French “et donc”, superposed with *so that*, which exists in English. Learners might be merging both phenomena resulting in an embryo of grammatical construction. Many cases show that learners' use of the string of words includes either of the two following features: a pause after *so* or an absence of reduction for *that*. Consequently, it appears that the use of this construction is different from *so that* expressing consequence. In fact, the occurrences seem to reveal two types of use. Firstly, in some utterances, students explain the cause of an assertion and they use expressions such as *so that's why.* Secondly, the students seem to be concluding their message. They were asked to give accounts of personal experiences such as a trip abroad. They were also asked to give their opinions on various paintings.

Overall, they were in a position of describing entities and finally concluding by referring back to these entities with the pro-form. *That,* in these cases, is used for its concluding meaning effect as described by (Lapaire and Rotgé 1998, 64). The speakers close the topic by emitting a judgement on the entity they have just focused on. The use of *so* would confirm this as it introduces a logical connection between what has been said and what is to be concluded. This strategy may be linked to the spoken mode or the genre which trigger a need for referential processes in the course of a presentation. The journalistic genre of the WSJ does not include such a use, which may indicate an effect due to the L1. This strategy may also be linked to the type of task given to students causing the need to conclude. Comparisons with a corpus based on map tasks—in which learners read a map and give geographical directions to another learner—may reveal that, in this case, concluding remarks are ruled out.

6.3 Case Study 2: Exploring the pro-form microsystem across corpora

The pro-form microsystem described in Section 3.2.2.2.1 indicates that learners tend to substitute the forms in various contexts. Studies in the previous section reveal that the functional realisation as a variable is significant in certain conditions. Consequently, a statistical analysis of *this* and *that* pro-forms, together with *it,* might cast light on the matter as it will help quantify the uses of the forms in terms of frequencies of use but also the predictions that can be extrapolated and thus the effects of specific predictor variables. This is where the annotation of each of the corpora brings its full leverage as the setup makes it possible to produce a corpus sample which is displayed with the different annotation levels (position, context, PoS-function) as predictor variables.

One way of exploring the microsystem is to ask the question of the correlation between the use of the forms and a certain number of predictor variables such as the type of corpus, the type of context and the type of syntactic position. Taking

into consideration the data frame structure and the names of the variables, two competing hypotheses might be formulated:

- H0: There is no correlation between TOKENS (which includes three levels corresponding to the three forms) and the predictor variables which are CORPUS, POSITION and CONTEXT.

- H1: There is a correlation between TOKENS (which includes three levels corresponding to the three forms) and the predictor variables which are CORPUS, POSITION and CONTEXT.

6.3.1 Multinomial regression model

The sample for this case study is extracted from the full corpus following a method described in Section 6.1, in which *it*, *this* and *that* in their pro-form function have been selected together with their annotated features. The sample is loaded as a data frame whose structure is presented here. The number of levels[32] can be read for each variable. The last line of the table, for instance, details the CORPUS variable which includes three levels or values, *i.e. noce, wsj* or *did*. The first set of values which are found in the dataset for this variable are the third ones, *i.e. did.*

```
> str(wsj.noce.did.proforms)
'data.frame':      108 obs. of  18 variables:
 $ DIDID: Factor w/ 73 levels "DID0014-S001.seq4R",..: 1 19 27 30 6 13
28 5 36 21 ...
 $ TOKENS: Factor w/ 3 levels "it","that","this": 1 2 1 1 3 2 2 2 1
1 ...
 $ TAGS         : Factor w/ 2 levels "PRP","TPRON": 1 2 1 1 2 2 2 2 1
1 ...
 $ TOKENS3BEFORE: Factor w/ 81 levels ",",".","a","added",..: 40 25 40
18 80 13 40 77 40 35 ...
 $ TAGS3BEFORE: Factor w/ 26 levels ",",".","CC","CD",..: 9 22 9 21 14
10 9 7 9 25 ...
 $ TOKENS2BEFORE: Factor w/ 81 levels ",",".","1","1988",..: 23 69 23 59
```

[32] See footnote page 263

```
62 81 23 11 23 42 ...
 $ TAGS2BEFORE: Factor w/ 23 levels ",",".","CD","DT",..: 19 8 19 15 7
12 19 8 19 11 ...
 $ TOKENS1BEFORE: Factor w/ 54 levels "``",",",":","!",..: 38 28 38 5 54
40 38 5 38 8 ...
 $ TAGS1BEFORE  : Factor w/ 22 levels "``",",",":",".",..: 12 7 12 4 16
20 12 4 12 11 ...
 $ TOKENS1AFTER : Factor w/ 51 levels ",",".","1929",..: 34 7 34 34 38
11 46 34 34 29 ...
 $ TAGS1AFTER   : Factor w/ 18 levels ",",".","CC","CD",..: 18 3 18 18
11 14 17 18 18 7 ...
 $ TOKENS2AFTER : Factor w/ 73 levels "","``","=","--",..: 69 64 61 40
70 10 27 38 42 29 ...
 $ TAGS2AFTER   : Factor w/ 29 levels "","``",",",":",..: 17 22 8 13 25
19 14 16 17 23 ...
 $ TOKENS3AFTER : Factor w/ 83 levels " "," ,"," .",..: 24 40 57 25 82 7
46 13 20 66 ...
 $ TAGS3AFTER   : Factor w/ 26 levels "",",",".","CC",..: 13 26 13 22 17
19 15 19 8 18 ...
 $ CONTEXT      : Factor w/ 2 levels "ENDO","EXO": 1 1 2 2 1 1 1 1 1
2 ...
 $ POSITION     : Factor w/ 2 levels "NOMI","OBLI": 1 2 1 1 2 2 1 1 1
1 ...
 $ CORPUS       : Factor w/ 3 levels "wsj","noce","did": 3 3 3 3 3 3 3 3
3 3 ...
```

After loading the sample, the choice of the statistical model is carried out on the basis of the nature of the variables. In this case, the response/dependent variable comprises three levels (*it, this* and *that*) and the predictor variables are all categorical (TOKENS, TAGS, TAGSxAFTER, TAGSxBEFORE, TOKENSxAFTER, TOKENSxBEFORE, CONTEXT, POSITION and CORPUS). A multinomial model seems appropriate for such variables. Its formulation can be done following a stepwise strategy seeking the lowest AIC indicator[33] as a criterion (Gries 2013 [2009], 260). A first model is computed with CORPUS, POSITION, CONTEXT and all their interactions. R's *stepAIC* function is used to automate the selection process

[33] In a nutshell, the regression analysis means that we try to predict the role of the various variables xi (CORPUS, POSITIONS, CONTEXT) finding the coefficients bi that account for the tokens (y) in an equation of the type: y = a +b1x1 + b2x2 + bixi + bnxn. The AIC is the Akaike Information Criterion, which assesses how well the model fits the data. It incorporates penalisation for each supplementary variable taken into account and follows the formula: AIC= -2log L(M) + $2n_{M.}$
The smaller the AIC, the better. L(M) is the loglikelihood of the model. The more variables we include, the more the model is penalised (n_M is the number of coefficients). Ultimately, this is meant to avoid too many variables in the model ('overtfitting' the model). For details on the stepwise method, see (Johnson 2008, 89–91).

on the basis of the lowest AIC. After several tested combinations some predictor variables are dropped to provide the lowest AIC model:

```
Step:  AIC=167.43
TOKENS ~ CORPUS + POSITION

# weights:  9 (4 variable)
initial  value 118.650127
iter  10 value 80.734204
final  value 80.732286
converged
# weights:  12 (6 variable)
initial  value 118.650127
iter  10 value 79.541924
final  value 79.541732
converged
           Df    AIC
<none>        167.44
- CORPUS    4 169.47
- POSITION  2 171.08
Residual Deviance: 151.4348
AIC: 167.4348
```

This model is kept in order to be estimated. It is important to note that the CONTEXT predictor is not included in the final model. The implications of such a statistical choice are discussed in Section 6.3.2. Before running the model, it is necessary to mention that *it* is the level of the dependent variable used as the default option. Coefficients show the propension of a specific *this* or *that* level to occur compared with *it*. In addition, *wsj* is the default level of the CORPUS variable. In other terms, it is a baseline to measure tendencies of other levels such as NOCE and DID. A one-unit increase in the CORPUS variable is associated with either a decrease or an increase of the chances of *this* or *that* v. *it*. This choice is linguistically motivated as the goal is to study how learners deviate from native use. Consequently, the obtained coefficients give indications on the way each of the groups of learners deviates from the 'norm' set by the *Wall Street Journal*. The syntactic complexity linked to native English may also have consequences such as overfitting the model. Complex structures such as extrapositions—which are not the default structure used by learners—could introduce extra parameters that would prove too many in relation to the number of observations, thus leading the model to focus on minor aspects of the data.

Running the model gives the results whose summary is presented in Table 56. It includes confidence intervals for each outcome variable. The coefficients are expressed in log odds.

```
> summary(model.final)
Call:
multinom(formula = TOKENS ~ CORPUS + POSITION, data =
wsj.noce.did.proforms)

Coefficients:
     (Intercept) CORPUSnoce  CORPUSdid POSITIONOBLI
that   -2.556750  0.5702129  1.6818209    1.4375997
this   -2.244729  0.4443304 -0.8335519   -0.1027629

Std. Errors:
     (Intercept) CORPUSnoce CORPUSdid POSITIONOBLI
that   0.6166837  0.7072140 0.6759528    0.5403641
this   0.6377268  0.8140742 1.1924576    0.8729388

Residual Deviance: 151.4348
AIC: 167.4348

> confint(model.final)
, , that

                  2.5 %    97.5 %
(Intercept)  -3.7654280 -1.348072
CORPUSnoce   -0.8159011  1.956327
CORPUSdid     0.3569777  3.006664
POSITIONOBLI  0.3785056  2.496694

, , this

                 2.5 %     97.5 %
(Intercept)  -3.494650 -0.9948072
CORPUSnoce   -1.151226  2.0398864
CORPUSdid    -3.170726  1.5036221
POSITIONOBLI -1.813692  1.6081658
```

Table 56: Model summary and confidence intervals for multinomial regression model on pro-form microsystem

The summary shows two parts. One part is related to the coefficients and the other one shows the standard errors. In each part there are two lines corresponding to *this* and *that*. Each line compares a specific form (TOKENS = *this* or TOKENS = *that*) to the baseline form TOKENS = *it*.

To read the coefficients, the log odds must be taken into account. For instance, the log odds of getting a *that* v. a *it* will increase by 0.57 if moving from native English (WSJ) to L1 Spanish learners of English (NOCE). The log odds of getting a *that* v. a *it* will increase by 1.68 if moving from the WSJ to the Diderot-LONGDALE. The intercept values show the log odds of *that* and *this* when the independent variables are set to their default values (*wsj* and *NOMI)*. Similarly, the log odds of getting a *that* v. *it* will increase if moving from a nominative to an oblique position. Still on POSITION, the log odds of getting a *this* v. *it* will decrease if moving from a nominative to an oblique position.

The ratio of the probability of choosing one outcome category over the probability of choosing the baseline category is computed by exponentiating the coefficients and the results are presented in Table 57. This ratio is also called the relative risk ratio (RRR). It measures the odds of getting a specific value of a variable in comparison with its default values. The intercept column shows the probabilities of obtaining *this* or *that* when the independent variables are set to default.

	(Intercept)	CORPUSnoce	CORPUSdid	POSITIONOBLI
that	0.07755	1.7686	5.3753	4.2105
this	0.1059	1.5594	0.4345	0.9023

Table 57: Relative risk ratios for outcomes per predictor

The above results can be plotted as follows:

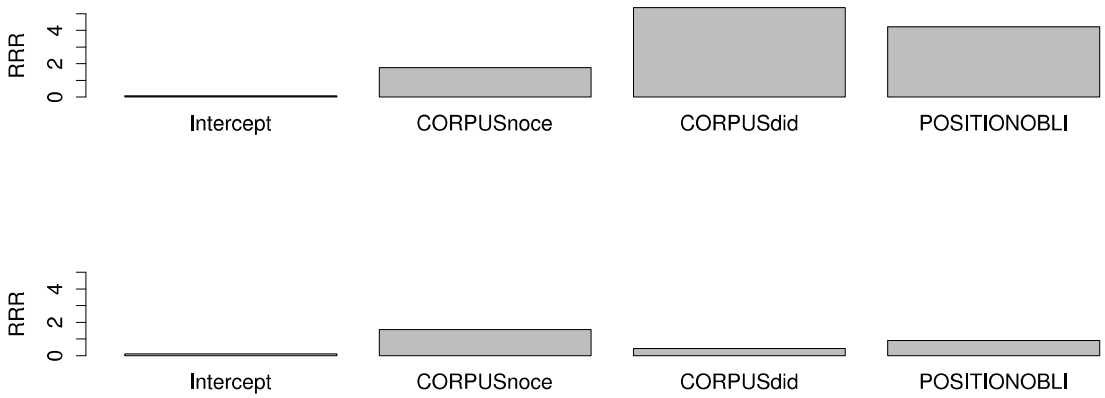

Figure 42: Relative Risk Ratio respectively for *this* and *that* according to their independent variables

By switching from the WSJ to the NOCE, the RRR is 1.7686 for using *that* instead of *it*. This means that L1 Spanish learners have a higher chance of using *that* over *it* than natives. By switching from the WSJ to the Diderot-LONGDALE, the RRR is 5.3753 for using *that* instead of *it*. This means that L1 French learners have an even higher chance of using *that* over *it* than natives and L1 Spanish learners. By switching from *it* to *this*, the RRR is 1.5594 for the NOCE v. the WSJ. This indicates that L1 Spanish learners have an increased chance of using *this* over *it* in comparison with natives. Conversely, by switching from *it* to *this*, the RRR is 0.4345 for the Diderot-LONGDALE v. the WSJ. This suggests that L1 French learners have a lower chance of using *this* over *it* than natives.

Regarding the POSITION predictor, the RRR, switching from the nominative to the oblique case is 4.2105 for using *that* v. *it*. This means that the oblique case has a higher chance of being found with *that* than *it*. Said differently, the chance of using *that* v. *it* increases when switching from nominative to oblique cases.

The p-value of the model is computed with a special R function which helps compare the model with a neutral one without any predictors.

```
> anova(model.final,multinom(TOKENS~1))
# weights:  6 (2 variable)
initial  value 118.650127
iter  10 value 83.625620
```

```
final  value 83.625614
converged
Likelihood ratio tests of Multinomial Models

Response: TOKENS
               Model Resid. df Resid. Dev   Test    Df LR stat.    Pr(Chi)
1                  1    214   167.2512
2 CORPUS + POSITION   208   151.4348 1 vs 2  6 15.81641 0.01477406
```

The results show that the model is significant overall (p-value = 0.014). On these grounds H0 can be rejected and the hypothesis of a correlation between the choice of pro-forms and the type of corpus and syntactic position, can be accepted. To determine the significance of each the coefficients, *i.e.* whether the effects of the variables are relevant, we calculate their p-values. To do so, we follow the method presented by the UCLA Statistical Consulting Group[34] and obtain the following values:

```
> p <- (1 - pnorm(abs(z), 0, 1))*2; p
      (Intercept) CORPUSnoce  CORPUSdid POSITIONOBLI
that 3.383827e-05  0.4200811 0.01284369  0.007804142
this 4.317246e-04  0.5851961 0.48453943  0.906289056
```

It appears that the values of the CORPUSdid and POSITIONobli variables are significant (p-value=0.012) and very significant (p-value=0.007) respectively when moving from *it* to *that*. The confidence intervals (shown in the summary of the model in Table 56 page 288) do not include 0, which indicates that the plausible values of the parameters cannot be null. Therefore, it means that the two variables can help infer that the French-L1 learners' choice of *that* and the oblique position of *that* are not due to chance. Rather, it seems that French learners favour *that* over *it* and that the oblique case acts as a strong criterion for the selection of *that* instead of *it*. This confirms the previous findings which show a correlation between the Diderot-LONGDALE and the use of *that*. However, these findings unveil the competition that exists between *that* and *it* and the preference of the first one over the second one.

[34] See the example of a Multinomial Logistic Regression at http://www.ats.ucla.edu/stat/r/dae/mlogit.htm (Last access March 31, 2016)

The model can also be tested in terms of classification accuracy. By fitting the model on the existing data, and by converting the ratios to category names, the resulting predictions can then be tabulated.

```
pred.prob.cat <-colnames(pred.prob)[max.col(pred.prob)]; pred.prob.cat
```

	Predicted values		
	it	that	this
it	73	3	0
that	20	4	0
this	7	1	0

Table 58: Predictions for *it*, *this* and *that* based on the statistical model

The rate of misclassified occurrences by the model is rather high:

```
> MC <- sum(pred.prob.cat!=TOKENS)/nrow(wsj.noce.did.proforms); MC
[ 1] 0.287037
```

Summary statistics computed in R provide the following results with:

```
> model.statistics(TOKENS, pred.prob.cat,pred.prob)
$loglikelihood.null
[1] -83.62561
$loglikelihood.model
[1] -75.71741
$deviance.null
[1] 167.2512
$deviance.model
[1] 151.4348
$R2.likelihood
[1] 0.09456675
$R2.nagelkerke
[1] 0.1729988
```

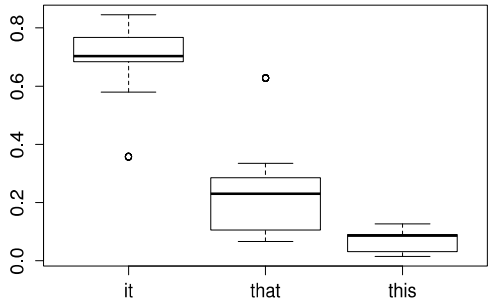

Figure 43: Box plot of the model predictions
for *it*, *this* and *that*

The predictions on the data sample (all corpora equally represented) can be presented in box plots (see Figure 43) to have an overview of the preferred outcome variable including its probabilities (with confidence intervals and means). These predictions show which form is mostly preferred in the whole sample, regardless of the type of corpus. *It* pro-form clearly appears as the preferred form. *This* pro-form shows a low predictive power, which explains the fact that the model fails to predict any of the occurrences as shown in Table 58. Confidence intervals and means show that the data points are gathered around the means. However, the means for *it* and *this* border the top of the quartiles, which indicates that a few probabilities are quite high and thus drive the mean upward (*this*) or downward (*it*).

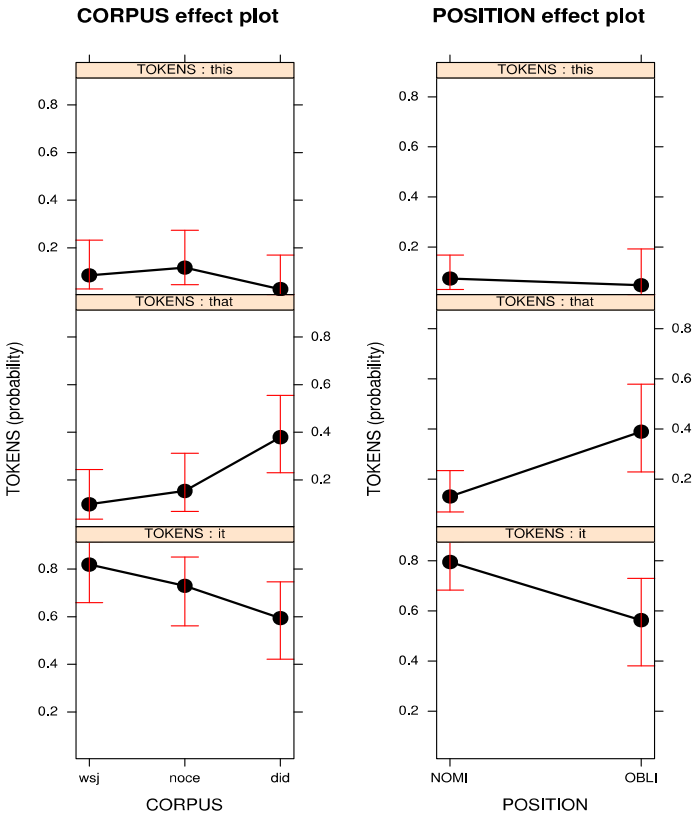

Figure 44: Effects of multinomial regression model on the selection of pro-forms

The effects of the model can be plotted for a graphical view of the way the pro-forms are used in relation to the corpus, or the syntactic position in which they appear (see Figure 44). The confidence intervals are represented with the vertical double square brackets. Regarding *that*, the intervals of the CORPUSdid variable do not overlap with CORPUSwsj, which confirms the significance of the variable with this value. Similarly, regarding *that*, the interval of the POSITIONobli variable does not overlap with POSITIONnomi, which confirms its significance with this value. For the other variable values, confidence intervals overlap, which indicates a lack of significance.

6.3.2 Discussion

The above analysis helps uncover specifics of the pro-form system among learners of English in comparison with natives. One first interesting aspect is the fact that the model selection process, based on the quality of the model and its component

predictors, drops the CONTEXT predictor. This suggests that, within the pro-form system, the endophoric/exophoric distinction is not a strong constraint for the selection of one of the forms. In other terms, it seems that choosing between *it, this* and *that* does not depend on the context whatever the L1.

The model shows that the POSITION variable is a significant predictor as it helps classify forms with a strong effect on *that*. The oblique case with *that* appears to be an interesting result since it shows a distinctive tendency of learners of English. The trend is more prevalent for L1-French learners than L1-Spanish learners, which may also be due to a change in mode from written to oral. Nevertheless, the fact remains that both learner corpora show a stronger preference for *that* v. *it* than the native corpus. This distinction can be seen as the preference of *that* over *it* when the L1 changes. It clearly suggests that learners tend to give *that* a rhematic value when using it as a pro-form. In other terms, they tend to comment on the entity which they refer to rather than place the entity as the main topic of their utterances. This may show that learners prefer the use of the pro-form *that* as a repository of their comments over what is being said about the discourse topic. Their preference for *that* in the oblique case may also confirm the confusion that exists within the pro-form system and signal errors in the selection of *that*. Learners seem to be attracted by *that* more than natives and this difference might translate into the level of errors observed in the predictions.

The work on the Relative Risk Ratios seems to bring evidence of the substitutions that exist in the pro-form microsystem described in Chapter 3. The fact that learners tend to prefer *this* or *that* to *it*, depending on their L1, supports the idea that some negative transfers are at play. The Spanish learners' preference for *this* may stem from the possibly predominant use of *esté* in Spanish. A Spanish corpus frequency analysis of the demonstrative would need to be performed to confirm this. The RRR regarding the oblique case appears to be another interesting finding. Oblique *that* pro-forms tend to replace *it* in learner English language. This supports the idea on substitutions and shows the directions these take.

The model also shows that the CORPUS predictors are correlated with the use of *this* or *that* over *it*. Strong significant corpus effects appear when the selection switches from *it* to *that* for the Diderot-LONGDALE corpus. In comparison with natives, the choice of *that* is stronger among French L1 learners than Spanish L1 learners. This may be due to a mode difference and is consistent with Biber's finding on the overuse of *that* in conversation (Biber *et al.* 1999, 349). In other terms, this may indicate that the genre plays a significant role rather than the learners' L1. Spoken data of Spanish L1 would be needed to confirm this.

Overall, the analysis shows that predictions indicate a preference for *it* when all corpora are taken into consideration. However, there are differences in preferences as *that* pro-forms are handled differently by learners in comparison with natives. There is a correlation between the choice of the pro-forms and the type of L1 and the syntactic position of the pro-form.

6.4 Summary

This chapter is directed towards the analysis of the pro-form microsystem that seems to govern the use of *it*, *this* and *that* by learners. By following a general-to-specific approach through two case studies, we bring evidence of L1-specific use of the pro-forms. Not only does our contribution highlight results on frequency tendencies, but we also manage to explore the substitutions that occur in learner uses of the pro-forms.

The logistic regression model applied to *this* and *that*, regardless of their function, shows that the selection of the forms is corpus-dependent. Results indicate that the variations of uses between L1s are linked to the functional realisation, which supports the possible existence of the two specific microsystems theorised in Chapter 3. The pro-form is a significant factor for the selection of *this* among Spanish speakers of English and for the selection of *that* among French learners. The determiner function for *this* is significantly favoured in learner corpora. The choice of the two forms is also correlated with the type of context. *This* tends to be

more strongly linked with exophoric rather than endophoric contexts. Interestingly enough, the oblique or nominative case is not significant when assessing all functions of the forms.

The exploration of the pro-form function confirms the existence of the pro-form microsystem among learners as differences appear between learners and natives. A multinomial regression model shows that, globally, *it* pro-form appears to be favoured by all speakers but learners do experience difficulties in using the forms in relation to *it.* There are underlying trends which can be reported. French learners have a higher chance of substituting *it* with *that* in comparison with natives and Spanish learners. Regarding *this,* Spanish learners are more likely to substitute *it* with the form in comparison with natives whilst it is the opposite for French learners. The model shows that the positional variable is significant within the pro-form microsystem. It seems that oblique *that* is the main point of difficulty for learners. The oblique case increases the chances of selecting *that* rather than *it.* Conversely, it decreases the chances of getting *this* rather than *it.*

These results advocate for the need to approach learner corpus research with multiple, interoperable and richly-annotated corpora so as to support multi-factorial quantitative analyses of various language markers in speech. In the next chapter, we report experiments based on the same corpus framework but instead of applying a regression modelling approach on limited samples or corpora, we use machine learning tools, called memory-based learners (MBL), to automatically analyse full learner and native corpora. This MBL system helps explore the corpora according to features whose significance is estimated with entropy-related indicators. In a different manner from the regression modelling approach, this technology probes corpora to explore the pro-form microsystem and to contribute to automatically detecting learner errors.

Because we have annotated many parameters, statistical modelling can be very complex (Bates *et al.* 2015). We have kept a low profile in the regression modelling

techniques, sampling our observations to preserve their independence and using a stepwise method to limit the number of variables. Chapter 7 will allow us to explore all the various features we have coded. This view from above nevertheless offers insights as to what learners 'do' and this kind of probabilistic model contributes to modelling learners tendencies.

Chapter 7 Machine Learning for automated analysis of *this* and *that*

This chapter shows how a specific family type of machine learning tools, called memory-based learners (MBL), can be used to automatically analyse learner corpora. We argue that NLP and more specifically machine learning tools can provide a strong support in the domain of learner English. With the growing size of learner corpora, they help sort and predict data in ways that have not been possible so far. Flach defines machine learning as "the systematic study of algorithms and systems that improve their knowledge or performance with experience." (2012, 3). Our assumption is that if an MBL can predict learner language, its internal rules somehow simulate human speakers' rules. By using an MBL system on learner language tasks, we intend to analyse the learning strategies of the system. Consequently, exploring the algorithms that help the system build knowledge might help uncover learner-specific uses of English. In sum, the principle is to use this technology to simulate human cognitive tasks:

> Memory-based learning is founded on the hypothesis that performance in cognitive tasks is based on reasoning on the basis of similarity of new situations to stored representations of earlier experiences, rather than on the application of mental rules abstracted from earlier experiences (Daelemans *et al.* 2010, 20).

Concerning *this, that* and *it*, it can be argued that humans' way of choosing one of the forms may rely on similarity reasoning. Consequently, we use an MBL system to learn to connect forms with their contexts of occurrence. The system's decisions can then be read to understand the prediction errors and the correct choices. In this chapter, an MBL system is used in two different experiments with two distinct purposes. Firstly, the system is used to successively model native and learner

language in order to investigate the specifics of pro-forms attached to each L1. The tool's internal "intelligence" built during successive learning phases is used to explore the pro-form microsystem. In the second experiment, the system is used to model native language to learn canonical uses of pro-forms. By running the system on learner language, the experiment's purpose is to detect idiosyncrasies.

The MBL tool we use is TiMBL and it was briefly described in Section 3.1.2.2.1. In addition to belonging to the machine-learning family type, there are two reasons why this tool is chosen in our experiments. Firstly, TiMBL has already been applied successfully in several language research areas such as syntax or morpho-phonology. For instance, Arndt-Lappe's research (2011) on compound stress shows positive results in which TiMBL outperforms every other tested tool. Secondly, the TiMBL data format is now part of the CESAX output format, which means that, in our case, anaphora-related annotation of *it*, *this* and *that* could be extracted for subsequent analysis in the MBL system. TiMBL's working principle relies on two components. The learning component is memory-based and the performance component is similarity-based. In the learning phase, the system adds training instances to its memory. An instance is a sequence of feature-value pairs followed by the class to assign to that particular sequence. The performance component classifies new instances on the basis of their similarity with already existing instances. The most frequent category among a set of most similar examples is assigned to the new instance, hence the term *classification*. In other terms, similarity between the different utterances, re-interpreted as a line incorporating the various parameters to analyse, is computed with specific metrics.

In this chapter, we use TiMBL to process corpora and to achieve two tasks. Firstly, in Section 7.1, the learning component of the system is successively used on native and learner language to explore the different behaviours with regard to the pro-form microsystem of *it*, *this* and *that*. Secondly, in Section 7.2, the performance component is used to detect learner errors on *this* and *that* when used as determiners or pro-forms.

7.1 Machine learning for the exploration of the pro-form microsystem

In this section, the principle is to use NLP tools to "provide specific analyses of the learner language in the corpus" (Meurers 2015, 537). We use the TiMBL system to analyse the pro-form microsystem and to see if we obtain similar results to those of the study presented in Section 6.3. The approach principle remains the same as in the previous chapter, *i.e.* to identify relevant linguistic features related to the selection of a form. Nevertheless, it is different because, instead of using small corpus samples to test hypotheses—as in the previous chapter—we choose large corpus subsets which undergo a classification process on the basis of distance metrics that compute similarity. These metrics provide information on the relevance of linguistic features and their impact on the selection of a form. By ordering features, it is possible to narrow down the most relevant features. In this case, the classifier is not simply used to classify with the highest degree of performance but its mechanism provides an insight into the features that are used for the classification. In this approach, we intend to see if we obtain the same results on the pro-form microsystem as in the regression-based approach. By using an approach that relies on entropy and Information Gain rather than multinomial regression, the data are sorted differently but the observations on the pro-form system are expected to be corroborated. Section 7.1.1 covers the algorithm and metrics chosen for the use of the classifier. In Section 7.1.2, the pro-form microsystem is explored within each corpus and the relevant features are compared.

7.1.1 Memory-based learning: selecting the algorithm and metrics for classification

The objective of this work is to explore the interactions of *this* and *that* with each other as well as *it*. In the previous chapter, the deictic system was analysed statistically by using logistic regression modelling. This method relied on decomposing an utterance into a set of features such as words, PoS, contextual and positional annotation. For instance, the following utterance from the Diderot-

Longdale corpus “If I do this as a singer” (DID0038-S001) is decomposed as the following set of features: *if IN I PRP do VBP this TPRON as IN a DT singer NN ENDO OBLI did* (see Section 6.1 on how features were used as variables whose value points were spread into matrices). The problem is that the number of features—used as variables—that such a model can handle makes interpretation more and more complex with the increasing number of interactions between variable values. By using an MBL system such as TiMBL (Daelemans *et al.* 2010), it is possible to handle more variables, also called *features*, and to extract information regarding the relevance of these features. During the training and classifying phases of TiMBL, relevance values are assigned to features. These values provide information on the features that mostly contribute to the classification performance. In short, TiMBL allows linguists to fine-tune the parameters taken into account by means of a technique called "gain ratio". This (somewhat technical) subsection explains the basics of this computation of the distance between instances.

TiMBL is a memory-based learner program that implements several methods and algorithms in order to carry out the classification task. The user can choose from a wide selection of algorithms ranging from k-nearest neighbours to incremental edited-memory based learning (Daelemans *et al.* 2010, 20–33). As well as algorithms, the program is also set to provide a range of distance metrics that can be used to compare new instances to already stocked instances in the memory. Among these distance metrics, it is possible to choose specific weighting options in order to introduce the variable impact that the features play on the selection of classes.

For the purpose of our experiments, TiMBL is used with specific settings, among which the algorithm. TiMBL offers an improved implementation of the IB1 algorithm initially developed by Aha, Kibler, and Albert (1991). In its original version, IB1 measures the distance between two instances by summing up the differences between the features. IB1 relies on the k-nearest neighbour algorithm in which an instance is compared with a *k* number of most similar examples (the k-

nearest neighbours). However, the original IB1 algorithm does not include a weight differentiation between the features that constitute the instance. In their implementation of IB1, called IB1-IG, TiMBL's authors have introduced the possibility to add weights to features that reflect domain knowledge. In other terms, TiMBL can be set to compute the weights of the features depending on their relevance in the training dataset. The way this relevance is computed is grounded in information theory and the notion of information gain exemplified in Quinlan's ID3 and C4.5 decision tree algorithms (Quinlan 1986, 87-92). The idea is to measure the worth of a feature in the classifying process. The role of Information Gain (IG hereafter) is to "measure how well a given attribute [a feature] separates the training examples according to their target classification". IG depends on the notion of entropy which "characterizes the (im)purity of an arbitrary collection of examples" (Mitchell 1997, 55). Entropy is computed with the following formula for a set of examples S:

$$Entropy(S) = \sum_{i=1}^{c} (-p_i \log_2 p_i)$$ (Mitchell 1997, 57)

where *c* is the number of classes, *pi* is the proportion of a specific value for a specific class. If entropy is 0 it means that all members of S belong to the same class. In the case of two classes the calculation could be written as: -1 . $\log^2$ 1 – 0 . $\log^2$ 0 = 0. If entropy is 1, it means that both classes include an equal number of elements. Therefore, the closer entropy is to 1 the more impure is the sample. In the case of *c* possible classes, maximum entropy is the result of $\log^2(c)$.

To illustrate this, let an arbitrary set of 14 instances include three classes: *it*, *this* and *that*. There are 9 instances classified as *it*, 2 as *this* and 3 as *that*. Entropy is calculated as:

Entropy ([9*it,* 2*this,* 3*that*])= -9/14 $\log^2$ (9/14) – 2/14 $\log^2$ (2/14) – 3/14 $\log^2$ (3/14) = 1.287

In this case, entropy is bounded by $\log^2(3)=1.584$. So the above example shows a high level of impurity in the sample. All values of the sample do not belong to the same class.

Let us focus on IG which relies on Entropy for its computation. In essence, it computes the difference in homogeneity between the entropies of a set of instances by including, or not, a specific feature. In doing so, the reduction of entropy is measured. In other words, IG shows the reduction in entropy when introducing a specific feature. IG of a feature F relative to the set of examples S is:

$$Gain(S, F) = Entropy(S) - \sum_{v \in Values(A)} \left(\frac{|S_v|}{|S|} Entropy(S_v)\right)$$ (Mitchell 1997, 58)

where *Values(A)* is the set of all possible values for feature F, and S*v* is the subset for which feature F has values v. The fraction of examples that belong to S*v* is defined as: $\frac{|S_v|}{|S|}$

For each feature of a set of instances, TiMBL implements the IG metric in order to compute by how much this particular feature reduces entropy. To illustrate this, let us take the arbitrary set of 14 instances seen above. It includes one feature called *noun,* with two different values and, again, three classes: *it*, *this* and *that*:

Values (Noun) = NN, "-" (the hyphen value means that the noun feature is inactivated/null).

The sample S can be summarised as:

S = [9*it,* 2*this,* 3*that*]

S_{NN} = [6it, 1this, 2that] (arbitrary)

S- = [3it, 1this, 1that] (arbitrary)

Gain (S, noun) = Entropy (S) – (9/14) . Entropy(S_{NN}) – (5/14) . Entropy(S-)

= 1.287 – (9/14) 1.224 – (5/14) 1.371

= 0.0103

For the *noun* feature IG is 0.0103 relative to the collection of 14 examples composed of one feature of two possible values. IG, however, poses a problem as it tends to "overestimate the relevance of features with large numbers of values" (Daelemans *et al.* 2010, 23). In the case of the datasets created by the sequencing program detailed in Section 5.3.1, this would be a problem as some features include many values when others only include a few. In order to resolve this issue, TiMBL also implements the Gain Ratio metric (Quinlan, 1993 reported in Daelemans *et al.* 2010, 23). This metric normalises IG for features with different numbers of values. The formula is defined as:

$$Gain\ Ratio = \frac{Gain(S, F)}{Entropy(S)}$$

As opposed to Quinlan's ID3 and C4.5 algorithms which use IG to identify the order of the features/nodes to build a decision tree, the TiMBL IB1-IG algorithm does not build a tree with IG indicators. Instead, it uses all the IG or Gain Ratio (GR hereafter) values computed for each feature to compute their sum. When a new instance of features is presented to the algorithm, the relevance weights of the features which are not identical are added. The result of the sum is compared to the results obtained for each instance in the memory. The most approximate value in the memory is matched to that of the new instance and the memorised instance class is also assigned to the new instance (Daelemans *et al.* 2010, 14–15).

During the training phase, TiMBL computes the metrics and, during the classifying phase, it compares new instances in light of these metrics. It then outputs results in terms of global accuracy, recall and precision. Confusion matrices can also be

produced. In the following pages, the idea is to use TiMBL to explore the behaviour of features so as to better understand how these relate to the selection of a class. With this purpose in mind, the following analytical method is followed. After obtaining the best classifying performance possible with specially selected features, GR results are looked at for each feature. This provides information on the most relevant features in the classification. By using TiMBL with different training sets from different corpora (native and non-native), it is possible to retrieve GR results for each training set and thus compare relevant features across corpora. By using the native English corpus as a training set for the classification of non-native English instances, global accuracy and confusion matrices inform on the error gap between NS and NNS.

7.1.2 *This, that* and *it* classification to retrieve linguistic features of the pro-form microsystem

In this section, we cover the use of an MBL system to operate a classification of *this*, *that* and *it* forms according to a number of features. By automatically weighting the importance of features with GR values, the MBL system highlights those that are relevant in the classification process. In Section 7.1.2.1, we present the dataset used for this experiment. Section 7.1.2.2 covers the classifying process and a discussion of the results is presented in Section 7.1.2.3.

7.1.2.1 Preparation of the dataset

As opposed to regression modelling, machine learning is carried out with large corpus subsets and not with randomly extracted samples. The dataset originates from the three corpora detailed in Section 4.2.1. For recollection purposes, it can be recalled that each corpus has received PoS-functional, contextual and positional annotations. Since each type of annotation process is not without flaws, manual verification is necessary. This labour-intensive task is repeated for all three annotation layers on three subsets of the corpora. The WSJ subset contains 40,000 words, the Diderot-LONGDALE 65,288 tokens (which include words and signs) and

the NOCE subset includes 14,517 tokens. As explained in Section 5.3.1, each of these subsets is pre-processed with a sequencing program in order to convert them into feature matrices. As a quick reminder, it can be said that pre-processing the annotated corpora outputs sequences of features with their class. Pre-processing is carried out on all form types of *it*, *this* and *that*.

Since the purpose of the experiment is to analyse interactions between pro-forms, the sequenced subsets must be stripped of forms and sequences that are not of this function. Consequently, non-referential *it* forms, determiner forms but also relative pronouns, adverbs and complementiser forms of *that* and *this* (when they apply) are eliminated from the dataset to only retain sequences that reflect the pro-form microsystem. Stripping the sets of non-pro-form instances is done via an R script (see script in Annex M) in which all forms are standardised in terms of number and lowercase.

As a result, the three subsets include the same number of features with the same possible values. The feature values are sourced from the same annotation data used through all three corpora. Table 59 summarises these features and gives details on their possible values. The table is an extension of the features analysed so far. It must be reminded that the feature specifications and their justifications are explained in Section 5.3.1.1.

Feature 1	3rd token to the left of the form
Feature 2	3rd PoS tag to the left of the form
Feature 3	2nd token to the left of the form
Feature 4	2nd PoS tag to the left of the form
Feature 5	1st token to the left of the form
Feature 6	1st PoS tag to the left of the form
Feature 7	1st token to the right of the form
Feature 8	1st PoS tag to the right of the form
Feature 9	2nd token to the right of the form
Feature 10	2nd PoS tag to the right of the form
Feature 11	3rd token to the right of the form
Feature 12	3rd PoS tag to the right of the form
Feature 13	Context (endophoric or exophoric)
Feature 14	Position (Nominative or oblique)
Feature 15	Present simple verb in 3-gram co-text of the form or not
Feature 16	Past simple verb in 3-gram co-text of the form or not
Feature 17	Negation in 3-gram co-text of the form or not
Feature 18	Coordination conjunction in 3-gram context of the form or not
Feature 19	Capital letter on the form or not
Feature 20	Punctuation in the 3-gram co-text of the form or not
Feature 21	Noun (plural , proper) in the 3-gram context of the form
Feature 22	Pro-form or pronoun in the 3-gram co-text of the form
Feature 23	Wh- form or relative pronoun in the 3-gram co-text of the form
Feature 24	Introductory preposition in 1-gram to the left of the form
Feature 25	Preposition in 1-gram to the right of the form

Table 59: List of features after pre-processing the pro-form subsets of the three corpora

All this makes the sets interoperable via the classifier. This illustrates the fact that corpus interoperability can not only be achieved at the annotation level by choosing the same annotation scheme, but also at feature level by creating matrices of identical structure and therefore comparable results.

7.1.2.2 Running the classifier and the results

Running the classifier ensures that the forms are adequately distinguished by the set of our 25 features presented in the previous section. The first thing to do is to "teach" the system enough information to properly classify data in a later stage. The standard procedure is to train the MBL system with a training set and then to

apply the trained system to a test set of new data and see how well it performs. There are different strategies for testing with a test set. The leave-one-out strategy, chosen in the following experiment, successively details the data either for training or testing, which allows the use of one sample for training and testing.

The experiment is carried out in two stages. The first one (Section 7.1.2.2.1) is to use the native subset as a standard for comparisons. The aim is to use the native subset for training before running the classifying process on each of the three subsets. This provides a measure of the difference in classification performance between subsets. The second stage (Section 7.1.2.2.2) focuses on the feature-ordering process carried out in the training phase. Each subset is used for training and the aim is to compare the respective orderings.

7.1.2.2.1 Training on native English and classifying native and learner English

The first stage relies on the same training file to classify instances. In other terms, the classifier is trained once to perform three classifying processes. The WSJ subset is used to train TiMBL before classifying new test instances. The first execution of the program is carried out on native instances of the WSJ. As the size of the WSJ subset is limited and, because we want to use the largest possible subset for training, the leave-one-out option is used with the WSJ subset (Mitchell 1997, 235). The advantage is that it maximises the size of the training file and still provides instances for testing. The second execution of TiMBL is on the NOCE subset (used as a test subset) and finally the program is run on the Diderot-LONGDALE (also used as a test subset). Table 60 summarises the overall accuracy results obtained for each subset (total number of correctly classified forms over the total number of forms).

Training subset	WSJ	WSJ	WSJ
Test subset	WSJ	NOCE	Diderot-LONGDALE
Global accuracy	0.731225	0.520000	0.700104

Table 60: Global accuracy results after TiMBL classification of three subsets with the same native training set

Classification on new WSJ instances shows 73.12% accuracy, which is rather low. When the classification is performed on the two non-native subsets of the corpora, the performance drops for the L1 Spanish subset (52%). It is stable for the L1 French subset. The rate obtained on the native subset can be used as a standard for comparison. If the learner instances that are tested were very close to those of the training file, *i.e.* WSJ instances, then the results should be expected to be the same as the classification of the WSJ instances. However, in this case, the NOCE instances are not classified as well as the tested WSJ instances. Consequently, there are two parts to look at in the error rate of the classifier. Firstly, errors may originate from the algorithm and, secondly, they may stem from learner errors. In other terms, if the algorithm's deficiency leads to errors, as with the native subset, other errors may be traced back to learner errors in the use of the forms. In fact, if learners make use of the forms in specific conditions that are different from native use, the classifier may not recognise those cases and produce errors. The following confusion matrices provide more details. Each column represents the instances in a predicted class, while each row represents the instances in their actual class (Daelemans *et al.* 2010, 18). For instance in Table 61, 7 indicates 7 forms of *this* that are predicted as *this* by the system and that actually are *this*. In other terms, the predictions are correct. Conversely, 2 shows that 2 actual occurrences of *that* are incorrectly predicted as *this*, which corresponds to type I errors (False Positives or FP when reading vertically). Number 4 shows that 4 actual occurrences of *this* are incorrectly classified as *that* (False Negatives or FN when reading horizontally).

		Predicted classes		
		This	That	It
Actual classes	This	**7**	4	11
	That	2	**17**	21
	It	9	21	**161**

Table 61: Confusion matrix after classification of WSJ instances

		Predicted classes		
		This	That	It
Actual classes	This	**5**	10	30
	That	2	**2**	18
	It	9	27	**97**

Table 62: Confusion matrix after classification of NOCE instances

		Predicted classes		
		This	That	It
Actual classes	This	**16**	13	29
	That	6	**64**	228
	It	51	247	**1260**

Table 63: Confusion matrix after classification of Diderot-LONGDALE instances

Our interpretation of the results relies on a comparative approach of the confusion matrices. We compare the results of the learner corpus matrices with those of the native corpus matrix. In the native matrix, the predicted classes correspond to the choices of the classifier and include classifying errors. In the learner matrices, the predicted classes can be interpreted as the native choices in comparison with what the learners actually chose. Results on the learner corpora are an extension of the WSJ classification to NOCE and Diderot-LONGDALE so they include two kinds of errors: those related to classifying errors, as is the case with the WSJ, and those related to learner errors. The learner error rates need to be investigated in comparison with the native error rates so as to see if there are differences indicating learner-specific errors. There are error rates for each form and their computations include counts of the two other competitor forms. For instance, in the case of *this* in the WSJ, 7 occurrences are correctly classified and a total of 2 *that* +9 *it* + 4 *that* + 11 *it* are FPs and FNs. As we are interested in the contribution of each form to the error rates, it is relevant to split the error rate according to each competitor form. We compute the error rate linked to a specific FP and FN of a form. For instance, in the WSJ, the 2.37% error rate linked to correct *this* includes *that* FPs and FNs. It is obtained by dividing (2 *that* FPs + *4 that* FNs*)* by the total number of occurrences of all forms, *i.e.* 253. This yields an error rate for the *this-that* pair. The *that-this* pair yields the same rate since *this* FPs and FNs remain the same. Table 64 presents the error rates for each pair in each matrix.

	WSJ	NOCE	Diderot-LONGDALE
This-that	2.37%	6.00%	0.99%
This-it	7.91%	**19.50%**	4.18%
That-it	16.60%	**22.50%**	**24.82%**

Table 64: Error rates for each form

This table shows which error rates of the learner corpora are similar to or distinct from the native corpus (we consider the figures in bold showing a difference of at least 5% with the native corpus). Caution must be taken as there may be a size effect due the high number of *it* forms in comparison with the number of *this* and *that* forms. In the case of the NOCE, it appears that the error rates are much higher than the WSJ for *this-it* and *that-it*. In other terms, this suggests that these rates include learner errors as well as misclassification. When comparing the Diderot-LONGDALE with the WSJ, only errors on *that-it* appear to be much higher than in the WSJ. This suggests that learner errors occur between *that* and *it* in the French-L1 corpus. In sum, these rates may be evidence of the main confusions that learners experience when establishing referential procedures. Nevertheless, these confusions do not give any indication of their directions, *e.g.* whether Spanish learners use *this* instead of *it* or vice versa. In addition, these rates are the result of the sums of FP and FN values, which in turn might hide strong specific FP or FN differences due to compensations.

Assuming the classifier emulates the grammar of natives, cases of misclassifications are likely to be (metaphorically) interpreted as non-native representations. Two cases need to be distinguished: false positives and false negatives. To investigate the directions within confusions, we can look at the False Positive Rates (FPR) and the False Negative Rates (FNR) in two two-dimensional tables crossing non-matching forms with the corpora (see Table 65 for FPR and 66 for FNR). FPR answers the following question: When the learners choose an actual form, how often do the natives choose another one? So FPR represents native expected preferences when learners choose a form. FNR answers the following question:

When a form is predicted as native, how often do the learners actually choose another form? So FNR represents learners' choices in native expected situations.

Learner *this*	WSJ	NOCE	Diderot-LONGDALE
FPR that	18.18%	22.22%	22.41%
FPR it	50.00%	**66.67%**	50.00%
Learner *that*			
FPR this	5.00%	9.09%	2.01%
FPR it	52.50%	**81.82%**	**76.51%**
Learner *it*			
FPR this	4.71%	6.77%	3.27%
FPR that	10.99%	**20.30%**	15.85%

Table 65: FPR for each actual learner form (WSJ used as gold standard)

Native *this*	WSJ	NOCE	Diderot-LONGDALE
FNR that	11.11%	12.50%	8.22%
FNR it	50.00%	56.25%	**69.86%**
Native *that*			
FNR this	9.52%	**25.64%**	4.01%
FNR it	50.00%	**69.23%**	**76.23%**
Native *it*			
FNR this	5.70%	**20.69%**	1.91%
FNR that	10.88%	12.41%	15.03%

Table 66: FNR for each actual native form (WSJ used a gold standard)

We compare the learner corpora with the native corpus. Some rates are 10 percentage points over the WSJ gold standard, which indicates an important difference. This may be explained by the presence of learner errors rather than just algorithm errors. The rates are shown in bold figures. For instance, for FPR, it appears that Spanish learners choose *this* when *it* should be favoured by natives in 66.67% of the cases (Table 65). This represents a 16.67% difference with the WSJ. In other terms, Spanish learners tend to replace native *it* by *this*. As regards FNR, when a native *this* is predicted, French learners actually choose *it* in 69.86% of the cases (Table 66). The WSJ rate is only 50%. In this case, learners tend to replace *this* with *it*.

By proceeding in this manner for each table and by crossing the results shown in bold, it is possible to highlight the main tendencies within confusions. Table 67

shows a synthesis of learners' confusions depending on the actual forms which should be selected in the same contexts.

Native	French	Spanish
this	it	
that	it	this, it
it	that	this, that

Table 67: Learners' preferences in comparison with natives

Regarding the *this-it* confusion French learners tend to replace *this* with *it* whereas Spanish learners tend to replace *it* by *this*. Regarding the *that-this* confusion, only Spanish learners tend to replace *that* with *this*. For the *that-it* confusion, French and Spanish learners tend to choose *that* instead of *it*.

7.1.2.2.2 Training and classifying on native and learner English

The second stage of the experiment focuses on the factors that impact the selection of a form. In this stage, we are interested in seeing how choices are made within each L1. In other terms, we want to see how the corpus features described in Table 59 (page 308) are classified, *i.e.* which GR weights are assigned to the features during the training phase. Since we want to see this process for each L1, the classifier needs to be re-trained with L1-specific subsets. By using the leave-one-out option, each subset is used for training and classifying. In total, there are three different training and classifying phases. Compared with the results presented in Table 60 (page 309), the global accuracy is better for the learner subsets but size effects due to the high number of *it* forms cannot be ignored. Table 68 gives the performance results of each classifying process. Training the classifier with an L1-specific subset helps improve its classification, which suggests that the feature ordering may be corpus specific.

Training subset	WSJ	NOCE	Diderot-LONGDALE
Test subset	WSJ	NOCE	Diderot-LONGDALE
Global accuracy	0.731225	0.565000	0.809300

Table 68: Global accuracy results after TiMBL classification of three corpus subsets with training sets of each of the corpora

During each training, a corpus-specific ordering of features as explained in Section 7.1.2.1, is carried out. The differences between the three orderings inform the researcher about the features that most impact each subset. What is necessary to examine is the computation of feature weights by TiMBL for each training subset. It must be recalled that GR is the metric used to compute the relevance weights of the features. There are three lists of GR values resulting from the training of the three subsets (see Table 69 for details). The values in bold correspond to those that belong to the top 25% weights of each list. For instance, the second line of the table shows that the feature called *3rd token to the left of the form* is among the 25% most important features after training each subset.

Features	WSJ	NOCE	DID
3rd token to the left of the form	**0.111**	**0.125**	**0.034**
3rd PoS tag to the left of the form	0.041	0.041	0.009
2nd token to the left of the form	**0.110**	**0.123**	**0.042**
2nd PoS tag to the left of the form	0.039	0.063	0.013
1st token to the left of the form	0.088	**0.105**	**0.033**
1st PoS tag to the left of the form	0.057	0.045	0.010
1st token to the right of the form	**0.096**	**0.114**	**0.056**
1st PoS tag to the right of the form	0.069	0.061	0.025
2nd token to the right of the form	**0.106**	**0.114**	**0.045**
2nd PoS tag to the right of the form	0.049	0.065	0.028
3rd token to the right of the form	**0.103**	**0.118**	**0.039**
3rd PoS tag to the right of the form	0.042	0.040	0.011
Context (endophoric or exophoric)	**0.175**	0.046	0.022
Position (Nominative or oblique)	0.001	0.012	**0.043**
Present simple verb in 3-gram co-text of the form or not	0.003	0.010	0.008
Past simple verb in 3-gram co-text of the form or not	0.016	0.038	0.002
Negation in 3-gram co-text of the form or not	0.005	0.004	0.004
Coordination conjunction in 3-gram context of the form or	0.024	0.017	0.008

not			
Capital letter on the form or not	0.089	0.006	0
Punctuation in the 3-gram co-text of the form or not	0.037	0.003	0.001
Noun (plural, proper) in the 3-gram context of the form	0.011	0.029	0.003
Pro-form or pronoun in the 3-gram co-text of the form	0.003	0.003	0.001
Wh- form or relative pronoun in the 3-gram co-text of the form	**0.114**	0.073	0.008
Introductory preposition in 1-gram to the left of the form	0.007	0.011	0.011
Preposition in 1-gram to the right of the form	0.045	0.044	0.0009

Table 69: List of features and their respective GR values in relation to the corpus subsets

In order to better understand the GR values for each subset, they can be visualised with two types of charts. Firstly, a strip chart of the GR distribution of each subset (Figure 45) shows how the data are spread along the same scale. For the computation of the graphs, the same jitter option adds a small value to each data point to avoid superimposed data points. From the graphs, it is evident that all three distributions differ. For each of the subsets, data points are split into two or three groups but the WSJ's and the NOCE's GR-value distribution spans over much of the entire spectrum whilst the Diderot-LONGDALE's data points are located on the first quarter of the x-axis. If we assume that high GR values indicate strong predictors, the Diderot-LONGDALE distribution raises the question of whether relevant features have been captured. This observation is in contradiction to the 0.80 overall accuracy reported for the subset in Table 68 as it implies that good predictors have been found to provide good classification.

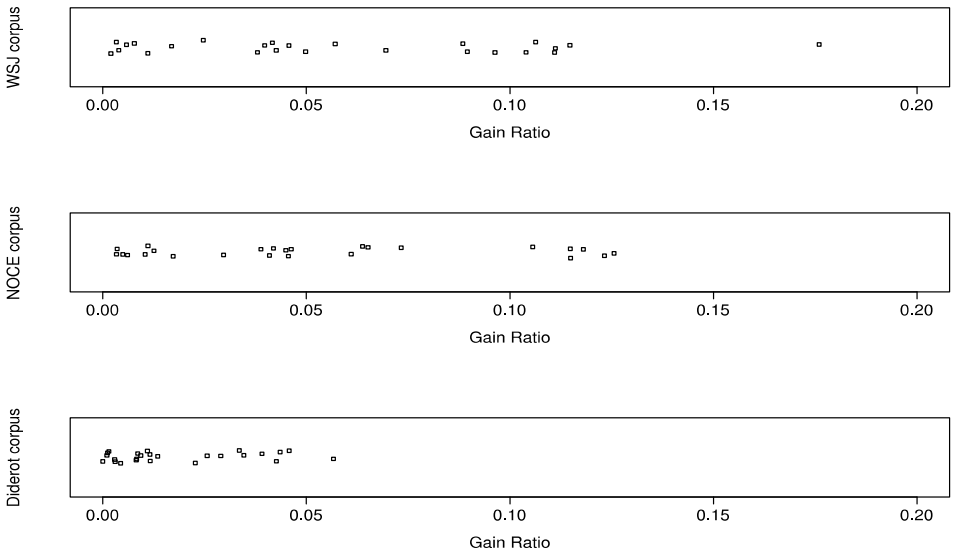

Figure 45: Strip charts to show the distribution of GR values across subsets

Secondly, a box plot chart (Figure 46) provides a more comprehensive understanding as it also allows visualisation of the medians and quartiles of the data values and distribution. Visualisation shows that while the WSJ's and NOCE's GR values are comparable in medians, they differ as far as the quartiles are concerned. The WSJ subset shows a greater top 25% of values than the NOCE subset. The Diderot-LONGDALE median is lower than that of the two other subsets. The smaller size of the box suggests that most of the data points are within the second and third quartile.

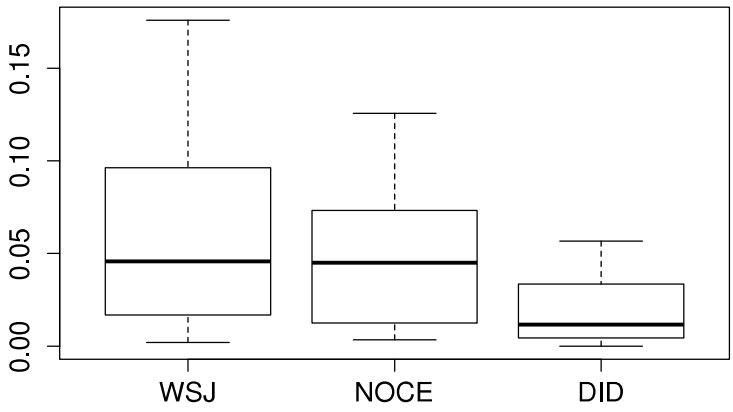

Figure 46: Box plot of GR values of features for each corpus

A more accurate visualisation of the quartiles in Table 70 shows the starting levels of values of each quartile. This helps with determining those values that most impact the classifying process. Among others, the 75% column shows the values that cut off the top 25% of the values of each corpus subset. In other terms, all the values above these particular values are the most influential in the classifying process. For instance, all values over 0.0963 in the WSJ subset form the highest quarter of GR values. This information is used later to order the features that match these values and observe possible differences. The values can be read in Table 69 but, for better reading, Table 71 (page 321) presents the top quartile results.

Quartiles	0%	25%	50%	75%	100%
WSJ	0.0020	0.0169	0.0457	0.0963	0.1760
NOCE	0.0034	0.0126	0.0450	0.0733	0.1256
Diderot-LONGDALE	0.0000	0.0044	0.0116	0.0335	0.0567

Table 70: Quartiles of GR values per corpus subset

Comparing the distribution of different GR values for the three subsets raises the question of the significance of the actual difference between these samples. In other words, the point is to know whether the difference in GR distribution between the WSJ and the NOCE is significant. This is tantamount to stating the following hypotheses:

H0: The distribution of the GR values (dependent variable) does not differ depending on whether they belong to one corpus subset in relation to another one (independent variable).

H1: The distribution of the dependent variable (GR) differs depending on the corpus variable.

To provide an answer, the two-sample Kolmogorov-Smirnov (K-S) test can be applied. It is based on computing the difference between empirical cumulative distributions (Gries 2013 [2009], 176). As its name indicates, it is a two-sample test which implies that it must therefore be applied to three different combinations of the three corpus subsets. The first combination corresponds to testing the WSJ and NOCE corpus subsets. The following results are obtained with R:

```
> ks.test(a$GR, b$GR)
Two-sample Kolmogorov-Smirnov test
data:  a$GR and b$GR
D = 0.16, p-value = 0.915
alternative hypothesis: two-sided
```

In this case the p-value is superior to 0.05 and thus H0 cannot be rejected and there is no significant difference between the two corpus subsets as far as GR values are concerned. The WSJ Diderot-LONGDALE combination can also be computed with the same test:

```
> ks.test(a$GR, c$GR)
Two-sample Kolmogorov-Smirnov test
data:  a$GR and c$GR
D = 0.48, p-value = 0.005614
alternative hypothesis: two-sided
```

In this case, the p-value is inferior to the 0.05 threshold and H0 can be rejected. The test confirms what can be observed in the first and third strip charts of Figure 47. The difference between the GR value of the WSJ subset and the Diderot-LONGDALE subset is very significant. The maximal absolute difference D is 0.48. To conclude with the differences between corpus subsets and their GR indicator, the difference between the NOCE and the Diderot-LONGDALE subsets must also be explored:

```
> ks.test(b$GR, c$GR)
Two-sample Kolmogorov-Smirnov test
data:  b$GR and c$GR
D = 0.44, p-value = 0.01484
alternative hypothesis: two-sided
```

The p-value is inferior to 0.05 and it shows that H0 can be rejected. Hence, the difference between both corpus-subset GR values is significant: D=0.44. Figure 47 summarises the findings. The Diderot-LONGDALE subset clearly differs from the two others in terms of cumulative values. For the Diderot-LONGDALE subset, GR values under 0.05 account for more than 90% of all values while it is not the case for the two other subsets. This explains why the K-S test showed a greater D between the Diderot-LONGDALE and either of the two subsets. It also raises the question of the correlation of these results with the 0.80 overall accuracy of classification: do more concentrated values lead to better classification?

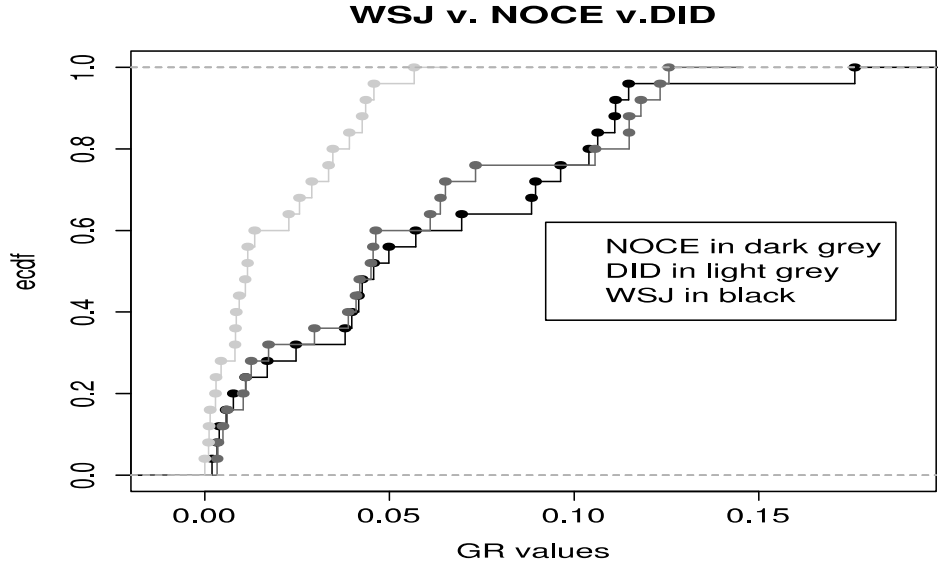

Figure 47: Empirical cumulative distribution functions of the GR values per corpus subset

With these results in mind, it is interesting to look at the top quartiles of the values for each subset and identify the features that most impact the classifying process. To present the results in a readable manner, Table 59 is reprinted with extra details regarding the top quartile (including the threshold values) in which they can be

found. The purpose is to link the feature to its rank in the quartile of a specific corpus subset. To do so, the rank of features is appended to the corpus abbreviation. For instance, WSJ2 refers to the second feature in order of GR importance in the WSJ subset.

Features	Description	Subsets for top quartile
Feature 1	3rd token to the left of the form	WSJ3; NOCE1; DID6
Feature 2	3rd PoS tag to the left of the form	
Feature 3	2nd token to the left of the form	WSJ4; NOCE2; DID4
Feature 4	2nd PoS tag to the left of the form	
Feature 5	1st token to the left of the form	NOCE6; DID7
Feature 6	1st PoS tag to the left of the form	
Feature 7	1st token to the right of the form	WSJ7; NOCE 5; DID1
Feature 8	1st PoS tag to the right of the form	
Feature 9	2nd token to the right of the form	WSJ5; NOCE4; DID2
Feature 10	2nd PoS tag to the right of the form	DID6
Feature 11	3rd token to the right of the form	WSJ6; NOCE3; DID5
Feature 12	3rd PoS tag to the right of the form	
Feature 13	Context (endophoric or exophoric)	WSJ1
Feature 14	Position (Nominative or oblique)	DID3
Feature 15	Present simple verb in 3-gram co-text of the form or not	
Feature 16	Past simple verb in 3-gram co-text of the form or not	
Feature 17	Negation in 3-gram co-text of the form or not	
Feature 18	Coordination conjunction in 3-gram context of the form or not	
Feature 19	Capital letter on the form or not	
Feature 20	Punctuation in the 3-gram co-text of the form or not	
Feature 21	Noun (plural, proper) in the 3-gram context of the form	
Feature 22	Pro-form or pronoun in the 3-gram co-text of the form	
Feature 23	Wh- form or relative pronoun in the 3-gram co-text of the form	WSJ2; NOCE7
Feature 24	Introductory preposition in 1-gram to the left of the form	
Feature 25	Preposition in 1-gram to the right of the form	

Table 71: Top quartiles of subsets in which features can be found with their rank

There are a number of observations that can be made from this table. First of all, we can focus on the features which are only common to all three top quartiles. The striking element is that the most relevant features calculated by the algorithm are those which are not PoS related. Whatever the subset, token-type features show high relevance weights. This means that, in all three subsets, lexicon is very relevant to predict forms. We may also wonder why the 3rd and 2nd tokens to the left are more relevant than the 1st one to the left. This may be due to the high proportion of *it* forms and the fact that such tokens act as better predictors. An ngram/cluster analysis might provide answers in this direction. It could give indications regarding the conditional probability of preceding markers of the forms.

Secondly, attention can be paid to those features that are found in the top quartile of only one subset. In this case, features 10, 13 and 14 stand out. Feature 10 shows that the second PoS-tag to the right of the form plays an important role in the French-L1 subset. A search for the second POS-tags after any of three forms in the Diderot-LONGDALE subset reveals that the second token to the right is most likely to be an adverb (RB), a pronoun (PRP), a determiner or article (DT) or an adjective (JJ). These findings suggest joint probabilities between forms used as subjects (which is confirmed by exploration) and predicates in which adverbs (including *not*), articles or adjectives can be found. For instance, “and *that*'s really shocking” (DID041-S003) shows a typical construction in which *that* is followed by an adverb and an adjective. It may be the sign of a lack of variety in the type of predicates introduced by the forms and thus a poor level of vocabulary when speaking. Conversely, pronouns in second position to the right may indicate a new predicative relation in which the pronoun is a subject, *e.g.* “so I realized I enjoyed to do *that* so I I wanted to became a journalist” (DID045-S001). Feature 13 shows that for the native corpus subset the endophoric or exophoric context is an important factor. This feature ranks first in the quartile, which shows that, for the native subset, classification greatly depends on it. Conversely, it is not primary for the non-native subsets. In fact, the feature is part of the second quartile in both learner subsets. This seems to corroborate findings in Section 6.3 in which the type

of context is not a significant variable in the pro-form microsystem. Nevertheless, it is not irrelevant either, especially when considering the GR-value difference between the first and second quartiles. Feature 14, which is the positional feature, only appears as very relevant for the French-L1 subset. It is in the third quartile of the Spanish L1 subset and last in the native subset. The difference in feature weight importance might be linked to the French-L1 type or to the spoken mode. In sum, for French-L1 learners the choice of a pro-form might be more dependent on the oblique or nominative cases. These findings also corroborate those of Section 6.3 in which the position variable is also significant and it is more prevalent for French speakers than Spanish speakers.

Finally, some top-quartile features are common to two subsets. Feature 5 shows that the first token to the left of the form plays a rather important role in the learner subsets only, making it a learner-specific feature. Feature 23 also stands out due to the fact that it is found in the first quartile of the WSJ and NOCE subsets. The presence of a WH form or a relative pronoun in the 3-gram co-text of an *it*, *this* or *that* form appears to be a relevant factor for classification. The use of *which*, *what*, *where* or *that* as a relative pronoun appears to be closely linked with the use of a form and the classifier uses this feature to perform the classification. For the native corpus subset, it is a primary source of relevance and for the Spanish L1 corpus subset, it ranks at the bottom of the first quartile, which denotes yet a difference in use of these forms between native English speakers and Spanish learners of English. The fact that this feature is not predominant with the Diderot-LONGDALE subset classification may show that the feature is either linked to the French L1 or to the written mode. The use of such forms as determiner *which*, relative *that*, *who* or *what*, possessive *whose* opens to more complexity in the utterances with the inclusion of hypotaxis. Parataxis, in turn, corresponds to the spoken mode and thus leads to less occurrences of these forms. Consequently, the classifier cannot include them as relevant features since they are not so present.

Globally, the NOCE and the WSJ are close to each other while the Diderot-LONGDALE stands out. This may suggest that the French L1 or the spoken mode influence the relevance weights. In other terms, the pro-form microsystem seems to be primarily governed by the speech mode of French speakers. Further comparative analysis is required to confirm this hypothesis. Comparisons with a native spoken corpus could yield results that might confirm or refute one of the alternatives.

7.1.2.3 Discussion

One first point of discussion is that, even though the features are weighted, the classifier simulates a human being by reasoning on the basis of similarity and selecting forms in context. For instance, when it is trained on native English, it can be argued that it somehow simulates a native speaker. This may be debated as there is more than similarity in the production of language. However, there is no denying that, to some extent, the memorisation of sequences helps the speaker to produce language. In the context of the pro-form microsystem, one research question might be about the constraints of use of *this* and *that* in this system v. those of the same forms in the determiner system. The research on formulaic language is a line of work that shows how learners make use of language by memorising blocks of language. Granger shows that to some extent learners memorise and make recurrent uses of phrases (Granger 2001, 154). (Hasselgren 1994) points out the existence of lexical "teddy bears" as safe words which learners grasp. They may appear as single words or sets of words as in recurrent 3-grams as reported by (Götz and Schilk 2011, 87). They are tantamount to reliability islands for learners and examples such as "that's it" can be found easily in the Diderot-LONGDALE corpus. In this respect, the classifier might be seen as emulating this behaviour insofar as it matches every new instance with pre-existing ones in memory on the basis of a set of features. The fact that n-gram tokens are part of the feature set encompasses this behaviour. For instance, the recurrent "that's it" 3-gram can be captured in all three corpora as the three tokens are all collected as

features surrounding the *that* and *it* forms. The MBL system will take the 3-gram into account as part of the sequences of features it relies on for classification.

The experiment has shown that confusions in the form of substitutions do exist at learner level. The initial hypothesis brought forward in Section 3.2.2.2.1 was grounded in the observation of a sample of errors in which learners presented confusions. By applying a methodical approach based on the consistent annotation and extraction of occurrences, it has been possible to show that confusions do exist. The aforedescribed experiment provides extra information in relation to the direction of the confusions. The results suggest that regarding the *this-it* confusion, French learners tend to replace *this* with *it* whereas Spanish learners tend to replace *it* by *this*. It is difficult to understand the case of Spanish learners. In Spanish *éste, ése, aquél* are the three forms of the demonstrative system (Gerboin and Leroy 1991, 45). It would be normal to expect learners to transfer the Spanish proximal forms to *this*. Instead they seem to wrongly choose *it* in some contexts and the reason remains to be found. Regarding the *that-this* confusion, Spanish learners tend to replace *that* with *this*. This may indicate an influence from their L1 insofar as the contexts of use of the Spanish pro-forms might be different from English, leading to errors. For the *that-it* confusion, French and Spanish learners tend to choose *that* instead of *it*. This may indicate confusions in the use of pro-forms as subjects of the verb *be* which is a largely used combination in both learner subsets. Compared with the results obtained in the regression analysis in Section 6.3.1, there are some common findings. In both analyses, confusions are found between *it* and the two other forms. It appears that learners tend to replace *it* with either *this* (Spanish learners) or *that* (French and Spanish learners). To conclude on confusions, there is evidence of learner substitutions and their main directions have been identified. We now need to analyse the specific utterance patterns in which these confusions occur. To do so, the feature orders prepared during the training phase provide a line of research.

The feature orders computed and used by TiMBL to classify give a first insight into these patterns. The orders of features give indications on some of the elements that play a role in the pro-form system of each L1. It appears that the endophoric/exophoric distinction is relevant only for the native English subset. This finding is consistent with that reported in Section 6.3 in which the same distinction, within the pro-form system, does not appear to be a strong constraint for the selection of one of the forms in learner corpora. The positional feature appears to be very relevant for the L1-French sample, which is partly consistent with the findings presented in the same other section. In the multinomial model, the positional feature is also significant in the L1-Spanish sample. One question regarding the order of importance of the features is whether it corresponds to mental hierarchies in the mind of the speakers much in the same way as Keenan and Comrie identified a hierearchy for relative clauses (Keenan and Comrie 1977). In their paper, they show evidence of an accessibility hierarchy concerning the positions of the nouns that can be relativised. For instance, the subject position is the most accessible position for a relative pronoun. In our case, our findings would mean that the speakers unconsciously take the features in order of relevance in order to make their choice. The disparate orders of features, depending on L1s, show that it might be a possibility. However, the overall accuracy results of each classification (see Table 68 page 314) show that these orderings could still be improved and, in doing so, new feature orders might be expected and new features could also be discovered.

The question of classification performance appears to be an issue. The 73% obtained on the native corpus can also be interpreted in light of other experiments based on classifying other forms according to contextual features as well. An experiment was carried out on the distinction of non-referential *it* forms (Boyd, Gegg-Harrison, Byron 2005). Training and testing were done on the BNC Sample Corpus and features rely mostly on textual and PoS annotation elements. The researchers report a 92% global accuracy. Similarly, Pradhan *et al.* worked on the classification of article errors in a non-native English writing corpus. They report a

91.89% accuracy (Pradhan *et al.* 2010). In the case of our work, the classification performance clearly needs to be improved to strengthen the validity of the observations made on the feature orderings.

Some of the observations made in this experiment rely on the difference between the overall accuracy obtained with the native corpus and that of the two non-native subsets. This difference is interpreted as the margin of learner errors. In the algorithm error rate, misclassification of forms can be imputed to the messiness of data—as it is on the native corpus—and to learners. To strengthen this approach though, it would be necessary to perform a classification on another native corpus with the initial native corpus (the WSJ in this case) as the training set. This would allow the measurement of the difference of accuracy between native corpora and, indirectly give an indication on the importance of the learner error margin.

One final point of discussion is related to the methodology applied for the sampling of the WSJ training subset. There are several issues that can be raised such as the relevance of the non-introduction of the singular/plural distinction of the demonstratives, the definition of oblique cases or the number of occurrences selected for each class. The latter issue is about the number of instances to include in the training file. One suggestion for the training phase might be to take an even number of instances of each form in each class. The purpose would be to provide no mathematical preference to either of the classes, *i.e.* equiprobability of forms in the data. This was tested but accuracy results of classification on the native corpus drastically drop (maybe due to the fact that the most numerous *it* form is better handled than the two others and so, by decreasing the number of *it* instances, the model becomes less predictive), which jeopardises the grounds for further analysis. The issue can also be regarded in relation to the terms of the equations of Entropy and Information Gain presented in Section 7.1.1. These two closely linked equations rely on fractions between the number of occurrences of a specific value and the total number of occurrences of all values. Results are altered if we choose an even number of forms per class instead of keeping the existing unequal number

of forms present in the corpus subset. In other terms, choosing an equal number of occurrences alters the proportions of a specific class in the training corpus subset. Entropy calculations, as well as IG, would thus be different leading to a different order of relevance for the weights of the features. Conversely, using a training set of instances that reflects the proportion of a class in a training corpus subset, maintains the proportion of that class, which translates into specific entropies for feature values, and thus a specific IG. In fact, IG includes the proportions of each class of the training subset and it seems coherent to use a training set of instances that reflects the corpus.

7.2 Towards error detection: the automatic linguistic analysis of learner data

In this section, we present two updated experiments inspired by (Gaillat, Sébillot, and Ballier 2014). The objective is to automatically detect errors on *this* and *that* in a learner corpus. We subsequently try to uncover the linguistic characteristics that influence the unexpected and expected uses of *this* and *that* in context. The principle, as explained in Section 5.3.1, is to pass on a representation of contexts to the classifier TiMBL in order to simulate the selection process of a *this* or a *that*.

Our work builds on previous work in several domains. Error tagging of learner corpora (Dagneaux *et al.* 1998; de Mönnink 2000) has shown that learner English requires specific processing, be it manual or computer-assisted (see Section 3.1.2.1.2). Recently, machine learning technologies have been applied to automatically detect various types of errors such as article selection (Han *et al.* 2006; Pradhan *et al.* 2010). The approach chosen for the two experiments reported in this section follows the same line of research. We use a multi-layer annotation scheme for automatic classification of a learner corpus. Instead of article selection, the focus is placed on the selection of demonstratives. In these experiments, NLP tools are used to provide specific analyses of the learner language in the corpus (Meurers 2015, 537).

The first experiment is an attempt to automatically detect errors. We try to measure the impact of certain distributional features of the demonstratives on their classification as *expected* or *unexpected* forms.[35] It discriminates *expected* uses of the two forms without distinction against *unexpected* uses of the same two forms. By selecting specific PoS-tags from surrounding contexts of *expected* and *unexpected* occurrences, a classifying process is implemented to see whether or not these selected features play a role in the distinction between *expected* and *unexpected* uses. Following in Pradhan *et al.'s* footsteps (2010), the second experiment's novelty lies in the nature of the dataset, as only *unexpected* uses of the demonstratives, in their close context, are considered. By using an automatic classifier with this dataset, the goal is to see what specific linguistic features play a role in the classification process of just *unexpected* uses of *this* or *that*. In other terms, the point is to see if the set of features in *unexpected* contexts helps to predict a particular *unexpected* form. After describing the preparation of the dataset for both experiments (Section 7.2.1), results are reported for each experiment (Section 7.2.2) and a discussion (Section 7.2.3) closes the section.

7.2.1 Preparation of the dataset

In this part, we describe the two components of the dataset and we explain how they are used in relation to the two experiments.

7.2.1.1 Native corpus subset

The first subset is a sample from the Penn Treebank WSJ corpus. It is made up of 40 randomly extracted occurrences of *this* and *that*: twenty singular and plural occurrences of each form, together with their close context composed of token PoS-tag pairs. The forty occurrences are selected from the annotated subset of the

[35] The term *unexpected* was favoured over the term 'error' after tests on natives. Non-native occurrences were shown to natives. The tests consisted in presenting actual non-native utterances to natives with gaps replacing *this* and *that*. Natives were first asked to fill in the gaps. When their choice contradicted the non-native choice, they were asked to judge the non-native choice. The tests showed that natives would classify choices in three categories: acceptable, unacceptable and acceptable as a second choice. The term *unexpected* covers both the unacceptable and second-choice categories.

corpus as described in Section 4.2.1. It ensures that contextual and positional annotation is also taken into account.

7.2.1.2 Learner corpus subset

The second subset of the data is an extract from the annotated Diderot-LONGDALE corpus. For the two experiments described in this section, forty occurrences of unexpected uses of *this* and *that* are identified manually and extracted from the transcripts. Each occurrence is selected with its surrounding context and annotation. When the context includes an occurrence of a demonstrative which is expected, the context is shortened so as to neutralise any expected use of the form. Without neutralisation, expected uses of the form would also be processed and, thus, introduce a bias into the homogeneity of the dataset. This sample includes 20 occurrences of *this* and 20 occurrences of *that*, which correspond to the pro-form and determiner functions. Both singular and plural forms are selected.

The selection of unexpected uses is performed manually, and verification is carried out with a native English speaker. A form is characterised as unexpected when the native speaker considers the choice of the demonstrative as not being the obvious one. As a reminder, Section 3.2.2.2 shows that alternatives would have been substitutions with the other demonstrative or with the pronoun *it* or the determiner *the*. In other terms, unexpectedness is due to two trends: either the learners swap the two words or they swap the form with an erroneous use of *the* or *it*. In our approach, errors are not perceived as a binary concept. Instead, the concept relies on a gradual scale of native expectedness composed of syntactic and pragmatic parameters. These parameters can be partially mastered by learners, which gives this perception of gradualism to the notion of errors.

7.2.1.3 Sequencing process: subsets, features and class assignment

In this section, we present the sequencing process. The subsets and the features are described. We also show what elements are selected as classes.

7.2.1.3.1 Subsets

For the first experiment, we use both subsets described above as we want the classifier to distinguish between learner-corpus demonstratives, characterised by unexpected uses, and native-corpus demonstratives. This would allow the identification of features that differentiate *expected* from *unexpected* uses. We do not distinguish between *this* or *that* at this point, but we introduce a balanced number of the forms in the samples extracted so that classification is not influenced by proportional differences. As opposed to the sample used in Section 7.1.2.1 (page 306), there is no reason to keep the proportion of the forms. In the present case, the *unexpected* and *expected* classes are not inherent to the corpus and, thus, do not need to be reflected by weights. The even number of occurrences of *this* and *that* forms in each subset gives a 50/50 baseline with which classification can be compared. The small size of the sample is due to the slow process of identifying unexpected uses manually. The classifying method explained below takes this into consideration so as to maximise training and test data.

For the second experiment, as already mentioned, we only use the learner subset in order to distinguish unexpected *this* from unexpected *that*. It is an insight into unexpected learner English only. This is why, in a similar approach to (Pradhan *et al.* 2010), the dataset is only composed of the Diderot-LONGDALE subset described above. In other terms, only unexpected uses are taken into consideration and the classification process is carried out so as to have a closer insight into the feature that led to the selection of a particular *unexpected* form.

7.2.1.3.2 Features

The sequencing extraction of features in both corpora is explained in Section 5.3.1. A PERL script is used to compile textual, contextual and positional information in the form of matrices of features. In addition to these features, learner-specific features are also collected since experience in correcting both oral and written

productions of students helped with the identification of grammatical issues that are found repeatedly amongst students. All the features are described in Table 72.

Some learner-specific features are linked to the fact that learners make mistakes on agreement between the forms in their determiner function and the nouns which they precede. As the sequencing process neutralises the grammatical number by converting all *these* and *those* to *this* and *that* forms, it is important to add a feature to mark the category for the number of the forms. The purpose is that the singular or plural patterns of determiners and even pro-forms be taken into account in the set of features. Consequently, a PLU feature is created and, if activated, it signals that the form is plural. Similarly, an NNS feature points to cases where a determiner form is followed by a plural noun.

Some other learner-specific features are linked to the fact that specific words are usually accompanied by learner difficulties. They are also isolated during the sequencing process. *For*, *since*, *despite*, *(in) order*, *(in) spite* can all be part of direct translations from French, and as such, may appear in *unexpected* uses. The verb *is* is also isolated, as combinations with the demonstratives are not that clear for learners. In all these cases, learners make typical mistakes and the introduction of the words is an attempt to capture the environment in which errors with *this* or *that* occur. For example, some learners tend to say “In order this happen”. By selecting the word *order* as a feature for the classifier, the idea is to see whether it helps with the improvement of the error classification process.

More words are also added to the list so as to maximise classification since the purpose is to identify as many unexpected uses as possible. They have been chosen according to several semantic groups that correspond to the notions identified in Section 5.3.1. Notions such as rejection (*i.e. no*, *never*), foreground/background information and interest (*i.e. want*, *hope*, *say*, *tell*, *first*, *second*), topic continuity/discontinuity (*i.e. after*, *however*, *then*) support the introspective choice of specific words. In addition, given the fact that the demonstratives are part of the

domain of deixis, it was decided to include the words that provide referential information made by the speaker (*i.e. here*, *there*, *this*, *that*, *now*). The list of words also includes tokens expected to be found next to *this* or *that* (*i.e. 's*, *all*, *like*, *of*).

Feature 1	Specific words in 3-gram context right of the form
Feature 2	Specific words in 2-gram context right of the form
Feature 3	Specific words in 1-gram context right of the form
Feature 4	Specific words in 3-gram context left of the form
Feature 5	Specific words in 2-gram context left of the form
Feature 6	Specific words in 1-gram context left of the form
Feature 7	PoS-function of the form
Feature 8	3rd token to the left of the form
Feature 9	3rd PoS tag to the left of the form
Feature 10	2nd token to the left of the form
Feature 11	2nd PoS tag to the left of the form
Feature 12	1st token to the left of the form
Feature 13	1st PoS tag to the left of the form
Feature 14	1st token to the right of the form
Feature 15	1st PoS tag to the right of the form
Feature 16	2nd token to the right of the form
Feature 17	2nd PoS tag to the right of the form
Feature 18	3rd token to the right of the form
Feature 19	3rd PoS tag to the right of the form
Feature 20	Context (endophoric or exophoric)
Feature 21	Position (Nominative or oblique)
Feature 22	Present simple verb in 3-gram co-text of the form or not
Feature 23	Past simple verb in 3-gram co-text of the form or not
Feature 24	Negation in 3-gram co-text of the form or not
Feature 25	Coordination conjunction in 3-gram context of the form or not
Feature 26	Capital letter on the form or not
Feature 27	Punctuation in the 3-gram co-text of the form or not
Feature 28	Noun (plural, proper) in the 3-gram context of the form
Feature 29	Pro-form or pronoun in the 3-gram co-text of the form
Feature 30	Wh- form or relative pronoun in the 3-gram co-text of the form
Feature 31	Introductory preposition in 1-gram to the left of the form
Feature 32	Preposition in 1-gram to the right of the form
Feature 33	Presence of plural noun after the form in its determiner function
Feature 34	Number of the form
Feature 35 (in the case of experiment 1 only)	*This* or *that* form

Table 72: List of features used for classification of unexpected uses of this and that

It is important to specify that no one feature can be seen as leading necessarily to the choice of a particular form. Instead, the experiment aims to test whether all the features, as a whole, have an influence or not on the choice of *this* or *that*. At this point in the study, it is not possible to indicate how the influence of feature x leads to the choice of *this* in one case, or *that* in another.

7.2.1.3.3 Class assignment

For the first experiment, the program extracts features from the two subsets described above. This sequence of features is then matched to a particular class: *expected* or *unexpected*. For all the lines of features extracted from the native subset, the *expected* class is assigned. For all the features extracted from the learner subset the *unexpected* class is assigned. Once the classes are assigned, both subsets are merged so as to finalise the training and test sets for the classifier. Figure 48 is a partial view of the output file following the sequencing process where linguistic characteristics are turned into lines of features (for the sake of readability, each actual line of features is printed on two lines on this page).

1. when - - - - - DT how WRB to TO speak VB language NN so RB when WRB ENDO OBLI - - - - - - - REFPRON - - - - - this unexpected
2. - - - - - - DT because IN you PRP visited VBD one NN em UH in IN ENDO OBLI - ED - - - - - REFPRON - - - - - this unexpected
3. like - - - - - TPRON a DT tree NN like IN very JJ highs NNS we PRP ENDO OBLI - - - - - - REFNN REFPRON - PREPINT - - - this unexpected

Figure 48: Partial view of feature sequences classified as *unexpected*

For instance, line 3 starts with the feature *like* as it is found three words before an occurrence of *this*, also printed as the second last feature of the same line. The hyphen sign after *like* means that none of the tokens added in the word list are found two words before the occurrence of *this*. When PoS tags are matched by the PERL program they are also printed. The TPRON tag shows the function of the *this* form. The words *a*, *tree*, *like*, *very*, *highs* and *we* are tokens found in the 3-gram context of *this*. The following tags correspond to the features described in Section 5.3.1. A hyphen denotes a non-existing feature for a given position before or after

the occurrence. The last element corresponds to the assigned class. When features are extracted from the native corpus subset described in 7.2.1.1, the *expected* class is printed instead of the *unexpected* one.

For the second experiment, the same program is run on the Diderot-LONGDALE subset described in 7.2.1.2 to create lines of features. However the classes are not expected and unexpected. The purpose is to see what factors lead to unexpected *this* or *that* forms so the assigned classes are *this* or *that*.

7.2.2 Classification and results

In this section, we explain the classifying method used by the classifier and how its performance is assessed. The second part deals with the results of the classification experiments.

7.2.2.1 Classification method

The machine-learning method used for the experiment applies the memory-based method, and the IB1 algorithm, as described in Section 7.1.1, is implemented in TiMBL. As a quick reminder, it can be recalled that any classifier such as an MBL system requires two types of data in order to classify sequences and verify the classification and its performance. The memory-based learner TiMBL first goes through a training phase before performing the classifying phase. In the training phase, it adds lines of features and their classes to its memory. Each line constitutes a vector of features. In the classifying phase, the classifier predicts the classes of new lines of features without the class information. The similarity between the new lines of features and all the examples in memory is computed using some distance metric. In this case, the IG metric is used as a measure to take into account the entropy of the data points and the number of values per feature, *i.e.* roughly, a degree of similarity between the different cases to be classified (see Section 7.1.2.1). The prediction is made by assigning the most frequent category within the k number of memorised lines in the training phase that are nearest to the line being

processed, hence the k-NN name for the method—NN stands for 'Nearest Neighbour'. To do so, the classifier computes a series of metrics (including Gain Ratio) in order to establish the order of the features to be taken into account in the decision process. It establishes a hierarchy of the features from most relevant to least relevant in the classifying process.

Due to the low volume of data, we use the leave-one-out option for training and testing on our dataset, which means that for each instance of the experiment, only one line of the file is used for testing and the other instances are used for training. This process is repeated for each pattern and the advantage is that, considering the small size of the samples, the leave-one-out option allows for greater robustness since this “classic” methodology maximises the training set without impairing the size of the test set. In order to evaluate the performance of the classification, precision and recall are calculated for each line due to the leave-one-out option. The results presented in 7.2.2.2 represent averages of each metric for successive classifying tests.

7.2.2.2 Results

We present the results of the experiments in two parts. First, we show the results of *unexpected* and *expected* classification and secondly the classification of only unexpected forms of *this* and *that*.

7.2.2.2.1 First experiment: expected v. unexpected TH- form classification

Running the classifier on the data yields a 0.9 global accuracy with 0.9 precision and recall. The equal number of *expected* and *unexpected* lines and an equal number of *this* and *that* occurrences in the dataset mean that random classification would provide overall accuracy of 50%. Considering this 50/50 baseline of our dataset, the extra 40% improvement margin (the actual accuracy minus the random accuracy) gives a measurement of the relevance of the feature set for the selection of expected *this* or *that*. This level is comparable to error detection tasks reported in

(Pradhan *et al.* 2010; Han, Chodorow, and Leacock 2006). The confusion matrix in Table 73 shows the distribution of classifying errors. The number of errors is identical between the two classes. This means that there are as many truly *unexpected* forms which are misclassified as there are truly *expected* forms which are misclassified.

The fact that not all lines obtain the correct class may be explained by a lack of exhaustiveness in the type of features. The immediate context of each occurrence, that is three tokens and three PoS-tags to the right and the left, may be seen as a limitation as there may be linguistic characteristics located further away that influence the selection of a class. Another limitation may find its source in the difference between the oral and written modes of the native and non-native subsets. The mode of the WSJ is written while that of the non-native subset is oral. A bias may have been introduced due to differences linked to distinct style and syntactic-complexity profiles. Classification between *expected* and *unexpected* uses determines the extent to which the features have an impact on the selection of the forms.

		Predicted classes	
		Unexpected	Expected
Actual classes	Unexpected	36	4
	Expected	4	36

Table 73: Confusion matrix for the classification of *unexpected* and *expected* forms of *this* and *that*

The mixed subset approach shows that it is possible to distinguish between *unexpected* and *expected* forms thanks to the selection of particular features that the classifier uses to categorise the abstraction of occurrences. Examples 62 and 63 show two occurrences of misclassification.

62) for - - - - - DT a DT lot NN in IN language NN but CC for IN ENDO OBLI - - - CC - - REFNN - - PREPINT - - - this unexpected expected

63) - - - - - - DT the DT problems NNS in IN at IN this DT moment NN EXO OBLI - - - - - - - - - - - - - this unexpected expected

The lines list the feature values (some are not activated) The second last feature indicates whether the form was expected or not in the context, *i.e.* whether it came from a native context or not. The last element indicates the class actually chosen by TiMBL. In both examples, the instances are extracted from non-native speech and are manually marked as *unexpected*. The classifier marks them as expected. It must be added that no index numbers were given to the occurrences of the forms, which impedes the identification of each line with the corpus text it comes from.

TiMBL allows the user to have access to the feature order set during the training phase. The Gain Ratio weight calculated for each feature shows the significance of each feature in the classification (see Table 74). Incidentally, it provides the linguist with significant information on each linguistic characteristic that has been abstracted in the feature vectors (the results are presented in descending order of the Gain Ratio values):

Feature 27	Punctuation in the 3-gram co-text of the form or not	0.3408851209
Feature 29	Pro-form or pronoun in the 3-gram co-text of the form	0.2808527832
Feature 22	Present simple verb in 3-gram co-text of the form or not	0.21598682
Feature 14	1st token to the right of the form	0.1816336533
Feature 24	Negation in 3-gram co-text of the form or not	0.1812140773
Feature 1	Specific words in 3-gram context right of the form	0.1788870398
Feature 7	PoS-function of the form	0.1713928371
Feature 30	Wh- form or relative pronoun in the 3-gram co-text of the form	0.167111053
Feature 16	2nd token to the right of the form	0.165320932
Feature 18	3rd token to the right of the form	0.1575989221
Feature 8	3rd token to the left of the form	0.1566775279
Feature 10	2nd token to the left of the form	0.1512577636
Feature 34	Number of the form	0.1509682931
Feature 17	2nd PoS tag to the right of the form	0.1445981486
Feature 15	1st PoS tag to the right of the form	0.1411174778
Feature 12	1st token to the left of the form	0.1324955
Feature 9	3rd PoS tag to the left of the form	0.1178563933
Feature 19	3rd PoS tag to the right of the form	0.1169901179
Feature 2	Specific words in 2-gram context right of the form	0.1140641243

Table 74: Gain Ratio weights of features computed by TiMBL for expected and unexpected forms

Nearly null and null GR values are not listed in the table. Only the features which show a significant level of relevance are reported. Null GR value features correspond to those that are not activated in the sequencing process. For instance, few or none of the words of the pre-established list are found in the 1- or 2-gram left context and none are found in the immediate right context of the form. Similarly, the GR value of feature 35 (see Table 72 page 333) is null due to the equal number of *this* and *that* values for the feature and thus a 0 entropy.

Globally, the values of Gain Ratio are unevenly distributed. The box plot presented in Figure 49 shows that half the values are above (median = 0.12) and a quarter of the values are within the 0.11-0.16 range. Conversely, a quarter of all the values are within the 0.01-011 range. The strip chart in Figure 50 shows the distribution of the same values by plotting every data point. The features are spread arbitrarily on the vertical axis. The data points are plotted for each feature according to their GR values (horizontal axis). This visualisation shows how features are spread according to their values. It confirms that the values are spread mostly around the 0.11 Gain Ratio threshold. Nevertheless, it also shows that there are three groups of values. The first two are located around the median and the third one includes outliers which indicate strong classifying effects.

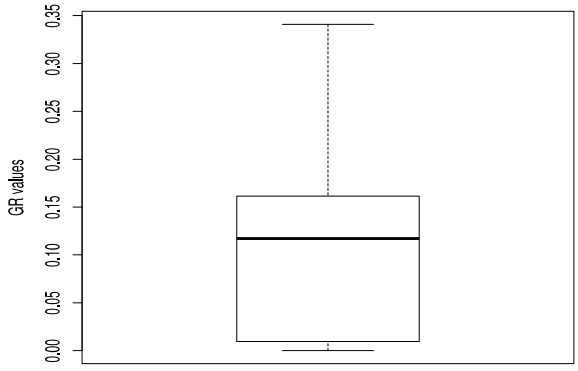

Figure 49: Box plot of GR features' distribution according to their GR values

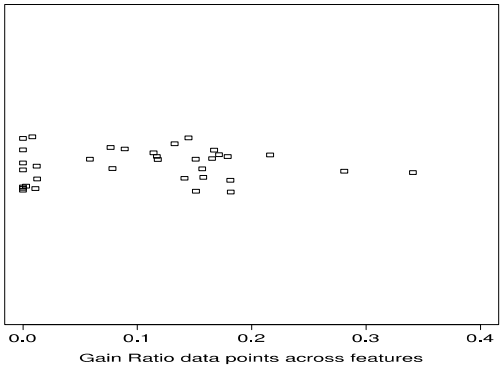

Figure 50: Strip chart of features' distribution according to their GR values

As previously said, the classifying process relies on the weights assigned to each feature by the classification process. By taking a look at the outliers, it appears that punctuation is the most important one. This value is linked to the nature of the corpora. The fact is that the oral corpus dataset does not include any full stop when the written corpus dataset does. This makes the punctuation feature a dividing criterion to split the data points across the two classes. Due to this difference between the two corpora, the feature is not relevant in terms of linguistic analysis and it may be disregarded. Two other features come into focus: pro-form or pronoun and present simple verb, all in the 3-gram contexts of the forms. These features are linguistically more significant as they can be found in both the native and non-native subsets of the dataset. The fact that *this* and *that* appear within the vicinity of a pronoun form is a strong incentive for the classification. By filtering the data points according to this criterion, 19 occurrences of the forms are extracted, 18 of which match the *unexpected* class. Similarly, by filtering the data points according to the present verb feature, 35 occurrences are retrieved, 30 of which match the *unexpected* class. These results show that these two features differentiate both subsets from each other. While they are present across the dataset, the classification results suggest that they play a significant role in helping the classifier differentiate an *expected* use of a demonstrative from an *unexpected* use. This could

be due to learner v. native differences but also to differences between the spoken and the written modes of the two subsets.

A group of other features (from feature 14 to feature 2 in descending order in Table 74) also plays a significant role in the classifying process, which reinforces the idea that such features designate differences between learner *unexpected* uses and native *expected* uses of the forms. As such, they form a knowledge base for the automated classification of learner uses of *this* and *that*.

One striking point is the irrelevance of the contextual and positional features. It may be due to the fact that there are no significant differences in use between the two native and non-native subsets leading to low entropy and thus low GR. This could indicate that French learners and native speakers have a similar way of using reference within their utterances. The low GR value for position may suggest that natives and non-natives differ slightly in their uses of the forms. Another interesting point is the irrelevance of the plural feature. It seems that the *this/these* misselection-error type, as reported in "these type of films" (Díaz-Negrillo 2007, 100), does not play a significant role in the detection in our dataset.

So overall, if this experiment mimics the cognitive selection process between expected and unexpected forms, then it seems that the important parameters of unexpected uses are: i) the present simple tense and ii) the presence of a pro-form.

7.2.2.2.2 Second experiment: unexpected form classification

In this experiment only unexpected occurrences of *this* and *that* in the Diderot-LONGDALE corpus are classified. After running the program, results show an overall accuracy of 0.875. Precision and recall values for *this* are 0.85 and 0.9 respectively. For *that,* the value are 0.89 and 0.85 respectively. The confusion matrix in Table 75 shows the details of the classification. True classes are shown horizontally while predicted classes are presented vertically.

		Predicted classes	
		this	that
Actual classes	this	18	2
	that	3	18

Table 75: Confusion matrix after the classification of unexpected *this* and *that* forms of the Diderot-LONGDALE corpus

Due to the fact that the feature set includes native- and learner-related features, the results show that predicting unexpected uses depends on native English and learner English. This may confirm that learners make use of native English features. There are 5 misclassified instances which are presetented below:

64) - - is like - - TPRON like VB </A> SYM <B> SYM is VBZ a DT good JJ EXO NOMI VBZ - - - - - - - - - - - - this that

65) - - - - - - TPRON my JJ grandma NN does VBZ by IN herself PRP French JJ ENDO OBLI VBZ - - - - - REFNN REFPRON - - PREPPOST - - this that

66) - - - - - - DT em UH I PRP love VBP language NN it PRP it PRP ENDO OBLI VBZ - - - - - - REFPRON - - - - - that this

67) - - for - - - TPRON I PRP 'm VBP doing VBG for IN him PRP </B> SYM ENDO OBLI VBZ - - - - - - REFPRON - - PREPPOST - - that this

68) there - - - - - DT and CC to TO discover VB country NN because IN there EX ENDO OBLI VBZ - - CC - - - - - - PREPPOST - - that this

Each instance represents a line of features. For instance, if we take line 65, the first six features are not activated, *i.e.* they take the "-" value. Feature 7, with the TPRON value, is activated, *i.e.* the PoS of the *this* form in this case. After a series of other feature values, the *this* in the second last position indicates the form taken from the context. The *that* in the last position indicates the class chosen by the classifier. In this case both forms differ, which indicates a misclassification. It must be noted that due to the absence of an indexing system, instances cannot be traced back to specific corpus texts.

The exploration of the features might cast light on the matter and their visualisation provides information on their distribution according to their GR values (see Figure 51).

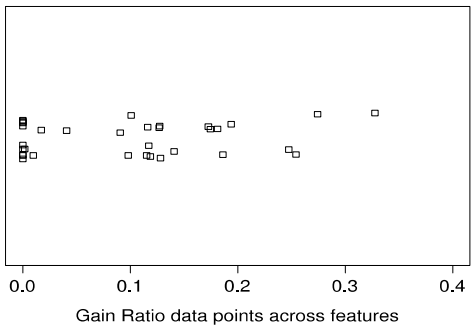

Figure 51: Feature distribution according to their GR values

The features appear to belong to four groups and range from 0 to over 0.3. Most of the values are located within the 0.1-0.2 range. Four values appear to be more significant than the others. A box plot view of the data (Figure 52) shows that half the data points are located above the 0.1 threshold. For comparitive purpose, in their study on diminutive formation in Dutch, Daelemans *et al.*(1997) classify words according to different suffix classes. To do so, they use a set of features. GR details of these features are given in (Daelemans *et al.* 2010) and show results ranging from 0.018 to 0.32.

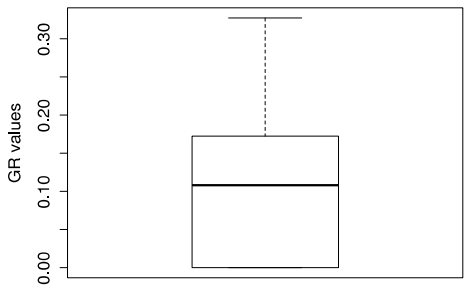

Figure 52: Box plot view of features' GR values

A detailed exploration of the features in descending order gives details on the features' GR values. Table 76 shows the list of the most significant features in descending order of GR values.

Feature 20	Context (endophoric or exophoric)	0.3275295689
Feature 22	Present simple verb in 3-gram co-text of the form or not	0.2742626684
Feature 3	Specific words in 1-gram context right of the form	0.2708145959
Feature 14	1st token to the right of the form	0.2541310006
Feature 6	Specific words in 1-gram context left of the form	0.2060031585
Feature 12	1st token to the left of the form	0.1938061295
Feature 8	3rd token to the left of the form	0.186120325
Feature 33	Presence of plural noun after the form in its determiner function	0.812140773
Feature 18	3rd token to the right of the form	0.1744689132
Feature 10	2nd token to the left of the form	0.1726666978
Feature 5	Specific words in 2-gram context left of the form	0.1605374578
Feature 4	Specific words in 3-gram context left of the form	0.1601358061
Feature 1	Specific words in 3-gram context right of the form	0.1530005162
Feature 16	2nd token to the right of the form	0.1408597915
Feature 2	Specific words in 2-gram context right of the form	0.1280180373
Feature 23	Past simple verb in 3-gram co-text of the form or not	0.127309254
Feature 17	2nd PoS tag to the right of the form	0.1268407724
Feature 21	Position (Nominative or oblique)	0.1187091008
Feature 19	3rd PoS tag to the right of the form	0.1170996989
Feature 13	1st PoS tag to the left of the form	0.1162996616
Feature 31	Introductory preposition in 1-gram to the left of the form	0.1152157945
Feature 15	1st PoS tag to the right of the form	0.1009294236

Table 76: Gain Ratio weights of features computed by TiMBL for unexpected *this* and *that* forms

The features that are not shown all return near-null or null values. The null value is due to the fact that the features are not found by the sequencing program when it scans the annotated texts. Consequently, the feature values are not activated and no value appears in the feature matrix, which in turn leads to null when computing GR.

The first four features show values which are well above those of the remainder of the list. Out of the four, two of them are also listed as significant features in the previous experiment: present simple verb in 3-gram co-text of the form and 1st

token to the right of the form. These two features help classify *unexpected* from *expected* occurrences and also help to choose wrong forms of *this* and *that*. In other terms, they are attached to unexpectedness and their activation concurs to the use of a specific erroneous *this* or *that*. This may indicate a learner error pattern. Observation of the data suggests that there is a distinction in the linguistic relevance of these two features. When filtering the data points according to the *1st token to the right* feature, it appears that this first token is the verb *be* in 17 out of 40 of the cases. However, the form is either followed by the full form or the contracted form of *be*. When it is followed by *'s,* the form is always *that.* When it is followed by *is,* it is always *this*. This distinction has a great impact on the classifying process as it is mathematically significant but it is not very relevant linguistically since the contraction can only be appended to *that*.

The present simple feature is of a different nature. Sorting the data points according to the *VBZ* value of the feature shows that most of the selected forms are *that* (19 out of 27). When sorting according to the inactivated sign of the feature, most of the selected forms are *this* (10 out of 11). This suggests that the activation of this feature triggers the choice of erroneous *that* among learners. This is consistent with theoretical findings presented in Section 2.5.2 in which it is argued that the present tense is more often linked to *this* when used by natives. Therefore, not using *this* with the present simple and using *that* with the present simple are strong factors for unexpected if not erroneous use.

The other significant feature that stands out, when classifying unexpected uses of *this* and *that*, is the contextual feature. The endophoric/exophoric distinction appears to be a strong dividing criterion for the selection process of the forms. At first sight, this is surprising as the feature does not appear as significant when it comes to selecting between *expected* and *unexpected* forms. This feature seems to be purely related to *unexpected* forms. The question is to know in what way the activation of the feature influences each form. The answer might reside in the data. Exploring the dataset and the assigned classes shows that the endophoric value of

the feature is assigned to *this* in 8 cases out of 8. Conversely, the exophoric value is assigned to *that in* 20 cases out of 32. This trend indicates that learners may have internalised a distinctive rule in the use of the two forms. For French-L1 speakers, the situational or textual referential processes seem to act as strong diacritical criteria for the choice of a form. This could be linked to teaching habits on the subject in which the endophoric/exophoric distinction acts as a strong paradigm. Cornish mentions an experiment carried out on a class of students of English as a major. He reports that many of the students use this paradigm to explain a series of utterances in which *this* and *that* are used (Cornish 1999). The evidence from the experiment presented in this section would therefore encourage teachers to also focus on other important linguistic criteria for the use of the forms.

Another one of the first four significant features is the presence of a pre-determined word in the 1-gram context right of the form. When exploring the sequenced dataset, it appears that the following words found in the 1-gram right of the form are: *very, day, is, moment, 's, for, that*. Nevertheless, *be* is predominant in this list with 17 occurrences out of 40. This makes the feature a very relevant one in the same manner as explained about the *first-token-to-the-right* feature.

The other top features that appear in the ranking correspond largely to words or tokens found in the n-gram context of the forms. This indicates that specific words may be linked to triggering the use of the forms. However, this may also be due to the size of the sample and overfitting might occur. This issue is discussed in the next section.

7.2.3 Discussion

As mentioned in the introduction paragraph of Section 7.2, these two experiments are updated versions of two similar experiments carried out previously (Gaillat, Sébillot and Ballier 2014). The initial experiments relied on the same learner subset and another randomly selected native subset. A different version of the sequencing PERL program implemented a different feature set. The contextual, positional and

referential features described in Table 59 (page 308) were not included initially. Instead, specific PoS features were used in an attempt to reflect several linguistic characteristics related to *this* and *that*. The same list of words to be searched for in the close context was also used in the initial experiment. A brief comparison between the results shows a difference of 0.2 in global accuracy (0.7 v. 0.9) between the first and second versions of experiment 1. The results remain stable with the second experiment. It seems logical to conclude that the selection of features had an impact. The PoS functional information used to inform on referential entities seems to be more relevant in the present study, especially for the identification of unexpected forms. One interesting point is that a present simple verb not only indicates an unexpected form, but it also points to the *unexpected* classification of a *that* form. Although the contextual feature is not relevant for *expected/unexpected* classification, it is very relevant in the second experiment. It shows the connection between unexpected uses and context level annotation. For finer detail, it would be interesting to see how the picture-description task plays a role (implying the use of the present simple) as opposed to the "summer" narration task (implying the use of the preterite). What we have done is monitor the role of the tenses with a past feature and a present feature.

If these experiments lack sufficient data to be completely satisfactory, at least, they give a first experimental overview of the different contributions of the different layers in the analysis:

- Positional features (parsing-based)
- Grammatical Adjacency features (PoS-based)
- Lexical-grammatical features (preceding and following words)
- Semantic features (exophoric/endophoric)

Of course, these categories could be refined, but this exploration of features has proven to be fruitful. The research agenda consists in a heuristics of features to analyse learner criterial features (Hawkins and Filipović 2012).

The current results show that there is still room for improvement. A line of research to explore could be linked to adding another layer of annotation. Discourse level annotation as explained in Section 4.2.2.5 could be added to the corpora. Tags could be related to notions such as focus, referent accessibility, and givenness of information. These extra tags could then be taken into account in the sequencing process of the features, and their relevance could be computed to see if it improves classification with TiMBL. The other line of research is about lexical contextual items, that could be systematically questioned by first running an n-gram analysis. In our experiment, *order* as a potential lexical feature has not proved to be relevant.

In the previous section, we mention the problem of overfitting the data. The fact that the datasets in the two experiments include a relatively low number of occurrences may appear as a limitation. This number is of the same order of magnitude as the number of features to describe the occurrences. In other terms, the number of training examples might be “too small to produce a representative sample of the true target function” used to classify the data (Mitchell 1997, 67). The risk is that some feature values predict just one class across all occurrences and this would be due to the small amount of occurrences. Consequently, the target function is capable of predicting with great accuracy but only due to the fact that the classifier's training is performed on skewed data. In order to verify the possibility of overfit data in this case, it would be necessary to increase the size of the dataset in order to have a larger amount of *unexpected* occurrences. This labour intensive process would provide the classifier with more occurrences. The features attached to these occurrences would likely take on a broader variety of values and their interactions with other feature values might improve the computation of a target function, giving it a more generalising power.

7.3 Summary

This chapter is grounded in machine learning with the use of TiMBL for memory-based learning. We show that learner language can be automatically analysed for two purposes. Firstly, the IB1 algorithm coupled with entropy-related relevance information helps identify specific linguistic features that play a role in the selection of a pro-form. Secondly, the MBL system can be used to detect unexpected uses of the forms.

The analysis of relevant linguistic features confirms some findings reported in the previous chapter and brings new evidence of the learner confusions that exist when selecting a pro-form. Results give details on the pro-form microsystem and the directions of "errors" made by learners. In fact, results suggest that learners make *this-it* confusions. French learners tend to replace *this* with *it* whereas Spanish learners tend to replace *it* by *this*. Regarding the *that-this* confusion, only Spanish learners tend to replace *that* with *this*. For the *that-it* confusion, French and Spanish learners tend to choose *that* instead of *it*. The pro-form microsystem theorised in Chapter 3 thus becomes more apparent.

Machine learning also helps with the more classical task of error classification. We show that distinguishing *expected* from *unexpected this* and *that* forms is possible. Features such as the presence of a present simple verb in the surroundings of a form or the presence of a referential pronoun play an important role in the classification of a form as *unexpected*. Further analysis also highlights the fact that the present simple verb feature influences learners in their unexpected uses of *that*.

All these results shed new light on the way learners use *this* and *that*. They show how PoS-functional, contextual, lexical and positional distinctions help compare corpora and which role learner-specific features have. As a result, evidence is given of systematic variations within the pro-form microsystem and learner errors can be predicted and partly explained. These results advocate for the need to approach learner corpus research with multiple, interoperable and richly-annotated corpora.

Such schemes support multi-factorial quantitative analysis of various language markers in speech.

Chapter 8 Conclusion

In this thesis, we have addressed the main question of reference in Interlanguage, *i.e.* Are there developmental patterns in the acquisition of reference in learner English? To explore this question, we limited our research to the deictic and anaphoric aspects of *this* and *that*. A preliminary study (Gaillat 2013a) showed possible difficulties for learners in their understanding of the syntagmatic, paradigmatic and pragmatic dimensions of the forms. Our research question was linked to the way learners implement deictic and anaphoric procedures, especially within the pro-form microsystem. We sought factors that influence learners in their use of *this* and *that* in referential processes and raised the question of their significance.

To answer this question, we tested three hypotheses. The first one was about the actual existence of the pro-form microsystem. The second hypothesis required the identification of factors in the use of the forms as pro-forms. The third hypothesis focused on error-patterns linked to specific factors in the use of pro-forms. Testing these hypotheses provided us with partial answers to the research question of reference in Interlanguage. As will be explained in Section 8.1 of this chapter, reference is not a simple issue for learners. It appears that referential markers such as *this* and *that* undergo a process of competition within their minds. Syntagmatic, paradigmatic and pragmatic parameters are not given the same order of preference in the minds of learners compared with native speakers. It seems that learners choose a marker based on different criteria from natives.

Our study adds to the legacy of previous research on the subject. As mentioned above, a number of findings were made and they corroborate a certain number of results from previous research (Petch-Tyson 2000; Lenko-Szymanska 2004; Zhang 2015). Petch Tyson used two corpora (ICLE[36] and LOCNESS[37]) and showed that learner confusions between both forms exist and that forms are used differently by natives and non-native speakers depending on their functional realisations. Lenko-Szymanska used two corpora (BNC and PELCRA[38]) to analyse the two forms. Her study showed differences at syntactic and pragmatic levels between natives and non natives. Zhang used one corpus of 17 argumentative essays by Chinese EFL learners. The author provided some frequency evidence regarding the pro-form microsystem. Overall, what was missing was a broad analysis of the forms encompassing the syntagmatic, paradigmatic and pragmatic dimensions across corpora of different L1s (natives and non-natives). Elaborate statistical methods also needed to be implemented not only to verify the significance of frequency observations, but to also provide an analysis of factors playing a role in the selection of the forms. This PhD completes the chain and corroborates some of the evidence in the three aforementioned dimensions. We have implemented specific statistical methods with machine learning techniques to provide an empirically grounded multi-corpus analysis which yields results in terms of NS-NNS and NS-NS differences.

In this chapter, we show how we have addressed the aforementioned research question. Hypotheses and related findings are covered in Section 8.1. In Section 8.2, we recapitulate the contributions we have made while designing our experimental framework. Because this jointly supervised PhD project tries to bridge a gap between NLP techniques and linguistic analysis, Section 8.3 critically looks back on the tools and discusses how this type of techniques can redefine the SLA/LCR research agenda. Section 8.4 points towards the future avenues of our

[36] The International Corpus of Learner English is made up of essays written by learners of English. See https://www.uclouvain.be/en-cecl-icle.html (Last access March 31, 2016)

[37] LOCNESS is a corpus of native English essays. See https://www.uclouvain.be/en-cecl-locness.html (Last access March 31, 2016)

[38] The PELCRA corpus consists of essays written by Polish students of the University of Lodi.

work both in the domains of automated annotation and Intelligent Tutoring Systems (ITS).

8.1 Initial hypotheses and findings

To bring answers to our research question, we tested three hypotheses with a number of findings in the domain of SLA and more specifically concerning acquisition patterns related to *this* and *that*. In the following paragraphs, we list the three hypotheses initially given in the introduction of this thesis and state the related findings.

i) Learners' patterns of use of the forms are dependent on the L1 due to influences from their L1.

To test this hypothesis, we carried out an approach in which the two forms were observed in their determiner and pro-form functions across several corpora. In this case, we found evidence of learner-specific patterns. Learners' syntagmatic choices of either of the forms rely on tendencies which vary according to the L1 and the forms' functional realisations. The pro-form realisation appears to be correlated with the French L1 of learners and the determiner realisation seems to be correlated with the Spanish L1. It also appears that learners tend to be aligned with natives in their use of *this* in exophoric contexts. All together, the combination of the L1 criterion with the functional realisation criterion appears to be a decisive parameter in the selection of the forms. This combination supports the idea of L1 dependence in specific functional realisations. It also speaks in favour of further exploration of the two microsystems in which *this* and *that* interact. We focused on the pro-form microsystem.

ii) A learner-specific pro-form microsystem of reference exists including specific linguistic features attached to *this* and *that* which are different from native English.

Two different analytical approaches (multiple regression and k-NN modelling) have confirmed its existence as learner specific variations can be noted. Multi-factorial methods give an insight into the significance of the features of the microsystem rather than just in terms of under- or overuse. The results suggest that learners' uses of the pro-forms significantly differ from natives' depending on the L1 and the oblique or nominative position. These two variables can be seen as two visible components of the microsystem. As they take on specific values the system's settings are modified and errors may be observed.

iii) Learner-error patterns are linked to specific linguistic features of the forms.

We adopted two approaches. Firstly, we compared the two forms with each other in a general way without taking heed of their functional distinction as determiners or pro-forms. Secondly, we focused on *this* and *that* as pro-forms and surveyed them as part of the aforementioned tripartite microsystem. In the first case, the machine learning based k-NN method shows the correlation between the presence of present simple verbs in the surrounding context and unexpected uses of either of the forms. This may suggest that it is mostly morphosyntactic features which impact the selection of forms. In the second approach, the two forms are compared with *it*. Results show that variations appear to be correlated with the position and the learners' L1. When analysed next to *this* and *that*, *it* clearly appears to be a competitor in the learners' minds. Errors seem to be correlated with the oblique value of the position variable of the system. Even if, globally, learners and natives all appear to prefer *it*, it seems that when confusions are made, they appear mostly in the oblique case. The L1 also introduces a distinction between learners as they show distinctive preferences. Figure 53 shows the confusions within the microsystem. For example, French learners tend to replace *it* with *that*.

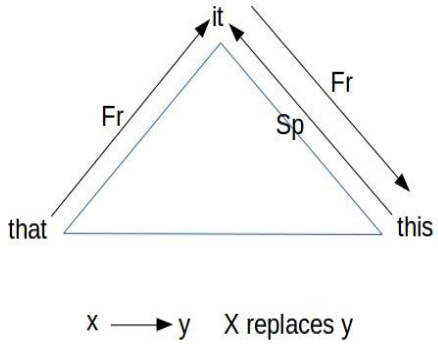

Figure 53: Learner confusions in the pro-form microsystem

The horizon of this non-linguistic investigation (French v. Spanish as L1) is a multidimensional representation of the parameters. Within this microsystem approach (learners make paradigmatic choices within a family of markers), this representation is focused on the L1 variable (note, in future research, logistic regressions could be visualised in a similar way).

The striking point appears to be the fact that confusions always involve either *this* or *that* with *it*. In fact, confusions between *this* AND *that* do not appear to be significant within the microsystem. Interestingly, the exophoric or endophoric distinction is not a determining factor in this microsystem. These main findings have implications in terms of teaching. Indeed, EFL teaching methods focus mostly on the difference between *this* and *that* but they completely overlook the fact that learners get confused between *it* and TH- forms. This learner behaviour must be recognised in order to be addressed.

All in all, our research has helped uncover learner-specific patterns in the acquisition of the forms. Our corpus-driven approach combined with multifactorial methods for the analysis of forms shows that comparing learner corpora with native corpora lays the ground for new evidence on the paradigmatic difficulties that learners have with *it* as a competitor form. In a broader perspective regarding reference, these results may indicate that learners implement reference differently

from natives. We have shown that learner difficulties are primarily linked to functional realisations in combination with specific syntactic features. This suggests that they mostly rely on the syntagmatic dimension of utterances. Conversely, the literature shows that natives tend to rely on pragmatic criteria to implement reference. This difference shows that learners have difficulties interpreting the pragmatic values of referential forms and that the articulation between the notions of deixis and anaphora is not completely grasped. Petch-Tyson (2000) and Lenko-Szymanska (2004) have shown that when pragmatic factors are used by learners to justify their choices, they consistently show deviations compared with natives. These are either due to over-reliance on the distal value of the forms or to the misuse of non-informative noun heads after a determiner—NS choose informative noun heads in this case. Because referring is a complex multidimensional process, learners seem to favour visible criteria of the co-text to make their choices.

8.2 Contributions

To test our hypotheses, we had to design a framework empirically grounded in the use of corpora, which relied on a theoretical model of reference. Our main contribution was to make the reference model explicit and to render three corpora interoperable in terms of data comparability. It is now possible to extract identical types of data from any of the three corpora so that results can be compared either in context or within matrices which support automatic analysis with machine learning technologies. The framework relies on three linguistic, SLA and technical pillars.

The first pillar of the framework is linguistic. A list of specific linguistic characteristics related to the use of *this* and *that* had to be established and the purpose was to prepare the ground for the annotation of the corpora. Our first contribution is a direct result of this work as it gives a clear understanding of deixis and anaphora. In this thesis, we provide a model on reference showing how referential processes are at work within the general framework of reference. We

bring together two seemingly antagonistic views on anaphora and deixis by integrating the situational paradigm into a discourse-functional approach. We explain how a discourse entity may be referred to and we show how retrieval processes such as endophora and exophora operate at context level to integrate the entity into discourse according to its 'given' or 'new' status. We also introduce the role of *this* and *that* in a synthetic view of referential processes. We demonstrate their ability to make a referent accessible and to be governed by the predicational context to produce various meaning effects. All in all, we have shown that the reference system is a dynamic model in which retrieval processes operate via several components, such as the levels of speech and elements of the predicational context. This model has served two purposes. It has helped interpret learner utterances including errors and identify some of the features which need to be introduced in the data format of the corpora.

The second pillar of the framework is grounded in SLA with a focus on the learners' acquisition processes of the forms. By using the model of the reference system, we analysed learner uses of the forms and identified common traits of unexpected uses. Our contribution is to provide a typology of uses in which two microsystems are revealed depending on the functional distinction of *this* and *that*. These two microsystems reveal specific learner difficulties based on paradigmatic confusions with two other competitor forms, *i.e. it* and *the*. Thanks to this typology, we identified more features which needed to be taken into account for a multi-corpus study of the forms and we focused our research question on the learner-specific pro-form microsystem.

The third pillar of the framework is technical and a number of contributions can be reported. In fact, analysing learner language meant the use of several corpora and the need to enrich, extract, manipulate and compare the forms in context. Consequently, we had to provide solutions in terms of annotation and in terms of corpus interoperability to make rich data comparable.

In terms of annotation, our contribution is a multi-layer annotation scheme which is now applicable to several corpora. This annotation scheme includes a fine-grained PoS annotation which was extended to include a functional distinction for the forms. The use of NLP tools has made this possible. The Penn Treebank tagset was modified to train a PoS tagger and tag two learner corpora (NOCE and Diderot-LONGDALE). A number of functional distinctions were introduced for each form as well as for *it* whose non-referential function was also distinguished by the tagger. As a result, any corpus can now be PoS-tagged automatically with special labels for *this, that* and *it*. The training file used for this process is now available online. The accuracy of the tagging processes was measured for native and non-native corpora showing a decrease in quality for non-native corpora. Manual checking was operated to ensure the best standards for subsequent analysis. It is worth mentioning that while modifying the tagset, a number of tagging errors were discovered in the WSJ gold standard regarding the PoS tagging of the two forms. We developed specific queries to identify and correct these errors automatically. Contextual annotation was also required for all forms of *this* and *that*. No NLP tools were found or developed to proceed with the task automatically. Nevertheless, three fairly sized subsets of the WSJ, NOCE and Diderot-LONGDALE corpora have been annotated manually. Positional annotation was also a requirement. For this purpose a specific PERL program, now available online[39], was developed. It implements a rule-based algorithm that automatically detects the forms and their grammatical categories so as to assign a specific positional tag. As a result of these mostly automatic processes, the subsets of three corpora are now annotated with several independent but inter-related annotation layers as recommended by Wynne (2005).

In terms of corpus interoperability, one requirement was the possibility of querying the corpora in order to retrieve data according to linguistic constraints. This meant that, after introducing annotation, the data had to be formatted in such a manner that a query tool could be used to read the data. This was achieved with NITE NXT

[39] See http://gaillat.free.fr/research/ (Last access March 31, 2016)

Search. As a contribution, we have provided a version of the corpora which is compliant with NITE XML. The data are placed into the NITE-model structure that describes different types of data elements: PoS-function, position and context. Complex queries can rely on several of these elements at the same time and they can be applied to any of the corpora to retrieve the forms and their close context. Another expected outcome in the domain of corpus interoperability was the possibility of retrieving the data in formats that would make comparability possible with the use of machine learning tools. As far as we know, the text search tools that exist on the market all focus on retrieving the data in their contexts of occurrence without any extra domain knowledge. Concordancers display all the instances of a word or phrase and they mostly focus on word frequencies. Although they provide statistics based on these words, what was needed in our work was to retrieve all occurrences and feature elements extracted from the co-text or any annotation layer. All these data, including domain knowledge, had to be displayed in such a way that they could be imported into machine learning tools. Our contribution was to develop a family of in-house PERL programs that sequence the data so as to provide instances of feature values. The programs plough through the multiple annotation layers of corpora to retrieve and display the data elements in identical structures or matrices. These structures support data comparability when imported in machine learning tools. All in all, our technical contribution has consisted in making the data that compose the corpora comparable. By adopting a common annotation scheme, by applying this annotation scheme in several layers, by formatting the data of each corpus with the same structures (XML or feature sequences), the corpora can interact together in the sense that they can be combined synchronously or successively to provide comparable results which in turn can be analysed. In other words, the corpora have been made interoperable.

8.3 Epistemological implications

The findings reported in Section 8.1 are the results of the analyses carried out with machine learning tools relying on exemplar-based approaches, *i.e.* instances of

features belonging to different classes. The machine learning approach has long been in use in the domain of corpus linguistics but it introduces a new challenge in the the domain of SLA/Learner Corpus Research. It implies a change of paradigm insofar as corpus texts are not only observed 'as is' any more (as with concordancers for instance). Instead, they are converted into abstractions of features which are assigned values and passed on into modelling equations. The purpose of machine learning tools is to compute models which are designed to account for the variability of the feature values. A modelling approach leads the researcher to regard and define linguistic issues as systems which are governed by series of criterion/value pairs. In doing so, machine learning technologies help model linguistic systems to explore multi-factorial discourse environments. As a metaphor, we could say that machine learning tools provide the researcher with a kind of microscope which helps observe mass data and highlight significant features—which, incidentally, may be at the expense of misclassified examples and outliers. Fine-tuning the microscope is of crucial importance to make the right observations and the same applies to machine learning tools like TiMBL. The choice of the methods to model linguistic realities, including in the domain of SLA, has an impact on the results the tools provide. It is therefore paramount to understand the implications of different methods. Consequently, grounding the approach in the use of such tools raises an epistemological question concerning their reliability in terms of the decision-making process. It is important to understand the settings of these tools especially the underlying algorithms that govern the learning processes. One question raised by Mitchell (1997, 15) can retain our attention: “Which algorithms perform best for which types of problems and representations?” In the field of corpus linguistics, this question can guide us when we try to understand how to deal with linguistic issues and their associated properties in the most optimal manner. Some aspects of the answer can be found by looking at the different families of algorithms.

A taxonomy of algorithms used in machine learning (Flach 2012, 38) gives an overview of the different branches that split the family. For Flach, machine learning

models can be categorised according to two groups of characteristics. The nature of the first group relies on mathematical reasoning. Models can be geometric, probabilistic or logical. Geometric models, such as linear models, rely on lines, planes and distances to split the set of all possible instances (sequences of features) describing the data (Flach 2012, 21). The target function—the function used by the program to compute the data points—acts as a dividing border between classes of instances. Probabilistic models rely on finding out the probability distribution of the data. In other terms, feature values that compose the instance space are distributed according to some underlying random process and the probabilistic approach consists in finding out the probability of a class attached to an instance. Logical models focus on learning sets of rules based on the *if-then* principle. The sum of all the rules is expected to match the underlying function that generates all the instances present in the training data. In the second group, model characteristics are linked to the methods used to handle the learned instances. For Flach, there are two strategies to apprehend the set of instances to be learned. They can either be grouped or graded. 'Grouping' refers to classifying instances in distinct categories whilst 'grading' corresponds to representing and exploiting every difference between instances, hence the possibility of assigning a unique score to a specific instance of the space. It must be noted that there can be hybrid methods. Figure 54 shows how several types of methods may be classified according to the aforementioned characteristics. It shows “a taxonomy describing machine learning methods in terms of the extent to which they are grading or grouping models, logical, geometric or a combination, and supervised or unsupervised. The colours indicate the type of model, from left to right: logical (red), probabilistic (orange) and geometric (purple)” (Flach 2012, 38).

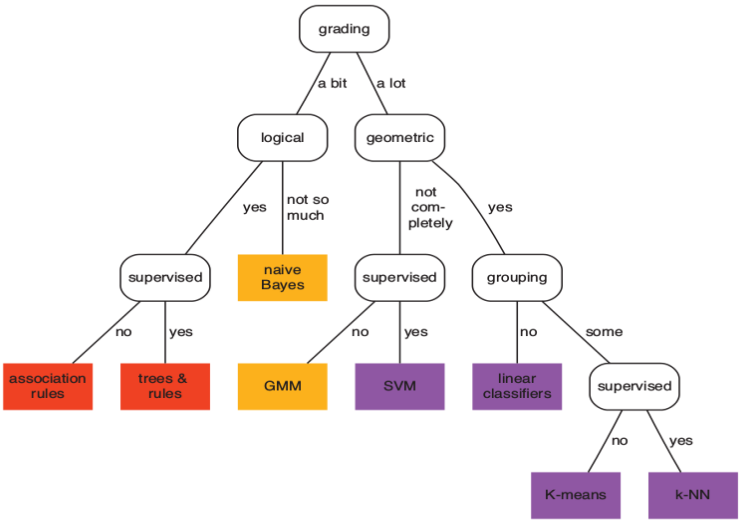

Figure 54: "A taxonomy describing machine learning methods in terms of the extent to which they are grading or grouping models, logical, geometric or a combination, and supervised or unsupervised." (Flach 2012, 38).

This PhD has exemplified two methods: regression and k-NN. For instance, the k-NN model (we implement the k-Nearest Neighbour model in Chapter 7) develops several characteristics. At a global level, it is a grading model, as it uses very accurate feature differences between instances. However, because it assigns classes to instances on the basis of a k number of nearest neighbours, it also complies with the grouping characteristic. As all data points are assumed to be part of a multi-dimensional space symbolised by feature vectors, the model is geometric. The fact that classes are assigned to instances in the training data places the model in the supervised learning category. In sum, there is a variety of models that rely on different characteristics and choosing a model implies matching the characteristics with those of the task which needs to be undertaken.

In the case of SLA/LCR, choosing a model and its settings may vary according to the characteristics of the task. Detecting errors in texts, selecting forms for specific contexts, classifying texts according to their topics or the learners' proficiency levels, identifying coreferential relationships are tasks which need to be evaluated according to the implied linguistic properties. These properties are the parameters of the linguistic microsystems in which the tasks are grounded and answering the

question of the choice of a model partly revolves around matching the properties of the linguistic system with the features chosen when designing the model.

It ensues that several conditions must be taken into account when selecting a method, *i.e.* modelling. If the linguistic task involves predicting specific categories, this will have an impact on the supervised or unsupervised machine learning methods. In supervised learning, training instances need to be mapped with category labels so that the performance module can produce predictions. Conversely, in unsupervised models, predictions are carried out intuitively as the model discovers the categories or scores to assign. If the task involves taking into account every small difference between instances, strong grading models are more appropriate. We can think of the difference between classifying a learner use of a word as an error on the basis of *if-then* rules rather than on the basis of every difference between the features that characterise the use of the word. In the first case, the rules divide the possible instances between errors and correct uses. In the second case, the feature differences are used to compute unique values for the instances and these values, *i.e.* probabilities, can be used as evidence of error or not.

If the task involves a large number of features, some models are more appropriate than others. In our two analytical approaches, we have seen that the k-NN model allows more features than the regression model as, typically with this kind of model, problems occur when a variable has four possible features or more. The number of features might also lead to interference issues, *i.e.* noise due to irrelevant features. Mitchell points out a disadvantage to many instance-based approaches, including nearest-neighbour approaches, insofar as "they typically consider *all* attributes of the instances when attempting to retrieve similar training examples from memory". For the author, "if the target concept (the model) depends on only a few of the many available attributes, then the instances that are truly most 'similar' may well be a large distance apart." (Mitchell 1997, 231). It seems that adding irrelevant features might have a counterproductive effect by

creating noise and “drowning” feature similarities within irrelevant features. If the task involves different types of features, their nature impacts the outcome of the predictions of the model. For instance, Arndt-Lappe (2011, 587) reports experiments in which constituent features have a positive impact on the performance classification of stress in noun-noun compounds with the k-NN algorithm, whereas syntactic and semantic features do not. All these conditions show that the feature abstraction, which encapsulates a linguistic issue and its properties, needs to be adequately adjusted to the type of model employed. More empirical research is required to understand the role of feature representations in models. A challenging research programme could address i) the implementation of proper ML features to account for linguistic properties (here, applied to SLA issues) and ii) the following question: How do algorithms contribute to the understanding of (linguistic) categories?

As well as feature representations, the methods implemented in machine learning also need to be tested. The way modelling is carried out may have an impact on the results provided by the tools and it is important to understand how the data points are manipulated in the process. As mentioned before, in this PhD, we present two methods to model our data. Tono *et al.* (2013) makes use of two other methods which are: Correspondence Analysis and Variability-based Neighbouring Clustering (VNC). These two methods are used to associate specific features of error types such as 'omissions' and 'additions' with different age brackets of learners of English. The purpose of the study is to identify groups of errors in relation to age. For instance, with the Correspondence Analysis, they observe that addition errors made by learners happen most often on nouns among children aged 7 or 8. At a later stage, learners do not repeat this type of errors. With VNC analysis, the researchers manage to refine the analysis as it helps them find information on how data points are grouped together meaningfully. For instance, this method links two development stages in the 'addition' error type with specific PoS used by learners. To sum up, machine learning methods provide different levels of interpretations of the data as they show different kinds of links between the data points. Testing

these methods is of crucial importance to better understand those which are most appropriate to the types of tasks they are used for, *e.g.* learner error detection, linguistic feature significance in learner specific microsystems. As pointed out by Schneider (2015), the avenues opened in (Díaz-Negrillo, Ballier, and Thompson 2013) on the role of automatic tools to process learner corpora call for further work. Hypotheses, approaches and algorithms need to be tested "in better controlled environments, using more computationally intense methods and interactive, semi-automatic approaches" (Schneider 2015).

8.4 Limitations and future developments

In this section, we cover some of the possible developments that stem from the work presented in this thesis. In 8.4.1, we present how the data processing chain can be improved to construct corpus abstractions. In 8.4.2, we show what kind of future applications could rely on an automated linguistic analysis.

8.4.1 In linguistic data analysis

The first issue that would require immediate attention is that of discourse-level annotation. In Chapter 2, we endorsed a discourse-functional approach in the analysis of referential processes in which the givenness of referents is an important point for their interpretation. In Chapter 3, we identified the importance of adding this kind of discourse annotation in the structuring process of corpora. It is clear that the endophoric/exophoric distinction and the new/known distinction should be added in two different layers. In spite of the existence of a tool (CESAX) that automates the task of adding givenness information to text units, this solution was ruled out. Nevertheless, it is necessary to overcome technical limits to insert discourse annotation. It would supply data sequencing processes with more potential features. This would open the door to more detailed statistical experiments whereby the new features would be weighted in relation to others. The importance and significance of the givenness feature could be assessed.

The second issue is a direct consequence of the first one. Reference requires a comprehensive approach in terms of features. We conducted analyses that left the givenness feature out, which at best ignores the feature but, at worst, ignores the impact of the feature on other features. In other terms, by entering this feature as a variable in our analytical methods, it might appear to be significant when a currently significant feature might become insignificant. This has implications on the observations made on referential patterns related to the two forms.

A third issue relates to the types of corpora that we used in our experiments. The WSJ and the NOCE corpora were of the written mode whilst the Diderot-LONGDALE is of a spoken mode. These modes are distinct in terms of speech characteristics such as repetitions, backchannels, onomatopoeia, interjections or punctuation. These distinctions are not without consequences on the observations extracted from the corpora and the way texts are processed by machine learning tools. Consequently, our analytical methods might exclude variable values which are significant in terms of explanation. Spoken or written features might be related to specific uses of the forms but this is not detectable in the current form of our analytical framework. Further efforts should be made to amend this problem. By adding more spoken and written corpora in the database, the mode could be taken into account as a variable, and so evaluated with regard to other variables. This would give clear answers to the influence of this factor on the use of the forms. Corpus types also depend on the tasks which are required from subjects. Our analysis should be refined in relation to these tasks, which has not been investigated as a variable in this PhD. Because of the LONGDALE protocol (Meunier *et al.* 2008), it was not possible to film learners. Learner multimodal data may probably offer groundbreaking cues for the analysis of tasks in which deixis might be more central (*e.g.* for picture description or map task). The analysis of gestures would be instrumental in the analysis of deictics. Ostensional features are probably underestimated in our analysis. It should be noted, however, that our solution adopted with NXT is compatible with multimodal corpora. Similarly, social information in the form of meta-data concerning subjects could also have been

taken into account (see Annex O for the questions asked to students) as variables in the system. The aforementioned developments in the analysis of linguistic data could be used to supply information to a module dedicated to error detection in an Intelligent Tutoring System (ITS).

8.4.2 Future applications

An ITS is a system dedicated to mimicking the behaviour of a human tutor and to interacting with a learner by adapting the situation to the user, *i.e.* to provide tutoring services to support learning (Nkambou *et al.* 2010; Danna 1997, 17). The system adapts its behaviour to the learner's by taking into account a number of parameters used for information collection. The data are subsequently used to compute interactions and contents dynamically supplied to the user. The case of *this* and *that* could be used to feed an ITS to guide learners in their acquisition of the correct uses of the forms.

An ITS is traditionally composed of several models:

- The expert model
- The learner's model
- The tutoring model
- The user interface

The purpose of each model is clearly defined. The expert model is designed to give the ITS an expert's knowledge in the domain. It is expected to behave like an expert by, for instance, providing the solution to a problem. The learner's model is composed of all possible information about the learner. This includes his/her state of knowledge and interest on the topic, his/her pedagogical preferences regarding the way the topic should be taught and any other type of information (such as age) that can help adjust the interactions of the system with the learner. The tutoring model is in charge of managing the interactions with the learner by receiving input from the expert and learner models. The order of activities, their levels of difficulty

and any kind of feedback is decided in this model so that information is dynamically provided to the learner at the most relevant time in his/her learning process. For such decisions, the tutoring system relies on pedagogical strategies which have been formalised during the programming stage. The user interface takes charge of the communication between the system and the learner.

In the case of *this* and *that*, the work on the detection of relevant referential features and on the automatic detection of unexpected uses could be used in the the tutoring and expert models of an ITS. The choice of the model would depend, respectively, on whether the integration of this work were to be made at decision level regarding tutoring strategies such as error annotation and detection, or at knowledge level with concepts and rules—such as learner errors and native typical uses—learned by the expert model.

The expert model appears as an avenue for further exploration due to the fact that it holds the knowledge of the domain. For Nkambou *et al.* an ITS must "possess a domain-specific expert module that is able to generate and resolve domain problems and provide access to such knowledge in order to facilitate dissemination" (Nkambou 2010, 16). The elaboration of an expert module could be achieved by following an approach focused on improving domain knowledge. It could be captured by using data mining techniques as exemplified by Fournier-Viger *et al*. They report on the acquisition of an ill-defined task, *i.e.* the articulation of a robot arm of the International Space Station. An ill-structured task is a problem characterised by its complexity "with indefinite starting points, multiple and arguable solutions, or unclear strategies for finding solutions" (Fournier-Viger *et al.* 2011, 478). In the ITS they develop, they implement an approach grounded in data mining. The underlying principle is to use algorithms to automatically extract a 'partial task model' from existing user solutions. For instance, the system records the values of a specific number of events such as the decreasing of the rotation value of a joint by two units or increasing the zoom of a camera by two units. So

this model is made of sequences of actions that are elicited according to different dimensions, *i.e.* feature-value pairs that characterise an action.

Fournier-Viger *et al.*'s approach is very similar to our error-detection approach in which a corpus of native English is used to learn correct uses of *this* and *that*. Just like the authors' system classifies actions according to various dimensions, applying machine learning on a native corpus helps extract a 'partial task model' from user solutions, *i.e.* correct uses of natives. This task model is used to make decisions and can be made accessible to help learners discover the features that most impact their decision-making process. Such a principle is exemplified by Dickinson *et al.* (2008, 370) who report on the design of an Intelligent Computer Assisted Language Learning system (ICALL)—a form of Intelligent Language Tutoring System—in which particle usage feedback is provided in real time to learners involved in an online chat. This approach shows that by feeding the expert module with specific linguistic knowledge, the challenge of real time learner tutoring is possible.

In sum, an ICALL system could be developed with a specialisation on referential processes. Models based on SLA linguistic microsystems such as pro-forms could be introduced as independent blocks in the expert module of a tutoring system dedicated to language writing. Learners' written input could be assessed and corrected with feedback. Teachers could have access to a database of student writings classified according to errors and linguistic issues. The system would provide indications on areas of reference which would need reinforcement.

With the combination of machine learning technologies and automatic learner corpus analysis (Antoniadis *et al.* 2009), language learning tutoring systems can evaluate input immediately, which is going to provide new diagnostic tools to teachers and learners. Language learning sessions will increasingly rely on tripartite strategies in which the teacher, the learner and the machines fed by corpora, will interact with each other to guide and accompany speech performance.

Bibliography

1) Aarts, Bas, Gerald Nelson & Sean Wallis. 2007. Using Fuzzy Tree Fragments to Explore English Grammar. *English Today* 23(2): 27–31.

2) Aarts, Jan, Hans Van Halteren & Nelleke Oostdijk. 1998. The Linguistic Annotation of Corpora: The TOSCA Analysis System. *International Journal of Corpus Linguistics* 3(2): 189–210.

3) Aha, David W., Dennis Kibler & Marc K. Albert. 1991. Instance-based Learning Algorithms. *Machine Learning* 6(1): 37–66.

4) Antoniadis, Georges, Sylviane Granger, Olivier Kraif, Claude Ponton, Julia Medori& Virginie Zampa. 2009. Integrated Digital Language Learning. In Nicolas Balacheff et al. (eds.). *Technology-Enhanced Learning*: 89–103. Dordrecht: Springer

5) Ariel, Mira. 1994. Anaphoric Expressions: A Cognitive Versus a Pragmatic Approach. *Journal of Linguistics* 30(1): 3–42.

6) Arndt-Lappe, Sabine. 2011. Towards an Exemplar-based Model of Stress in English Noun–Noun Compounds. *Journal of Linguistics* 47(03): 549–585.

7) Artstein, Ron & Massimo Poesio. 2008. Inter-coder Agreement for Computational Linguistics. *Computational Linguistics* 34(4): 555–596.

8) Ballier, Nicolas & Philippe Martin. 2013. Developing Corpus Interoperability for Phonetic Investigation of Learner Corpora. In Ana Díaz Negrillo, Nicolas Ballier & Paul Thompson (eds.), *Automatic Treatment and Analysis of Learner Corpus Data*: 33–64. (Studies in Corpus Linguistics 59). Amsterdam: John Benjamins Publishing Co.

9) Barlow, Michael. 2005. Computer-based Analyses of Learner Language. In Rod Ellis & Gary Barkhuizen (eds.), *Analysing Learner Language*: 335–357. Oxford: Oxford University Press.

10) Biber, Douglas. 1993. Representativeness in Corpus Design. *Literary and Linguistic Computing* 8(4): 243–257.

11) Biber, Douglas, Stig Johanson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Longman.

12) Bies, Ann, Mark Ferguson, Karen Katz, Robert MacIntyre, Victoria Tredinnick, Grace Kim, Mary Ann Marcinkiewicz & Britta Schasberger. 1995. *Bracketing Guidelines for Treebank II Style Penn Treebank Project*. Philadelphia: CIS, University of Pennsylvania.

13) Bordet, Geneviève. 2011. « This » comme marqueur privilégié du genre : le cas des résumés de thèses. *Discours* (9). http://discours.revues.org/8506. (27 March, 2016.)

14) Boucher, Paul, Frédéric Danna & Pascale Sébillot. 1993. *Compounds: an Intelligent Tutoring System for Learning to Use Compounds in English*. Rapport de recherche. INRIA.

15) Bouscaren, Janine. 1991. *Linguistique anglaise : initiation à une grammaire de l'énonciation.* Paris: Ophrys.

16) Boyd, Adriane, Whitney Gegg-Harrison & Donna Byron. 2005. Identifying Non-referential *it*: a Machine Learning Approach Incorporating Linguistically Motivated Patterns. *Proceedings of the Second ACL Workshop on Effective Tools and Methodologies for Teaching NLP and CL*: 40-47. University of Michigan - Ann Arbor, Michigan, USA: Association for Computational Linguistics.

17) Boyd, Adriane, Markus Dickinson & W. Detmar Meurers. 2008. On Detecting Errors in Dependency Treebanks. *Research on Language and Computation* 6(2): 113–137.

18) Bresnan, Joan, Anna Cueni, Tatiana Nikitina & R. Harald Baayen. 2007. Predicting the Dative Alternation. In Bouma Gerlof, Irene Kramer & Joost Swarts (eds.), *Cognitive Foundations of Interpretation*: 69–94. Amsterdam: Royal Netherlands Academy of Arts and Sciences.

19) Bresnan, Joan, and Tatiana Nikitina. 2009. On the Gradience of the Dative Alternation. In Linda Uyechi& Lian-Hee Wee (eds.), *Reality Exploration and Discovery: Pattern Interaction in Language & Life*: 161–84. Stanford CA: Center for the Study of Language and Information.

20) Brill, Eric. 1992. A Simple Rule-based Part of Speech Tagger. *Proceedings of the Third Conference on Applied Natural Language Processing (ANLC '92)*: 152–155. Stroudsburg, PA: Association for Computational Linguistics.

21) Brutt-Griffler, Janina, and Keiko K. Samimy. 2001. Transcending the Nativeness Paradigm. *World Englishes* 20(1): 99–106.

22) Burnard, Lou. 2005. Metadata for Corpus Work. In Martin Wynne (ed.), *Developing Linguistic Corpora: A Guide to Good Practice*: 30-46. Oxford: Oxbow. https://ota.ox.ac.uk/documents/creating/dlc/chapter3.htm. (27 March, 2016.)

23) —. 2007. *The British National Corpus, version 3 (BNC XML Edition)*. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. http://www.natcorp.ox.ac.uk/. (27 March, 2016.)

24) Buscail, Laurie. 2013. *Étude comparative des pronoms démonstratifs neutres anglais et français à l'oral : référence indexicale, structure du discours et formalisation en grammaire notionnelle dépendancielle*. Thèse de doctorat. Toulouse: Université de Toulouse 2.

25) Calhoun, Sasha, Jean Carletta, Jason M. Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman & David Beaver. 2010. The NXT-format Switchboard Corpus: a Rich Resource for Investigating the Syntax, Semantics, Pragmatics and Prosody of Dialogue. *Language Resources and Evaluation* 44(4): 387–419.

26) Carletta, Jean. 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics* 22(2): 249–54.

27) Carletta, Jean, Jonathan Kilgour, Tim O'Donnell, Stefan Evert, & Holger Voormann. 2003. The NITE Object Model Library for Handling Structured Linguistic Annotation on Multimodal Data Sets. *Proceedings of the EACL Workshop on Language Technology and the Semantic Web*. Budapest: Association for Computational Linguistics. http://homepages.inf.ed.ac.uk/jeanc/nlpxml2003.final.pdf. (27 March, 206.)

28) Carletta, Jean, Stefan Evert, Ulrich Heid & Jonathan Kilgour. 2006. The NITE XML Toolkit: Data Model and Query Language. *Language Resources and Evaluation* 39(4):313–334.

29) Chanquoy, Lucile, Andre Tricot, & John Sweller. 2007. *La Charge cognitive: théorie et applications*. (Collection U). Paris: A. Colin.

30) Charniak, Eugene, Don Blaheta, Niyu Ge, Keith Hall, John Hale & Mark Johnson. 2000. BLLIP 1987-89 WSJ Corpus Release 1. Philadelphia: Linguistic Data Consortium.

31) Chomsky, Noam. 1957. *Syntactic Structures*. (Janua linguarum 4). The Hague: Mouton.

32) —. 1966. *Aspects of the Theory of Syntax*. Cambridge (Mass.), USA: Massachusetts Institute of Technology Press.

33) —. 2000. *New Horizons in the Study of Language and Mind*. Cambridge: Cambridge University Press.

34) Cohen, Jacob. 2009. *Statistical Power Analysis for the Behavioral Sciences*. New York: Routledge.

35) Corder, Stephen Pit. 1967. The Significance of Learner's Errors. *International Review of Applied Linguistics in Language Teaching* 5(4): 161–169.

36) Cornillon, Pierre-André, Arnaud Guyader, François Husson, Nicolas Jégou, Julie Josse, Maela Kloareg, Éric Matzner-Lober & Laurent Rouvière. 2010. *Statistiques avec R*. 2e éd. aug. Rennes: Presses Universitaires de Rennes.

37) Cornish, Francis. 1999. *Anaphora, Discourse, and Understanding. Evidence from English and French*. Oxford: Oxford University Press.

38) —. 2010. Anaphora: Text-based or Discourse-dependent? Functionalist vs. Formalist accounts. *Functions of Language* 17(2): 207–241.

39) —. 2011. Indexical Reference within Discourse Context: Anaphora, Deixis, "Anadeixis" and Ellipsis. Paper presented at the Journée d'étude "Ellipse et anaphore", Institut Charles V, Université Paris-Diderot. October 15, 2011.

40) Cortes, Viviana. 2007. Issues in the Design and Analysis of Learner Language Corpora. In Esther Usó-Juan & Maria Noelia Ruiz-Madrid (eds.), *Pedagogical Reflections on Learning Languages in Instructed Settings*: 94–111. Newcastle: Cambridge Scholars Publishing.

41) Cotte, Pierre. 1993. De l'étymologie à l'énonciation. *Travaux de Linguistique et de Philologie*, XXXI : 43–89.

42) —. 1997. Étude d'un texte de D. H. Lawrence. *Grammaire linguistique* :153–158. (CNED-Didier Concours). Paris: Didier érudition CNED.

43) Daelemans, Walter & Antal van den Bosch. 1992. Generalization Performance of Backpropagation Learning on a Syllabification Task. *Proceedings of the 3rd Twente Workshop on Language Technology*: 27–38. Enschede: Universiteit Twente.

44) Daelemans, Walter, Peter Berck & Steven Gillis. 1997. Data Mining as a Method for Linguistic Analysis: Dutch Diminutives. *Folia Linguistica* XXXI(1-2): 57–75.

45) Daelemans, Walter, Jakub Zavrel†, Ko van der Sloot & Antal van den Bosch. 2010. *TiMBL: Tilburg Memory-based Learner Version 6.3 Reference Guide*. ILK 10-01. Tilburg: Induction of Linguistic Knowledge, Tilburg University and CLiPS, University of Antwerp.

46) Dagneaux, Estelle, Sharon Denness & Sylviane Granger. 1998. Computer-aided Error Analysis. *System* 26(2): 163–174.

47) Danna, Frédéric. 1997. *Modélisation de l'apprenant dans un logiciel d'enseignement intelligemment assisté par ordinateur – application à un tutoriel intelligent dédié aux composés anglais*. Thèse de doctorat. Rennes: Université de Rennes 1.

48) Danon-Boileau, Laurent. 1984. That is the Question. In Almuth Grésillon & Jean-Louis Lebrave (eds.), *La Langue au ras du texte* : 31–55. Lille: Presses Universitaires de Lille.

49) —. 1992. Ce que "ça" veut dire: les enseignements de l'observation clinique. In Laurent Danon-Boileau & Mary-Annick Morel (eds.), *La Deixis : Colloque en Sorbonne (8-9 juin 1990)* : 415–425. Paris: Presses Universitaires de France.

50) Dasgupta, Rana. 2005. *Tokyo Cancelled*. New York: Black Cat.

51) Davies, Marl. 2009. The 385+ Million Word Corpus of Contemporary American English (1990–2008+): Design, Architecture, and Linguistic Insights. *International Journal of Corpus Linguistics* 14(2): 159–190.

52) Díaz-Negrillo, Ana & Jésus Fernandez-Domingez. 2006. Error Tagging Systems for Learner Corpora. *Spanish Journal of Applied Linguistics* (RESLA) (19): 83–102.

53) Díaz-Negrillo, Ana. 2007. *A Fine-grained Error Tagger for Learner Corpora*. Ph.D. Thesis. Jaen: University of Jaen.

54) —. 2009. *EARS: a User's Manual*. Vol. 1–2. Munich: Lincom Academic Reference Books.

55) Díaz-Negrillo, Ana, Nicolas Ballier & Paul Thompson (eds.) 2013. *Automatic Treatment and Analysis of Learner Corpus Data*. (Studies in Corpus Linguistics 59). Amsterdam: John Benjamins Publishing Co.

56) Díaz-Negrillo, Ana, Detmar Meurers, Salvador Valera & Holger Wunsch. 2010. *Towards Interlanguage POS Annotation for Effective Learner Corpora in SLA and FLT*. Language Forum 36(1-2): 139–154.

57) Dickinson, Markus & Detmar Meurers. 2003. Detecting Errors in Part-of-Speech Annotation. *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics EACL 03*, 1:107. Budapest: Association for Computational Linguistics.

58) Dickinson, Markus, Soojeong Eom, Yunkyoung Kang, Chong Min Lee & Rebecca Sachs. 2008. A Balancing Act: How Can Intelligent Computer-generated Feedback Be Provided in Learner-to-learner Interactions? *Computer Assisted Language Learning* 21(4): 369–382.

59) Ellis, Rod. 1994. *The Study of Second Language Acquisition*. Oxford: Oxford University Press.

60) Ellis, Rod & Gary Barkhuizen. 2005. *Analysing Learner Language*. Oxford: Oxford University Press.

61) Fagin, Ronald, Phokion G. Kolaitis, Renée J. Miller & Lucian Popa. 2005. Data Exchange: Semantics and Query Answering. *Theoretical Computer Science* 336(1): 89–124.

62) Flach, Peter A. 2012. *Machine Learning: the Art and Science of Algorithms that Make Sense of Data*. Cambridge: Cambridge University Press.

63) Fournier-Viger, Philippe, Roger Nkambou, André Mayers, Engelbert Mephu Nguifo & Usef Faghihi. 2011. An Hybrid Expert Model to Support Tutoring Services in Robotic Arm Manipulations. In Ildar Batyrshin & Grigori Sidorov (eds.), *Advances in Artificial Intelligence*: 478–489. (Lecture Notes in Computer Science 7094). Berlin: Springer.

64) Francis, W. Nelson & Henry Kučera. 1964. *Brown Corpus Manual - Manual of Information to Accompany a Standard Corpus of Present-day Edited American*

English, for Use with Digital Computers. Providence, Rhode Island: Department of Linguistics, Brown University.

65) Fraser, Thomas & André Joly. 1979. Le Système de la deixis - esquisse d'une théorie d'expression en anglais. *Modèles linguistiques* 1(2) : 97–157.

66) —. 1980. Le Système de la deixis (2) - esquisse d'une théorie d'expression en anglais. *Modèles linguistiques* 2(2) : 22–49.

67) Frei, Henri. 2011 [1929]. *La Grammaire des fautes*. Rennes: Presses Universitaires de Rennes.

68) Gaillat, Thomas. 2013a. *This* and *That* in Native and Learner English: From Typology of Use to Tagset Characterisation. In Sylviane Granger, Gaëtanelle Gilquin & Fanny Meunier (eds.), *Twenty Years of Learner Research: Looking Back, Moving Ahead*: 167–177. (Corpora and Language in Use). Louvain-la-Neuve: Presses Universitaires de Louvain.

69) —. 2013b. *This* et *that* dans les domaines spécialisés du corpus ICE-GB : quelles caractéristiques distributionnelles ?. *ASp. la revue du GERAS* (64) : 161–183.

70) —. 2013c. Annotation automatique d'un corpus d'apprenants d'anglais avec un jeu d'étiquettes modifié du Penn Treebank. *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)* : 271–284. Les Sables d'Olonne : Association pour le Traitement Automatique des Langues.

71) Gaillat, Thomas, Pascale Sébillot & Nicolas Ballier. 2014. Automated Classification of Unexpected Uses of *This* and *That* in a Learner Corpus of English. In Lieven Vandelanotte, Kristin Davidse, Caroline Gentens & Ditte Kimps (eds.), *Recent Advances in Corpus Linguistics: Developing and Exploiting Corpora*: 309–324. Amsterdam: Rodopi.

72) Gard, B. Jenset. 2008. *Basic statistics for corpus linguistics*. Unpublished MS written as course material for a methods in linguistics seminar taught at the University of Bergen in 2008.

73) Garside, Roger. 1987. The CLAWS Word-tagging System. In Geoffrey Sampson, Roger Garside & Geoffrey Leech (eds.), *The Computational Analysis of English: A Corpus-based Approach,* 30-41. London: Longman.

74) Gerboin, Pierre & Christine Leroy. 1991. *Grammaire d'usage de l'espagnol contemporain*. Paris: Hachette éducation.

75) Götz, Sandra & Marco Schilk. 2011. Formulaic Sequences in Spoken ENL, ESL and EFL: Focus on British English, Indian English and Learner English of Advanced German Learners. In Joybrato Mukherjee & Marianne Hundt (eds.), *Exploring Second-language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*: 79–100. (Studies in Corpus Linguistics). Amsterdam: John Benjamins Publishing Co.

76) Granger, Sylviane. 1993. International Corpus of Learner English. In Jan Aarts, Pieter de Haan & Nellake Ostdijk (eds.), *English Language Corpora: Design, Analysis and Exploitation : Papers from the Thirteenth International Conference on English Language Research on Computerized Corpora, Nijmegen 1992*: 57–72. Amsterdam: Rodopi.

77) —. 1994. The Learner Corpus: a Revolution in Applied Linguistics. *English Today* 10(39-3): 25–29.

78) —. 1996. From CA to CIA and Back: An Integrated Approach to Computerized Bilingual and Learner Corpora. In Karin Aijmer, Bengt Altenberg & Mats Johansson (eds.), *Languages in Contrast. Text-based Cross-linguistic Studies*, vol. 88:37–51. Lund: Lund University Press.

79) —. 1998. *Learner English on Computer. Studies in Language and Linguistics.* Geoffrey Leech & Jenny Thomas (eds.). Harlow: Longman.

80) —. 2001. Prefabricated Patterns in Advanced EFL Writing: Collocations and Formulae. In Anthony P. Cowie (ed.), *Phraseology: Theory, Analysis, and Applications*: 145–160. Oxford: Oxford University Press.

81) —. 2002. A Bird's Eye View of Learner Corpus Research. In Joseph Hung, Stephanie Petch-Tyson, and Sylviane Granger (eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*: 3–33. (Language learning and language teaching 6). Amsterdam: John Benjamins Publishing Co.

82) —. 2008. Learner Corpora in Foreign Language Education. In Nancy H. Hornberger (ed.), *Encyclopedia of Language and Education*, vol. 4: 337–351. Boston: Springer.

83) Granger, Sylviane, Gaëtanelle Gilquin & Fanny Meunier (eds.). 2015. *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press.

84) Gries, Stefan Th. 2003. *Multifactorial Analysis in Corpus Linguistics: A Study of Particle Placement*. London: Bloomsbury Publishing.

85) —. 2008. Corpus-based Methods in Analysis of Second Language Acquisition Data. In Peter Robinson & Nick C. Ellis (eds.), *Handbook of Cognitive Linguistics and Second Language Acquisition*. London: Routledge.

86) —. 2010. Useful Statistics for Corpus Linguistics. In Aquilino Sánchez & Moisés Almela (eds.), *A Mosaic of Corpus Linguistics: Selected Approaches*, 269–291. Frankfurt am Main: Peter Lang.

87) —. 2013 [2009]. *Statistics for Linguistics with R: a Practical Introduction*. *2nd revised ed.*. Berlin: De Gruyter Mouton.

88) Gries, Stefan Th. & Andrea L. Berez. [Forthcoming]. Linguistic Annotation in/for Corpus Linguistics. http://www.linguistics.ucsb.edu/faculty/stgries/research/InProgr_STG_ALB_LingAnnotCorpLing_HbOfLingAnnot.pdf (27 March, 2016.)

89) Gries, Stefan Th. & Sandra C. Deshors. 2014. Using Regressions to Explore Deviations between Corpus Data and a Standard/Target: Two Suggestions. *Corpora* 9(1). 109–136.

90) Grosz, Barbara J., Scott Weinstein & Aravind K. Joshi. 1995. Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics* 21(2): 203–225.

91) Guillemin-Flescher, Jacqueline. 1993. Étude contrastive de la deixis en anglais et en français. In Laurent Danon-Boileau & Jean Louis Duchet (eds.), *Opérations énonciatives et interprétations de l'énoncé* : 181–208. Paris: Ophrys.

92) de Haan, Pieter. 2000. Tagging Non-native English with the TOSCA-ICLE Tagger. In Christian Mair & Marianne Hundt (eds.), *Corpus Linguistics and Linguistic Theory*: 69-80. Amsterdam: Rodopi.

93) Halliday, M. A. K. & Ruqaiya Hasan. 1976. *Cohesion in English*. (English Language Series). Harlow: Pearson Education Limited.

94) Hasselgren, Angela. 1994. Lexical Teddy Bears and Advanced Learners: a Study into the Ways Norwegian Students Cope with English Vocabulary. *International Journal of Applied Linguistics* 4(2): 237–258.

95) Hawkins, John A. & Paula Buttery. 2010. Criterial Features in Learner Corpora: Theory and Illustrations. *English Profile Journal* 1(01).

96) Hawkins, John A. & Luna Filipović. 2012. *Criterial Features in L2 English: Specifying the Reference Levels of the Common European Framework*. Suffolk: Cambridge University Press.

97) Heid, Ulrich, Holger Voorman, Jan-Torsten Milde, Ulrike Gut, Katrin Erk & Sebastian Pado. 2004. Querying both Time-aligned and Hierarchical Corpora with NXT Search. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*: 1455–1458. Lisbon: European Language Resource Association.

98) Huddleston, Rodney & Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Beccles: Cambridge University Press.

99) Izumi, Emi, Kiyotaka Uchimoto & Hitoshi Isahara. 2005. Error Annotation for Corpus of Japanese Learner English. *Proceedings of the Sixth International Workshop on Linguistically Interpreted Corpora*: 71–80. Jeju Island: Asian Federation of Natural Language Processing.

100) Johnson, Keith. 2008. *Quantitative Methods in Linguistics*. Malden, MA: Wiley-Blackwell.

101) Kamp, Hans & Reyle Uwe. 1993. *From Discourse to Logic (Vol. 42)*. Dordrecht: Springer.

102) Keenan, Edward L. & Bernard Comrie. 1977. Noun Phrase Accessibility and Universal Grammar. *Linguistic Inquiry* 8(1): 63–99.

103) Kleiber, Georges. 1991. Anaphore-deixis : où en sommes-nous ? *L'Information grammaticale* 51(1): 3–18.

104) —. 1992. Anaphore-deixis : deux approches concurrentes. In Laurent Danon-Boileau & Mary-Annick Morel (eds.), *La Deixis*: 613–626. Paris: Presses Universitaires de France.

105) —. 1994. *Anaphores et pronoms*. Louvain-la-Neuve: Duculot.

106) —. 2001. *L'Anaphore associative*. Paris: Presses Universitaires de France.

107) Komen, Erwin R. 2012. Coreferenced Corpora for Information Structure Research. VARIENG 10. (Studies in Variation, Contacts and Change in English). http://www.helsinki.fi/varieng/series/volumes/10/komen/. (27 March, 2016.)

108) —. 2013. CESAX : Coreference Editor for Syntactically Annotated XML corpora - Reference Manual. Nijmegen: Radboud University.

109) Lapaire, Jean-Rémi. 1987. Du notionnel-lexical au métaopérationnel modal : étude des opérateurs 0, a, th-e/-is/-at/en/-ere en Anglais. Thèse de doctorat. Paris: Université de Paris 3.

110) Lapaire, Jean-Rémi & Wilfrid Rotgé. 1991. *Linguistique et grammaire de l'anglais*. Toulouse: Presses Universitaires du Mirail.

111) Lappin, Shalom. 2005. A Sequenced Model of Anaphora and Ellipsis Resolution. In Antonio Branco, Tony McEnery & Ruslan Mitkov (eds.), *Anaphora Processing Linguistic, Cognitive and Computational Modelling*, vol. 263: 3-16 (Current Issues in Linguistic Theory IV). Amsterdam: John Benjamins Publishing Co. http://www.dcs.kcl.ac.uk/staff/lappin/papers/daarc_chap.pdf (27 March, 2016.)

112) Leacock, Claudia. 2010. *Automated Grammatical Error Detection for Language Learners*. California: Morgan & Claypool Publishers.

113) Leech, Geoffrey. 2005. Adding Linguistic Annotation. In Martin Wynne (ed.), *Developing Linguistic Corpora: a Guide to Good Practice*: 17–29. Oxford: Oxbow Books.

114) Lenko-Szymanska, Agnieszka. 2004. Demonstratives as Anaphora Markers in Advanced Learners' English. In Guy Aston, Sylvia Bernardini & Dominic Stewart (eds.), *Corpora and Language Learners*: 84–108. (Studies in Corpus Linguistics 17). Amsterdam: John Benjamins Publishing Co.

115) Levy, Roger & Andrew Galen. 2006. Tregex and Tsurgeon: Tools for Querying and Manipulating Tree Data Structures. *5th International Conference on Language Resources and Evaluation (LREC 2006)*. Genoa, Italy.

116) Liang, Xia. 2009. A Corpus-based Study of Developmental Stages of Demonstratives in Chinese English Majors' Writing. *Asian Social Science* 5(11): 117–125.

117) Lozano, Cristóbal & Amaya Mendikoetxea. 2013. Learner Corpora and Second Language Acquisition: The Design and Collection of CEDEL2. *Automatic Treatment and Analysis of Learner Corpus Data*: 65–100. (Studies in Corpus Linguistics 65). Amsterdam: John Benjamins Publishing Co.

118) Lüdeling, Anke & Hagen Hirschmann. 2015. Error Annotation Systems. In Sylviane Granger, Gaëtanelle Gilquin, & Fanny Meunier (eds.), *The Cambridge Handbook of Learner Corpus Research*: 135-158. Cambridge: Cambridge University Press.

119) Lüdeling, Anke, Maik Walter, Emil Kroymann, & Peter Adolphs. 2005. Multi-level Error Annotation in Learner Corpora. In *Proceedings of Corpus Linguistics Conference Series*, vol. 1: 105-115. Birmingham: Centre for Corpus Research. https://www.linguistik.huberlin.de/de/institut/professuren/korpuslinguistik/forschung/falko/pdf/FALKO-CL2005.pdf. (27 March, 2016.)

120) Lyons, John. 1968. *Introduction to Theoretical Linguistics*. Cambridge: Cambridge University Press.

121) Marcus, Mitchell P., Mary Ann Marcinkiewicz & Beatrice Santorini. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics* 19(2): 313–330.

122) McEnery, Tony, Richard Xiao & Yukio Tono. 2006. *Corpus-based Language Studies: an Advanced Resource Book*. (Routledge Applied Linguistics). New York: Routledge.

123) Meunier, Fanny, Sylviane Granger, Damien Littré & Magali Paquot. 2009. *The LONGDALE (Longitudinal Database of Learner English)*. UCL-CECL. .

124) Meurers, Detmar. 2012. Natural Language Processing and Language Learning. In Carol A. Chapelle (ed.), *The Encyclopedia of Applied Linguistics*: 4193–4205. Blackwell Publishing Ltd.

125) —. 2015. Learner Corpora and Natural Language Processing. In Sylviane Granger, Gaëtanelle Gilquin & Fanny Meunier (eds.), *The Cambridge Handbook of Learner Corpus Research*: 537-565. Cambridge: Cambridge University Press.

126) de Mönnink, Inge, 2000. Parsing a Learner Corpus. In Christian Mair & Marianne Hundt (eds.), *Corpus Linguistics and Linguistic Theory*: 82–90. Amsterdam: Rodopi.

127) Moulin, Michel, Henri Odin & Janine Bouscaren. 1996. *Pratique raisonnée de la langue: initiation à une grammaire de l'énonciation pour l'étude et l'enseignement de l'anglais*. Paris: Ophrys.

128) Na-Rae Han, Martin Chodorow & Claudia Leacock. 2006. Detecting Errors in English Article Usage by Non-native Speakers. *Natural Language Engineering* 2(12): 115–129.

129) Nelson, Gerald, Sean Wallis & Bas Aarts. 1998. *The British Component of the International Corpus of English (ICE-GB) and ICECUP software (CD-ROM)*. London: University College London.

130) Nicholls, Diane. 2003. The Cambridge Learner Corpus - Error Coding and Analysis for Lexicography and ELT. In Dawn Archer, Paul Rayson & Tony McEnery (eds.), *Proceedings of the Corpus Linguistics 2003 conference*, vol. 16. (UCREL Technical Paper). Lancaster: UCREL Lancaster University.

131) Nkambou, Roger. 2010. Modeling the Domain: an Introduction to the Expert Module. In Roger Nkambou, Jacqueline Bourdeau & Riichiro Mizoguchi (eds.), *Advances in Intelligent Tutoring Systems*: 15–32. (Studies in Computational Intelligence 308). Berlin: Springer.

132) Nøklestad, Anders & Ashild Søfteland. 2007. Tagging a Norwegian Speech Corpus. In Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek & Mare Koit (eds.), *NODALIDA 2007 Conference Proceedings*: 245–248. Tartu: Northern European Association for Language Technology

133) Paice, C. D. & G. D. Husk. 1987. Towards the Automatic Recognition of Anaphoric Features in English Text: the Impersonal Pronoun “it.” *Computer Speech and Language*(2): 109–132.

134) Petch-Tyson, Stephanie. 2000. Demonstrative Expressions in Argumentative Discourse. In Simon Botley & Anthony Mark McEnery (eds.), *Corpus-based and Computational Approaches to Discourse Anaphora*. (Studies in Corpus Linguistics). Amsterdam: John Benjamins Publishing Co.

135) Pradhan, Aliva M., Aparna S. Varde, Jing Peng & Eileen M Fitzpatrick. 2010. Automatic Classification of Article Errors in L2 Written English. *Proceedings of the Twenty-Third International FLAIRS Conference*: 259-264. Florida: Association for the Advancement of Artificial Intelligence (AAAI).

136) Quinlan, John Ross. 1986. Induction of Decision Trees. *Machine Learning* 1(1): 81–106.

137) Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A Grammar of Contemporary English*. London: Longman.

138) R Core Team. 2012. *R: A language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

139) Rooy, Bertus van & Lande Schafer. 2003. An Evaluation of Three POS Taggers for the Tagging of the Tswana Learner English Corpus. In Dawn Archer, Paul Rayson, Andrew Wilson & Tony McEnery (eds.), *Proceedings of the Corpus Linguistics 2003 conference*, vol. 16: 835–844. Lancaster: University Centre For Computer Corpus Research On Language.

140) Schmid, H. 1994a. *How to Use the TreeTagger*. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/. (27 March, 2016.)

141) —. 1994b. Probabilistic Part-of-speech Tagging Using Decision Trees. *Proceedings of the International Conference on New Methods in Language Processing*: 14–16. Manchester. ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tree-tagger1.pdf. (27 March, 2016.)

142) Schneider, Gerold. 2015. Ana Díaz-Negrillo, Nicolas Ballier and Paul Thompson (Eds.), *Automatic Treatment and Analysis of Learner Corpus Data*. [Book review]. *International Journal of Learner Corpus Research* 1(1): 172–177. https://benjamins.com/#catalog/journals/ijlcr.1.1.07sch/fulltext. (27 March, 2016.).

143) Selinker, Larry. 1972. Interlanguage. *International Review of Applied Linguistics in Language Teaching* 10(3): 209.

144) Sérasset, Gilles, Andreas Witt, Ulrich Heid & Felix Sasaki. 2009. Multilingual Language Resources and Interoperability. *Language Resources and Evaluation* 43(1): 1-14.

145) Sinclair, John. 2005. Corpus and Text: Basic Principles. In Martin Wynne (ed.), *Developing Linguistic Corpora: a Guide to Good Practice*: 1-16 Oxford: Oxbow. http://www.ahds.ac.uk/guides/linguistic-corpora/chapter1.htm. (27 March, 2016.)

146) Soon, Wee Meng, Hwee Tou Ng & Daniel Chung Yong Lim. 2001. A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Computational Linguistics* 27(4): 521-544.

147) Stefanowitsch, Anatol & Stefan Th. Gries. 2003. Collostructions: Investigating the Interaction of Words and Constructions. *International Journal of Corpus Linguistics* 8(2): 209–243.

148) Stirling, Lesley & Rodney Huddleston. 2002. Deixis and Anaphora. In Rodney Huddleston & Geoffrey K. Pullum (eds.), *The Cambridge Grammar of the English Language*: 1449–1564. Beccles: Cambridge University Press.

149) Strauss, Susan. 2002. *This*, *that*, and *it* in Spoken American English – a Demonstrative System of Gradient Focus. *Language Sciences* 24(2): 131–152.

150) Tanguy, Ludovic & Nabil Hathout. 2007. *Perl pour les Linguistes*. (Tic et Sciences Cognitives). Paris: Lavoisier.

151) Tarone, Elaine & Betsy Parrish. 1988. Task-Related Variation in Interlanguage: The Case of Articles. *Language Learning* 38(1): 21–44.

152) Tono, Yukio. 2013. Automatic Extraction of L2 Criterial Lexicogrammatical Features across Pseudo-longitudinal Learner Corpora: Using Edit Distance and Variability-based Neighbour Clustering. In Camilla Bardel, Christina Lindqvist & Batia Laufer (eds.), *L2 Vocabulary Acquisition, Knowledge and Use: New Perspectives on Assessment and Corpus Analysis*: 149–176. (Eurosla Monographs Series 2). The European Second Language Association. http://www.eurosla.org/monographs/EM02/EM02tot.pdf. (27 March, 2016.)

153) Venables, William N. & Brian D. Ripley. 2000. *S Programming*. New York: Springer.

154) Vogel, Klaus. 1995. *L'Interlangue: la langue de l'apprenant*. Toulouse: Presses Universitaires du Mirail.

155) Wynne, Martin (ed.). 2005. *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow.

156) Young, Richard. 1996. Form-Function Relations in Articles in English Interlanguage. In Robert Bayley & Dennis R. Preston (eds.), *Second Language Acquisition and Linguistic Variation*: 135-175. (Studies in Bilingualism 10). Amsterdam: John Benjamins Publishing Co.

157) Zhang, Jing. 2015. An Analysis of the Use of Demonstratives in Argumentative Discourse by Chinese EFL Learners. *Journal of Language Teaching and Research* 6(2): 460–465.

Table of Annexes

Annex A - Utterance comparisons: learner v. natives

In Chapter 3, we use the Diderot-LONGDALE corpus to extract learner occurrences of *it*, *this* and *that*. To carry out an error analysis, we select unexpected or erroneous occurrences. We evaluate the error annotation process in comparison with natives. In order to assess our annotation predictions of erroneous uses, it is necessary to identify what the correct/expected uses are in the same contexts as the learners'. For this purpose eleven erroneous learner utterances were proposed to a panel of six natives and they were requested to fill the gaps that correspond to the forms. This annex shows the utterances which were submitted together with their introductory paragraph dedicated to explaining the questionnaire's purpose.

The following contexts were pronounced by different learners of English. In order to see the difference between learner and native use of specific words, could you, please, fill the gaps? Please, use all the words that you think could be used. (Reformulation might be your choice but could you, nevertheless, apply the most appropriate word, please?)

(DID0160-S0001) I liked (er) . the Independence Hall ……..1…….. was really interesting . and (er) . I went to: the: U Penn U Penn University and I really liked it .

(DID0121-S001) we we see a (em) a romance for (em) the guy’s eyes because most of the time ……..2……..(be) the girl who is telling the story about was bad and and blablabla

(DID0115-S001) <A (native speaker)> would you consider pizza an Italian food </A><B (learner)>(em) yes but it’s not it’s not really f= it’s typic but it’s not (em) we can eat that everyday everywhere now and . but (em) my grandma does ……..3…….. by herself

(DID0158-S0001) first of all what I loved the most of the world is ice hockey . just but I I loved ……..4…….. (reference to hockey) (er) I loved ……..5…….. (reference to hockey again) before going there (Canada)

(DID0164-S001) so at the end she's an old lady she writes a book . and actually the book the scene that we saw when she apologize is (er) . she wrote that story on her book so ……..6…….. was her way to (eh) try to (em) to apologize to them through the book but (er) </B>

(DID0118-S001) French French is very proud actually and they say yeah we're very open minded we can yeah but that that's not true we can see ……..7…….. with all the problems in this at ……..8…….. moment

(DID0157-S001) (Student talking about her trip to NYC) (er) you know you can talk with people in the streets (er) it's (er) . it's really nice I'm not used to ……..9…….. in France so (...)

(DID0112-S001) <A> okay how old were you when you visited for the first time </A><B> I was twelve years old and I went there with my mum because my family lives there my grandm my grandmother and my my uncles and my cousins (laughter) </B><A> (mhm) And do you have any particular memories from that first trip </A><B> oh yes because I didn't know how to speak ……..10…….. language so when I went there I didn't even know how how to say . hello to my grandmother

(DID0155-S001) I remember there was this little comic book store just just . like one block away from the . the: the Empire State Building so every time I like had an hour I would go and buy my coffee at Starbucks <begin laughter> <end laughter> anyhow (er) . Greenwich Village was very nice (er) I remember spending a lot of time at (er) ……..11…….. bookstore called Barnes and Nobles . and . those were . like it was like huge in the middle of Greenwich Village . so I spend a lot of time over there . (er) .

Annex B - Corpus specifications

The following table gives a recap on the specifications of the three corpora as they were used in our studies.

Corpora	Diderot-LONGDALE (Meunier *et al.* 2008)	NOCE (Díaz-Negrillo 2007)	WSJ (Marcus *et al.* 1993)
# of tokens	131,361	39,015	1,039562
# of tokens in manually checked subsets	64,858	13,143	55,765
# of different speakers in subsets	12	45	NA
# files in subsets	36	45	40
# of recording/sampling sessions	4	3	NA
Nature	Longitudinal	Cross-sectional	Cross-sectional
Types of task	Description[40] of personal experiences and pictures	Expressing personal opinion on topical issues	Writing news articles
Registers	Conversation	Descriptive, narrative, argumentative or free-writing essay	News
PoS-tagged	Penn Treebank	Penn Treebank	Penn Treebank
Error annotation	No	Yes (Díaz-Negrillo 2009)	No
Syntactic parsing	No	No	Yes
Type of data format	Text	XML and text	Bracketed text

Note: At this stage the LONGDALE corpus is only accessible to members of the LONGDALE project (see https://www.uclouvain.be/en-cecl-longdale.html)

[40] The description does not include the last recordings which consist of a map task in which learners had to interact.

Annex C - LINDSEI Transcription guidelines

The LINDSEI guidelines give details on the various aspects to take into consideration when transcribing oral recordings for the LONGDALE corpus.

1. Interview identification

Each interview is preceded by a code of this type: <h nt="FR" nr="*FR+three-figure number*">

e.g.: <h nt="FR" nr="FR004"> (4th interview with French mother tongue student)

Examples of country codes:

DUTCH = DU001

GERMAN = GE001

SWEDISH = SW001

FINNISH = FI001

SPANISH = SP001

CZECH = CZ001

NORWEGIAN = NO001

2. Speaker turns

Speaker turns are displayed in vertical format, i.e. one below the other. Whilst the letter 'A' enclosed between angle brackets always signifies the interviewer's turn, the letter 'B' between angle brackets indicates the interviewee's (learner's) turn. The end of each turn is indicated by either </A> or </B>.

e.g.: <A> okay so which topic have you chosen </A>

<B> the film or play that I thought was particularly good or bad really </B>

3. Overlapping speech

The tag <overlap /> (with a space between "overlap" and the slash) is used to indicate the beginning of overlapping speech. It should be indicated in both turns.

e.g.: <B> yeah I went on a bus to London once and I'll never <overlap /> do it again </B>

<A> <overlap /> that's even worse </A>

4. Punctuation

No punctuation marks are used to indicate sentence or clause boundaries.

5. Empty pauses

Empty pauses are defined as a blank on the tape, i.e. no sound, or when someone is just breathing. The following three tier system is used: one dot for a 'short' pause (< 1 second), two dots for a 'medium' pause (1-3 seconds) and three dots for 'long' pauses (> 3 seconds).

e.g.: <B> erm .. it's a British film there aren't many of those these days </B>

6. Filled pauses and backchannelling

Filled pauses and backchannelling are marked as (eh) [brief], (er), (em), (erm), (mm), (uhu) and (mhm). No other fillers should be used.

e.g.: <B> yeah . well Namur was warmer (er) it was (eh) a really little town </B>

7. Unclear passages

A three tier system is used to indicate the length of unclear passages: <X> represents an unclear syllable or sound up to one word, <XX> represents two unclear words, and <XXX> represents more than two words.

e.g.: <B> <X> they're just begging <XX> there's there's honestly he did a course .. for a few weeks </B>

If transcribers are not entirely sure of a word or word ending, they should indicate this by having the word directly followed by the symbol '<?>'.

e.g.: <B>: I went to see a<?> friend at university there and stayed </B>

Unclear names of towns or titles of plays for example may be indicated as '<name of city>' or '<title of play>'.

e.g.: <B>: where else did we go er <name of city> it's in Bolivia </B>

8. Truncated words

Truncated words are immediately followed by an equals sign.

e.g.: <B> it still resem= resembled the theatre </B>

9. Contracted forms

All standard contracted forms are retained as they are typical features of speech.

10. Non-standard forms

Non-standard forms that appear in the dictionary are transcribed orthographically in their dictionary accepted way: *cos, dunno, gonna, gotta, wanna* and *yeah*.

11. Foreign words and pronunciation

Foreign words are indicated by <foreign> (before the word) and </foreign> (after the word).

e.g.: <B> we couldn't go with er knives and so on <foreign> enfin </foreign> we were er </B>

As a rule, foreign pronunciation is not noted, except in the case where the foreign word and the English word are identical. If in this case the word is pronounced as a foreign word, this is also marked using the <foreign> tag.

e.g.: <B> I didn't have the erm . <foreign> distinction </foreign> </B>

12. Acronyms

If acronyms are pronounced as sequences of letters, they are transcribed as a series of upper-case letters separated by spaces.

e.g.: <B> yes not really I did sort of basic G C S E French and German </B>

If, on the other hand, acronyms are pronounced as words, they are transcribed as a series of upper-case letters not separated by spaces.

e.g.: <A> mhm er you're doing a MAELT </A>

13. Dates and numbers

Figures have to be written out in words. This avoids the ambiguity of, for example, “1901”, which could be spoken in a number of different ways.

e.g.: <B> an awful lot of people complain and say well the grants were two thousand two hundred </B>

14. Nonverbal vocal sounds

Nonverbal vocal sounds are enclosed between angle brackets.

e.g.: <B> I hope so I’ve I’ve got some <coughs> friends out there </B>

e.g.: <B> so I went back into Breda . and sat down again <imitates the sound of a guitar> </B>

15. Contextual comments

Non-linguistic events are indicated between angle brackets only if they are deemed relevant to the interaction (if one of the participants reacts to it, for example).

e.g.: <A> no it’s true it’s nice to have your own bathroom </A>

<somebody enters the room>

<B> hi </B>

16. Prosodic information: voice quality

If a particular stretch of text is said laughing or whispering for instance, this is marked by inserting <begin laughter> or <begin whisper> immediately before

the specific stretch of speech and <end laughter> or <end whisper> at the end of it.

e.g.: <B> <begin laughter> I don't have to assess it I only have to write it <end laughter> </B>

17. Phonetic features

(a) Syllable lengthening

A colon is used to indicate that the preceding syllable is lengthened. Colons should not be inserted inside words.

e.g.: <B> that's something I'll I'll plan to: to learn </B>

(b) Articles

-when pronounced as [ei], the article 'a' is transcribed as 'a[ei]';

e.g.: <B> and it's about erm . life in a[ei] eh public school in America I think </B>

-when pronounced as [i:] the article 'the' is transcribed as 'the[i:]'.

e.g.: <B> and the[i:] villa we were staying in was in one of the valleys </B>

18. Tasks

The three tasks making up the interview (set topic, free discussion and picture description) should be separated from each other. This is done using the following tags: <S> (before the set topic), </S> (after the set topic), <F> (before the free discussion), </F> (after the free discussion), <P> (before the picture

description), </P> (after the picture description). These tags should occupy a separate line and should not interrupt a turn.

e.g.: <S>

<A> did you . manage to choose a topic </A>

19. End

All interviews should end with the following tag (on a separate line): </h>

20. Questions?

If you have any questions regarding these transcription guidelines, don't hesitate to get in touch with us!

Annex D - Description of the tasks given to students

This annex gives a description of the tasks and topics of conversations that were given to the learners. Each task included conversations between a native language assistant and a learner. Each learner was recorded several times in two years.

Spoken tasks S001 (October 2009)

Task 1: modelled on LINDSEI directives: spontaneous, monologue + conversational mode

Part 1 (Set task): Discuss one of the following topics

- Topic 1: A good or a bad experience
- Topic 2: A country you visited
- Topic 3: A film, a play, a sport event, a concert you saw

Part 2(Free task) : Informal conversation

Life interests, hobbies and academic choices

Task 2: Reading aloud Task (The Selfish Giant, Oscar Wilde) (unplanned/one minute preparation)

Spoken tasks S002 (June-October 2010)

Topics for oral expression: Choose one of the following subjects and be ready to talk about it for a few minutes.

1) Topic one : Suppose you have time and money to travel or move to a different country/city, where will you go ? Why ? How will you organise your new life ?

2) Topic two : Can you tell me about an important event, experience or meeting which has made a difference or changed your life in the past six months ?

3) Topic 3 : Do you feel creative ? Tell me about a work of art you would like to create or participate in : a play, a film, a musical event, a book, a painting, a computer game, etc. How would you go about it ?

• The interviewer will ask you some additional questions.

• Reading task : Get ready to read aloud this speech by David Cameron

Spoken tasks S003 (March 2011)

Task 1: Commenting paintings followed by free part: discussion

You are going to see four works of art (paintings), one after the other. I'd like you to react to each of them quite spontaneously and tell me how you feel about them

Additional questions:

• Can you justify, explain why you like or dislike picture one, two, three, four ?

• Which of these four pictures would you like to have at home, in your room?

• If you were to take one of those pictures to illustrate a book you want to write, which one would you choose?

Task 2: Reading Task: The Landlady, Roald Dahl

Annex E - Penn Treebank PoS tagset

This is a description of the Penn Treebank PoS tagset (Taylor, Marcus, and Santorini 2003, 317)

POS Tag	Description
CC	coordinating conjunction
CD	cardinal number
DT	determiner
EX	existential there
FW	foreign word
IN	preposition/subordinating conjunction
JJ	adjective
JJR	adjective, comparative
JJS	adjective, superlative
LS	list marker
MD	modal
NN	noun, singular or mass
NNS	noun plural
NNP	proper noun, singular
NNPS	proper noun, plural
PDT	predeterminer
POS	possessive ending
PRP	personal pronoun
PRP$	possessive pronoun
RB	adverb
RBR	adverb, comparative
RBS	adverb, superlative
RP	particle
TO	to
UH	interjection
VB	verb, base form
VBD	verb, past tense
VBG	verb, gerund/present participle
VBN	verb, past participle
VBP	verb, sing. present, non-3d
VBZ	verb, 3rd person sing. present
WDT	wh-determiner
WP	wh-pronoun
WP$	possessive wh-pronoun
WRB	wh-abverb

Annex F - Tregex queries to identify occurrences of *that* in the Penn Treebank

This annex shows a comprehensive inventory of the queries that are used to identify and, subsequently, modify the Penn Treebank's PoS tags. We use Tregex query patterns which correspond to syntactic configurations in the Penn Treebank to identify all occurrences of *that*. These queries map the PoS tags to their syntactic constituent path. Green-coloured patterns indicate cases that were initially tagged otherwise (tagging errors) in the Penn Treebank. Each line corresponds to a query that matches the tag assigned in the corpus. The queries are used to correct the PoS tags in the Penn Treebank and to introduce a specific finer-grained tagset for *it, this* and *that*.

True class	Occurrences of THAT (Corrections of tagging errors in green)	Nb of occurrences
IN total	/^IN$/ < /^[T\|t]hat$/	5992
IN (complementiser) breakdown	/^IN$/ < /^[T\|t]hat$/ > /SBAR.*/	5408
	/^IN$/ < /^[T\|t]hat$/ > (/SBAR.*/ > /^NP.*/)	827
	/^IN$/ < /^[T\|t]hat$/ > (/SBAR.*/ > /^VP.*/)	3883
	/^IN$/ < /^[T\|t]hat$/ > (/SBAR.*/ > /^SBAR*/)	408
	/^IN$/ < /^[T\|t]hat$/ > (/SBAR.*/ > /UCP/)	4
	/^IN$/ < /^[T\|t]hat$/ > (/SBAR.*/ > /^ADJP.*/)	182
	/^IN$/ < /^[T\|t]hat$/ > (/SBAR.*/ > /^S$/)	41
	/^IN$/ < /^[T\|t]hat$/ > (/SBAR.*/ > /^PRN.*/)	4
	/^IN$/ < /^[T\|t]hat$/ > (/SBAR.*/ > /^FRAG.*/)	7
	/^IN$/ < /^[T\|t]hat$/ > (/SBAR.*/ > /^ADVP.*/)	23
	/^IN$/ < /^[T\|t]hat$/ > (/SBAR.*/ > /SINV.*/)	10
	/^IN$/ < /^[T\|t]hat$/ > (/SBAR.*/ > /PP.*/)	10
	/^IN$/ < /^[T\|t]hat$/ > (/SBAR.*/ > /SQ.*/)	4
	/^IN$/ < /^[T\|t]hat$/ > (/SBAR.*/ > / /)	3
	/^DT$/ < /^[T\|t]hat$/ > /^SBAR.*/	233
	/^DT$/ < /^[T\|t]hat$/ > (/SBAR.*/ > /^NP.*/)	23
	/^DT$/ < /^[T\|t]hat$/ > (/SBAR.*/ > /^VP.*/)	188
	/^DT$/ < /^[T\|t]hat$/ > (/SBAR.*/ > /^SBAR.*/)	20
	/^DT$/ < /^[T\|t]hat$/ > (/SBAR.*/ > /^ADJP.*/)	1
	/^DT$/ < /^[T\|t]hat$/ > (/SBAR.*/ > null)	1
		233
	/^WDT$/ < /^[T\|t]hat$/ > /^SBAR.*/	8
	Sub-total	5649
	/^IN.*/ < /^[T|t]hat$/ > /^WHADVP.*/	40
	/^IN*/ < /^[T|t]hat$/ > /^ADVP.*/	5

	/^IN*/ < /^[T|t]hat$/ > X	1
	/^DT$/ < /^[T\|t]hat$/ > /^WHADVP.*/	1
	Sub-total	47
	Total that complementiser	5696
WDT total	/^WDT.*/ < /^[T|t]hat$/	2483
WDT (relatives) breakdown	/^WDT$/ < /^[T\|t]hat$/ > /^WHNP.*/	2409
	/^WDT.*/ < /^[T|t]hat$/ > (/^WHNP.*/ >(/^SBAR.*/ > /^NP.*/))	2320
	/^WDT.*/ < /^[T|t]hat$/ > (/^WHNP.*/ >(/^SBAR.*/ > /^SBAR.*/))	21
	/^WDT.*/ < /^[T|t]hat$/ > (/^WHNP.*/ >(/^SBAR.*/ > /^VP.*/))	60
	/^WDT.*/ < /^[T|t]hat$/ > (/^WHNP.*/ >(/^SBAR.*/ > /^S$/))	1
	/^WDT.*/ < /^[T|t]hat$/ > (/^WHNP.*/ >(/^SBAR.*/ > /^S\-/))	1
	/^WDT.*/ < /^[T|t]hat$/ > (/^WHNP.*/ >(/^SBAR.*/ > /^SQ.*/))	1
	/^WDT.*/ < /^[T|t]hat$/ > (/^WHNP.*/ >(/^SBAR.*/ > /^PP.*/))	1
	/^WDT.*/ < /^[T|t]hat$/ > (/^WHNP.*/ >(/^SBAR.*/ > /^X-CLF/))	1
	/^WDT.*/ < /^[T|t]hat$/ > (/^WHNP.*/ >(/^SBAR.*/ > /^NX/))	1
	/^IN.*/ < /^[T|t]hat$/ > /^WHNP.*/	430
	/^IN.*/ < /^[T|t]hat$/ > (/^WHNP.*/ >(/^SBAR.*/ > /^NP.*/))	415
	/^IN.*/ < /^[T|t]hat$/ > (/^WHNP.*/ >(/^SBAR.*/ > /^SBAR.*/))	5
	/^IN.*/ < /^[T|t]hat$/ > (/^WHNP.*/ >(/^SBAR.*/ > /^VP.*/))	10
	/^DT$/ < /^[T\|t]hat$/ > /^WHNP.*/	58
	/^VBP.*/ < /^[T|t]hat$/	1
	/^DT$/ < /^[T\|t]hat$/ > (/^ADJP.*/ > (NP > NP))	1
	Total that relative pronoun	2899
DT total	/^DT$/ < /^[T\|t]hat$/	1918
Assigned DT	/^DT$/ < /^[T\|t]hat$/ > /^NP.*/	1614
Pro-form	/^DT$/ < /^[T\|t]hat$/ >- /^NP.*/	801
Determiner	/^DT$/ < /^[T\|t]hat$/ > /^NP.*/ !>- /^NP.*/	813
Pro-form	/^DT$/ < /^[T\|t]hat$/ > /^INTJ.*/	3

Assigned IN (complementiser)	/^IN.*/ < /^[T|t]hat$/ > /^NP.*/	106
Pro-form	/^IN.*/ < /^[T|t]hat$/ >- /^NP.*/	71
Determiner	/^IN.*/ < /^[T|t]hat$/ > /^NP.*/ !>-/^NP.*/	35
Assigned WDT (relatives)	/^WDT$/ < /^[T\|t]hat$/ > /^NP.*/	66
Pro-form	/^WDT$/ < /^[T\|t]hat$/ >- /^NP.*/	66
Determiner	/^WDT.*/ < /^[T|t]hat$/ > /^NP.*/ !>-/^NP.*/	0
Assigned NN		1
Pro-form	/^NN.*/ < /^[T|t]hat$/	1
	Total that determiner and pro-form	1790
RB total	/^RB$/ < /^[T\|t]hat$/	26
RB breakdown	/^RB$/ < /^[T\|t]hat$/ > /^ADVP.*/	8
	/^RB$/ < /^[T\|t]hat$/ > /^ADJP.*/	11
	/^RB$/ < /^[T\|t]hat$/ > /^SBAR.*/	5
	/^RB$/ < /^[T\|t]hat$/ > /^NP.*/	2
	/^DT$/ < /^[T\|t]hat$/ > /^ADVP.*/	5
	/^DT$/ < /^[T\|t]hat$/ > (/^ADJP.*/ > (NP > VP))	3
	/^IN.*/ < /^[T|t]hat$/ > /^ADJP.*/	2
	Total that adverbial	36
	TOTAL	10421
	Total for tagging errors	903
	Error tagging %	8.67
	Total counts for /^[T|t]hat$/	10421

Annex G - PERL script to compute accuracy

This script counts all occurrences of each form depending on its function in a TreeTagger formatted file. It then computes precision and recall.

```
# Copyright Thomas Gaillat – Université de Paris-Diderot
use strict;
use locale;
#script by Thomas Gaillat. Calculate confusion matrix and precision and
recall for specific tags in TT formated files.
# Arguments: First arg: file to evaluate - Second arg: reference file
my ( @tokens, @tags, @lemmata );
my ( @tokensRD, @tagsRD, @lemmataRD );

my $countit=0;
my $counttagsit=0;
my $counttagsRDit=0;

my $countitPRP=0;
my $counttagsitPRP=0;
my $counttagsRDitPRP=0;
my $counttagsitPRPbutRDPNR=0;

my $countitPNR=0;
my $counttagsitPNR=0;
my $counttagsRDitPNR=0;
my $counttagsitPNRbutRDPRP=0;


my $countthis=0;
my $counttagsthis=0;
my $counttagsRDthis=0;

my $countthat=0;
my $counttagsthat=0;
my $counttagsRDthat=0;

my $countthisDT=0;
my $counttagsthisDT=0;
my $counttagsRDthisDT=0;
my $counttagsthisDTbutRDTPRON=0;
my $counttagsthisDTbutRDRB=0;



my $countthisTPRON=0;
my $counttagsthisTPRON=0;
my $counttagsRDthisTPRON=0;
my $counttagsthisTPRONbutRDDT=0;
my $counttagsthisTPRONbutRDRB=0;


my $countthisRB=0;
my $counttagsthisRB=0;
my $counttagsRDthisRB=0;
my $counttagsthisRBbutRDDT=0;
my $counttagsthisRBbutRDTPRON=0;
```

```
my $counttagsthatRBbutRDTCOM=0;
my $counttagsthatRBbutRDTREL=0;



my $countthatDT=0;
my $counttagsthatDT=0;
my $counttagsRDthatDT=0;
my $counttagsthatDTbutRDTPRON=0;
my $counttagsthatDTbutRDTCOM=0;
my $counttagsthatDTbutRDTREL=0;
my $counttagsthatDTbutRDRB=0;

my $countthatTPRON=0;
my $counttagsthatTPRON=0;
my $counttagsRDthatTPRON=0;
my $counttagsthatTPRONbutRDDT=0;
my $counttagsthatTPRONbutRDTCOM=0;
my $counttagsthatTPRONbutRDTREL=0;
my $counttagsthatTPRONbutRDRB=0;

my $countthatRB=0;
my $counttagsthatRB=0;
my $counttagsRDthatRB=0;
my $counttagsthatRBbutRDDT=0;
my $counttagsthatRBbutRDTPRON=0;

my $countthatTCOM=0;
my $counttagsthatTCOM=0;
my $counttagsRDthatTCOM=0;
my $counttagsthatTCOMbutRDDT=0;
my $counttagsthatTCOMbutRDTPRON=0;
my $counttagsthatTCOMbutRDTREL=0;
my $counttagsthatTCOMbutRDRB=0;

my $countthatTREL=0;
my $counttagsthatTREL=0;
my $counttagsRDthatTREL=0;
my $counttagsthatTRELbutRDDT=0;
my $counttagsthatTRELbutRDTPRON=0;
my $counttagsthatTRELbutRDTCOM=0;
my $counttagsthatTRELbutRDTRB=0;


open IN, $ARGV[0] or die "can't open", $ARGV[0], "\n";
while(my $line=<IN>)
{
          chomp $line;
        my @array = split ( /\t/, $line );
          my @arraybis=split(/\s/, $array[1]);
    push ( @tokens, $array[0] );
    push ( @tags, $arraybis[0] );
    push ( @lemmata, $arraybis[1] );
}
```

```
close (IN);

open (IN, "<", $ARGV[1]);
while(my $line=<IN>)
{
          chomp $line;
         my @array= split ( /\t/, $line );
          my @arraybis=split(/\s/, $array[1]);
    push ( @tokensRD, $array[0] );
    push ( @tagsRD, $arraybis[0] );
    push ( @lemmataRD, $arraybis[1] );
}
close (IN);

###################COUNT TOKENS & TAGS####################
########################IT################################
#precision and recall for "it"
#count of Nc
for ( my $i = 0 ; $i <= ($#tags) ; $i++ ) {
   if((  $tags[$i] eq $tagsRD[$i]) and ($tokens[$i] =~/^[I|i]t$/) and
($tokensRD[$i] =~/^[I|i]t$/)) {
          $countit++;
    }
        #count of Na
   if((($tags[$i]) eq "PRP") or (($tags[$i]) eq "PNR") ) and
($tokens[$i] =~/^[I|i]t$/) ) {
             $counttagsit++;
   }
   #count of Nr
             if((($tagsRD[$i]) eq "PRP") or (($tagsRD[$i]) eq "PNR") )
and ($tokensRD[$i] =~/^[I|i]t$/)) {
             $counttagsRDit++;
   }
}


#precision and recall for "it" as PRP
#count of Nc
for ( my $i = 0 ; $i <= ($#tags) ; $i++ ) {
   if((  $tags[$i] eq $tagsRD[$i]) and ($tokens[$i] =~/^[I|i]t$/) and
($tagsRD[$i] =~/^PRP$/) and ($tokensRD[$i] =~/^[I|i]t$/)) {
          $countitPRP++;
    }

    #count of Na
   if((($tags[$i]) eq "PRP") and ($tokens[$i] =~/^[I|i]t$/) ) {
          $counttagsitPRP++;
   }
#count of Nr
             if(($tagsRD[$i]) eq "PRP") and ($tokensRD[$i] =~/^[I|
i]t$/)) {
            $counttagsRDitPRP++;
   }
```

```
          #count number of "it" tagged PRP but that are actually PNR in
the reference data
if(((($tags[$i]) eq "PRP") and ($tagsRD[$i] eq "PNR") and ($tokens[$i]
=~/^[I|i]t$/)) {
          $counttagsitPRPbutRDPNR++;
    }
}

#precision and recall for "it" as PNR
#count of Nc
for ( my $i = 0 ; $i <= ($#tags) ; $i++ ) {
   if((  $tags[$i] eq $tagsRD[$i]) and ($tokens[$i] =~/^[I|i]t$/) and
($tagsRD[$i] =~/^PNR$/) and ($tokensRD[$i] =~/^[I|i]t$/)) {
          $countitPNR++;
    }

    #count of Na
   if((($tags[$i]) eq "PNR") and ($tokens[$i] =~/^[I|i]t$/) ) {
          $counttagsitPNR++;
   }
#count of Nr
             if((($tagsRD[$i]) eq "PNR") and ($tokensRD[$i] =~/^[I|
i]t$/)) {
            $counttagsRDitPNR++;
   }

          #count number of "it" tagged PNR but that are actually PRP in
the reference data
if((($tags[$i]) eq "PNR") and ($tagsRD[$i] eq "PRP") and ($tokens[$i]
=~/^[I|i]t$/)) {
          $counttagsitPNRbutRDPRP++;
    }

}

############THIS#####################

#precision and recall for "this"
#count of Nc
for ( my $i = 0 ; $i <= ($#tags) ; $i++ ) {
   if((  $tags[$i] eq $tagsRD[$i]) and ($tokens[$i] =~/[T|t]his/) and
($tokensRD[$i] =~/[T|t]his/)) {
          $countthis++;
    }

    #count of Na
   if((((($tags[$i]) eq "DT") or (($tags[$i]) eq "TPRON") or (($tags[$i])
eq "RB")) and ($tokens[$i] =~/[T|t]his/) ) {
             $counttagsthis++;
   }
#count of Nr
             if((((($tagsRD[$i]) eq "DT") or (($tagsRD[$i]) eq "TPRON") or
(($tagsRD[$i]) eq "RB")) and ($tokensRD[$i] =~/[T|t]his/)) {
             $counttagsRDthis++;
   }
```

```
}

#precision and recall for "this" as DT
#count of Nc
for ( my $i = 0 ; $i <= ($#tags) ; $i++ ) {
   if((  $tags[$i] eq $tagsRD[$i]) and ($tokens[$i] =~/[T|t]his/) and
($tagsRD[$i] =~/^DT$/) and ($tokensRD[$i] =~/[T|t]his/)) {
          $countthisDT++;
    }

    #count of Na
   if(($tags[$i]) eq "DT") and ($tokens[$i] =~/[T|t]his/) ) {
          $counttagsthisDT++;
   }
#count of Nr
             if(($tagsRD[$i]) eq "DT") and ($tokensRD[$i] =~/[T|t]his/))
{
            $counttagsRDthisDT++;
   }

           #count number of this tagged DT but that are actually TPRON in
the reference data
if(($tags[$i]) eq "DT") and ($tagsRD[$i] eq "TPRON") and ($tokens[$i]
=~/[T|t]his/)) {
          $counttagsthisDTbutRDTPRON++;
    }
     #count number of this tagged DT but that are actually RB in the
reference data
if(($tags[$i]) eq "DT") and ($tagsRD[$i] eq "RB") and ($tokens[$i] =~/
[T|t]his/)) {
          $counttagsthisDTbutRDRB++;
    }
}

#precision and recall for "this" as TPRON
#count of Nc
for ( my $i = 0 ; $i <= ($#tags) ; $i++ ) {
   if((  $tags[$i] eq $tagsRD[$i]) and ($tokens[$i] =~/[T|t]his/) and
($tagsRD[$i] =~/^TPRON$/) and ($tokensRD[$i] =~/[T|t]his/)) {
          $countthisTPRON++;
    }

    #count of Na
   if(($tags[$i]) eq "TPRON") and ($tokens[$i] =~/[T|t]his/) ) {
          $counttagsthisTPRON++;
   }
#count of Nr
             if(($tagsRD[$i]) eq "TPRON") and ($tokensRD[$i] =~/[T|
t]his/)) {
            $counttagsRDthisTPRON++;
   }

          #count number of this tagged TPRON but that are actually DT in
```

```
the reference data
if(((($tags[$i]) eq "TPRON") and ($tagsRD[$i] eq "DT") and ($tokens[$i]
=~/[T|t]his/)) {
          $counttagsthisTPRONbutRDDT++;
    }
          #count number of this tagged TPRON but that are actually RB in
the reference data
if(((($tags[$i]) eq "TPRON") and ($tagsRD[$i] eq "RB") and ($tokens[$i]
=~/[T|t]his/)) {
          $counttagsthisTPRONbutRDRB++;
    }
}

#precision and recall for "this" as RB
#count of Nc
for ( my $i = 0 ; $i <= ($#tags) ; $i++ ) {
   if((  $tags[$i] eq $tagsRD[$i]) and ($tokens[$i] =~/[T|t]his/) and
($tagsRD[$i] =~/^RB$/) and ($tokensRD[$i] =~/[T|t]his/)) {
          $countthisRB++;
    }

    #count of Na
   if((($tags[$i]) eq "RB") and ($tokens[$i] =~/[T|t]his/) ) {
          $counttagsthisRB++;
   }
#count of Nr
             if((($tagsRD[$i]) eq "RB") and ($tokensRD[$i] =~/[T|t]his/))
{
            $counttagsRDthisRB++;
   }
#count number of this tagged RB but that are actually DT in the reference
data
if((($tags[$i]) eq "RB") and ($tagsRD[$i] eq "DT") and ($tokens[$i] =~/
[T|t]his/)) {
          $counttagsthisRBbutRDDT++;
    }

#count number of this tagged RB but that are actually TPRON in the
reference data
if((($tags[$i]) eq "RB") and ($tagsRD[$i] eq "DT") and ($tokens[$i] =~/
[T|t]his/)) {
          $counttagsthisRBbutRDTPRON++;
    }
}


#precision and recall for "that"
#count of Nc
for ( my $i = 0 ; $i <= ($#tags) ; $i++ ) {
   if((  $tags[$i] eq $tagsRD[$i]) and ($tokens[$i] =~/[T|t]hat/) and
($tokensRD[$i] =~/[T|t]hat/)) {
          $countthat++;
    }

    #count of Na
```

```
   if((((($tags[$i]) eq "DT") or (($tags[$i]) eq "TPRON") or (($tags[$i])
eq "RB") or (($tags[$i]) eq "TCOM") or (($tags[$i]) eq "TREL")) and
($tokens[$i] =~/[T|t]hat/) ) {
            $counttagsthat++;
   }
#count of Nr
            if((((($tagsRD[$i]) eq "DT") or (($tagsRD[$i]) eq "TPRON") or
(($tagsRD[$i]) eq "RB") or (($tagsRD[$i]) eq "TCOM") or (($tagsRD[$i]) eq
"TREL")) and ($tokensRD[$i] =~/[T|t]hat/)) {
            $counttagsRDthat++;
   }

}

#precision and recall for "that" as DT
#count of Nc
for ( my $i = 0 ; $i <= ($#tags) ; $i++ ) {
   if((  $tags[$i] eq $tagsRD[$i]) and ($tokens[$i] =~/[T|t]hat/) and
($tagsRD[$i] =~/^DT$/) and ($tokensRD[$i] =~/[T|t]hat/)) {
          $countthatDT++;
    }

    #count of Na
   if((($tags[$i]) eq "DT") and ($tokens[$i] =~/[T|t]hat/) ) {
          $counttagsthatDT++;
   }
#count of Nr
            if((($tagsRD[$i]) eq "DT") and ($tokensRD[$i] =~/[T|t]hat/))
{
           $counttagsRDthatDT++;
   }
   #count number of that tagged DT but that are actually TPRON in the
reference data
if((($tags[$i]) eq "DT") and ($tagsRD[$i] eq "TPRON") and ($tokens[$i]
=~/[T|t]hat/)) {
          $counttagsthatDTbutRDTPRON++;
    }
    #count number of that tagged DT but that are actually TCOM in the
reference data
    if((($tags[$i]) eq "DT") and ($tagsRD[$i] eq "TCOM") and ($tokens[$i]
=~/[T|t]hat/)) {
          $counttagsthatDTbutRDTCOM++;
    }

    #count number of that tagged DT but that are actually TREL in the
reference data
    if((($tags[$i]) eq "DT") and ($tagsRD[$i] eq "TREL") and ($tokens[$i]
=~/[T|t]hat/)) {
          $counttagsthatDTbutRDTREL++;
    }
   #count number of that tagged DT but that are actually RB in the
reference data
    if((($tags[$i]) eq "DT") and ($tagsRD[$i] eq "RB") and ($tokens[$i]
=~/[T|t]hat/)) {
          $counttagsthatDTbutRDRB++;
```

```
    }
}

################################THAT#########################
#precision and recall for "that" as TPRON
#count of Nc
for ( my $i = 0 ; $i <= ($#tags) ; $i++ ) {
   if((  $tags[$i] eq $tagsRD[$i]) and ($tokens[$i] =~/[T|t]hat/) and
($tagsRD[$i] =~/^TPRON$/) and ($tokensRD[$i] =~/[T|t]hat/)) {
         $countthatTPRON++;
    }

    #count of Na
   if((($tags[$i]) eq "TPRON") and ($tokens[$i] =~/[T|t]hat/) ) {
         $counttagsthatTPRON++;
   }
#count of Nr
            if((($tagsRD[$i]) eq "TPRON") and ($tokensRD[$i] =~/[T|
t]hat/)) {
           $counttagsRDthatTPRON++;
   }
      #count number of that tagged TPRON but that are actually DT in the
reference data
    if((($tags[$i]) eq "TPRON") and ($tagsRD[$i] eq "DT") and
($tokens[$i] =~/[T|t]hat/)) {
         $counttagsthatTPRONbutRDDT++;
}
      #count number of that tagged TPRON but that are actually TCOM in
the reference data
    if((($tags[$i]) eq "TPRON") and ($tagsRD[$i] eq "TCOM") and
($tokens[$i] =~/[T|t]hat/)) {
         $counttagsthatTPRONbutRDTCOM++;
}
      #count number of that tagged TPRON but that are actually TREL in
the reference data
    if((($tags[$i]) eq "TPRON") and ($tagsRD[$i] eq "TREL") and
($tokens[$i] =~/[T|t]hat/)) {
         $counttagsthatTPRONbutRDTREL++;
}
      #count number of that tagged TPRON but that are actually RB in the
reference data
    if((($tags[$i]) eq "TPRON") and ($tagsRD[$i] eq "RB") and
($tokens[$i] =~/[T|t]hat/)) {
         $counttagsthatTPRONbutRDRB++;
}
}

#precision and recall for "that" as RB
#count of Nc
for ( my $i = 0 ; $i <= ($#tags) ; $i++ ) {
   if((  $tags[$i] eq $tagsRD[$i]) and ($tokens[$i] =~/[T|t]hat/) and
($tagsRD[$i] =~/^RB$/) and ($tokensRD[$i] =~/[T|t]hat/)) {
         $countthatRB++;
    }
```

```
    #count of Na
   if(((($tags[$i]) eq "RB") and ($tokens[$i] =~/[T|t]hat/) ) {
         $counttagsthatRB++;
   }
#count of Nr
            if(((($tagsRD[$i]) eq "RB") and ($tokensRD[$i] =~/[T|t]hat/))
{
           $counttagsRDthatRB++;
   }
 #count number of that tagged RB but that are actually DT in the
reference data
    if(((($tags[$i]) eq "RB") and ($tagsRD[$i] eq "DT") and ($tokens[$i]
=~/[T|t]hat/)) {
         $counttagsthatRBbutRDDT++;
}
 #count number of that tagged RB but that are actually TPRON in the
reference data
    if(((($tags[$i]) eq "RB") and ($tagsRD[$i] eq "TPRON") and
($tokens[$i] =~/[T|t]hat/)) {
         $counttagsthatRBbutRDTPRON++;
}
 #count number of that tagged RB but that are actually TCOM in the
reference data
    if(((($tags[$i]) eq "RB") and ($tagsRD[$i] eq "TCOM") and ($tokens[$i]
=~/[T|t]hat/)) {
         $counttagsthatRBbutRDTCOM++;
}


#count number of that tagged RB but that are actually TREL in the
reference data
    if(((($tags[$i]) eq "RB") and ($tagsRD[$i] eq "TREL") and ($tokens[$i]
=~/[T|t]hat/)) {
         $counttagsthatRBbutRDTREL++;
}

}


#precision and recall for "that" as TCOM
#count of Nc
for ( my $i = 0 ; $i <= ($#tags) ; $i++ ) {
   if((  $tags[$i] eq $tagsRD[$i]) and ($tokens[$i] =~/[T|t]hat/) and
($tagsRD[$i] =~/^TCOM$/) and ($tokensRD[$i] =~/[T|t]hat/)) {
         $countthatTCOM++;
    }

    #count of Na
   if(((($tags[$i]) eq "TCOM") and ($tokens[$i] =~/[T|t]hat/) ) {
         $counttagsthatTCOM++;
   }
#count of Nr
            if(((($tagsRD[$i]) eq "TCOM") and ($tokensRD[$i] =~/[T|
t]hat/)) {
           $counttagsRDthatTCOM++;
```

```
   }
         #count number of that tagged TCOM but that are actually DT in
the reference data
    if((($tags[$i]) eq "TCOM") and ($tagsRD[$i] eq "DT") and ($tokens[$i]
=~/[T|t]hat/)) {
         $counttagsthatTCOMbutRDDT++;
}
         #count number of that tagged TCOM but that are actually TPRON in
the reference data
    if((($tags[$i]) eq "TCOM") and ($tagsRD[$i] eq "TPRON") and
($tokens[$i] =~/[T|t]hat/)) {
         $counttagsthatTCOMbutRDTPRON++;
}
     #count number of that tagged TCOM but that are actually TREL in the
reference data
    if((($tags[$i]) eq "TCOM") and ($tagsRD[$i] eq "TREL") and
($tokens[$i] =~/[T|t]hat/)) {
         $counttagsthatTCOMbutRDTREL++;
}
     #count number of that tagged TCOM but that are actually RB in the
reference data
    if((($tags[$i]) eq "TCOM") and ($tagsRD[$i] eq "RB") and ($tokens[$i]
=~/[T|t]hat/)) {
         $counttagsthatTCOMbutRDRB++;
}
}

#precision and recall for "that" as TREL
#count of Nc
for ( my $i = 0 ; $i <= ($#tags) ; $i++ ) {
   if((  $tags[$i] eq $tagsRD[$i]) and ($tokens[$i] =~/[T|t]hat/) and
($tagsRD[$i] =~/^TREL$/) and ($tokensRD[$i] =~/[T|t]hat/)) {
         $countthatTREL++;
    }

    #count of Na
   if((($tags[$i]) eq "TREL") and ($tokens[$i] =~/[T|t]hat/) ) {
         $counttagsthatTREL++;
   }
#count of Nr
            if((($tagsRD[$i]) eq "TREL") and ($tokensRD[$i] =~/[T|
t]hat/)) {
           $counttagsRDthatTREL++;
   }
        #count number of that tagged TREL but that are actually DT in the
reference data
    if((($tags[$i]) eq "TREL") and ($tagsRD[$i] eq "DT") and ($tokens[$i]
=~/[T|t]hat/)) {
         $counttagsthatTRELbutRDDT++;
}
        #count number of that tagged TREL but that are actually TPRON in
the reference data
    if((($tags[$i]) eq "TREL") and ($tagsRD[$i] eq "TPRON") and
($tokens[$i] =~/[T|t]hat/)) {
         $counttagsthatTRELbutRDTPRON++;
```

```
}
        #count number of that tagged TREL but that are actually TCOM in
the reference data
    if(($tags[$i]) eq "TREL") and ($tagsRD[$i] eq "TCOM") and
($tokens[$i] =~/[T|t]hat/)) {
         $counttagsthatTRELbutRDTCOM++;
}
        #count number of that tagged TREL but that are actually RB in the
reference data
    if(($tags[$i]) eq "TREL") and ($tagsRD[$i] eq "RB") and ($tokens[$i]
=~/[T|t]hat/)) {
         $counttagsthatTRELbutRDTRB++;
}

}
########################################COMPUTE PRECISION AND
RECALL############################
# recall = Nc/Nr
#precision = Nc/Na
#calculate precision and recall based on variables for "it"
#it
my $precisionit=$countit/$counttagsit;
my $recallit=$countit/$counttagsRDit;
my $fscoreit=2*$precisionit*$recallit/($precisionit+$recallit);

#it PRP
my $precisionitPRP=$countitPRP/$counttagsitPRP;
my $recallitPRP=$countitPRP/$counttagsRDitPRP;
my $fscoreitPRP=2*$precisionitPRP*$recallitPRP/($precisionitPRP+
$recallitPRP);

#it PNR
my $precisionitPNR=$countitPNR/$counttagsitPNR;
my $recallitPNR=$countitPNR/$counttagsRDitPNR;
my $fscoreitPNR=2*$precisionitPNR*$recallitPNR/($precisionitPNR+
$recallitPNR);


#calculate precision and recall based on variables for "this"
#this
my $precisionthis=$countthis/$counttagsthis;
my $recallthis=$countthis/$counttagsRDthis;
my $fscorethis=2*$precisionthis*$recallthis/($precisionthis+$recallthis);

#this DT
my $precisionthisDT=$countthisDT/$counttagsthisDT;
my $recallthisDT=$countthisDT/$counttagsRDthisDT;
my $fscorethisDT=2*$precisionthisDT*$recallthisDT/($precisionthisDT+
$recallthisDT);

#this TPRON
my $precisionthisTPRON=$countthisTPRON/$counttagsthisTPRON;
my $recallthisTPRON=$countthisTPRON/$counttagsRDthisTPRON;
my $fscorethisTPRON=2*$precisionthisTPRON*$recallthisTPRON/
($precisionthisTPRON+$recallthisTPRON);
```

```
#this RB (if any reactivate)
#my $precisionthisRB=$countthisRB/$counttagsthisRB;
#my $recallthisRB=$countthisRB/$counttagsRDthisRB;
#my $fscorethisRB=2*$precisionthisRB*$recallthisRB/($precisionthisRB+
$recallthisRB);



#calculate precision and recall based on variables for that
#that
 my $precisionthat=$countthat/$counttagsthat;
my $recallthat=$countthat/$counttagsRDthat;
 my $fscorethat=2*$precisionthat*$recallthat/($precisionthat+
$recallthat);

#that DT
my $precisionthatDT=$countthatDT/$counttagsthatDT;
my $recallthatDT=$countthatDT/$counttagsRDthatDT;
my $fscorethatDT=2*$precisionthatDT*$recallthatDT/($precisionthatDT+
$recallthatDT);


#that TPRON
 my $precisionthatTPRON=$countthatTPRON/$counttagsthatTPRON;
my $recallthatTPRON=$countthatTPRON/$counttagsRDthatTPRON;
my $fscorethatTPRON=2*$precisionthatTPRON*$recallthatTPRON/
($precisionthatTPRON+$recallthatTPRON);

#that RB (activate lines if any in test file)
#my $precisionthatRB=$countthatRB/$counttagsthatRB;
#my $recallthatRB=$countthatRB/$counttagsRDthatRB;
#my $fscorethatRB=2*$precisionthatRB*$recallthatRB/($precisionthatRB+
$recallthatRB);

#that TCOM
my $precisionthatTCOM=$countthatTCOM/$counttagsthatTCOM;
my $recallthatTCOM=$countthatTCOM/$counttagsRDthatTCOM;
my $fscorethatTCOM=2*$precisionthatTCOM*$recallthatTCOM/
($precisionthatTCOM+$recallthatTCOM);

#that TREL
my $precisionthatTREL=$countthatTREL/$counttagsthatTREL;
my $recallthatTREL=$countthatTREL/$counttagsRDthatTREL;
my $fscorethatTREL=2*$precisionthatTREL*$recallthatTREL/
($precisionthatTREL+$recallthatTREL);

#print messages

print "----------------------------IT-------------------------------\n";
print "Nc: There are ", $countit, " correctly assigned tags to 'it'
forms. \n";
print "Na: There are ", $counttagsit, " tags automatically assigned to
'it' forms in the data file. \n";
print "Nr: There are ",$counttagsRDit, " correct tags of 'it' forms
```

```
according to the reference file. \n";
print "Recall for occurrences of 'it' = ", $recallit, " \n";
print "Precision for occurrences of 'it' = ", $precisionit, "\n";
print "F-score for 'it' is ", $fscoreit, " \n \n";

print "----------------------------IT Pronoun
Personal---------------------------------\n";
print "Nc: There are ", $countitPRP, " correctly assigned tags to 'it
PRP' forms. \n";
print "There are ",$counttagsitPRPbutRDPNR, " PRP tags assigned to 'it'
when they should be PNR. \n";
print "Na: There are ", $counttagsitPRP, " tags automatically assigned to
'it PRP' forms in the data file. \n";
print "Nr: There are ",$counttagsRDitPRP, " correct tags of 'it PRP'
forms according to the reference file. \n";
print "Recall for occurrences of 'it PRP' = ", $recallitPRP, " \n";
print "Precision for occurrences of 'it PRP' = ", $precisionitPRP, "\n";
print "F-score it PRP is ", $fscoreitPRP, " \n \n";

print "----------------------------IT Pronoun Non
Referential-------------------------------\n";
print "Nc: There are ", $countitPNR, " correctly assigned tags to 'it
PNR' forms. \n";
print "There are ",$counttagsitPNRbutRDPRP, " PNR tags assigned to 'it'
when they should be PRP. \n";
print "Na: There are ", $counttagsitPNR, " tags automatically assigned to
'it' PNR forms in the data file. \n";
print "Nr: There are ",$counttagsRDitPNR, " correct tags of 'it' PNR
forms according to the reference file. \n";
print "Recall for occurrences of 'it PNR' = ", $recallitPNR, " \n";
print "Precision for occurrences of 'it PNR' = ", $precisionitPNR, "\n";
print "F-score it PNR is ", $fscoreitPNR, " \n \n";

print "------------------------Confusion matrix for
'it'---------------------------------- \n";
print " \tit\t|PRP\t|PNR\t|\n";
print " tagged PRP\t|", $countitPRP, "\t|",$counttagsitPRPbutRDPNR,
"\t|\n";
print " tagged PNR\t|",$counttagsitPNRbutRDPRP, "\t|", $countitPNR,
"\t|\n";


print "----------------------------
THIS---------------------------------\n";
print "Nc: There are ", $countthis, " correctly assigned tags to 'this'
forms. \n";
print "Na: There are ", $counttagsthis, " tags automatically assigned to
'this' forms in the data file. \n";
print "Nr: There are ",$counttagsRDthis, " correct tags of 'this' forms
according to the reference file. \n";
print "Recall for occurrences of 'this' = ", $recallthis, " \n";
print "Precision for occurrences of 'this' = ", $precisionthis, "\n";
print "F-score for this is ", $fscorethis, " \n \n";

print "Nc: There are ", $countthisDT, " correctly assigned tags to 'this
```

```
DT' forms. \n";
print "There are ",$counttagsthisDTbutRDTPRON, " DT tags assigned to
'this' when they should be TPRON. \n";
print "There are ",$counttagsthisDTbutRDRB, " DT tags assigned to 'this'
when they should be RB. \n";
print "Na: There are ", $counttagsthisDT, " tags automatically assigned
to 'this DT' forms in the data file. \n";
print "Nr: There are ",$counttagsRDthisDT, " correct tags of 'this DT'
forms according to the reference file. \n";
print "Recall for occurrences of 'this DT' = ", $recallthisDT, " \n";
print "Precision for occurrences of 'this DT' = ", $precisionthisDT,
"\n";
print "F-score this DT is ", $fscorethisDT, " \n \n";

print "Nc: There are ", $countthisTPRON, " correctly assigned tags to
'this TPRON' forms. \n";
print "There are ",$counttagsthisTPRONbutRDDT, " TPRON tags assigned to
'this' when they should be DT. \n";
print "Na: There are ", $counttagsthisTPRON, " tags automatically
assigned to 'this TPRON' forms in the data file. \n";
print "Nr: There are ",$counttagsRDthisTPRON, " correct tags of 'this
TPRON' forms according to the reference file. \n";
print "Recall for occurrences of 'this TPRON' = ", $recallthisTPRON, "
\n";
print "Precision for occurrences of 'this TPRON' = ",
$precisionthisTPRON, "\n";
print "F-score this TPRON is ", $fscorethisTPRON, " \n \n";

print "Nc: There are ", $countthisRB, " correctly assigned tags to 'this
RB' forms. \n";
print "Na: There are ", $counttagsthisRB, " tags automatically assigned
to 'this RB' forms in the data file. \n";
print "Nr: There are ",$counttagsRDthisRB, " correct tags of 'this RB'
forms according to the reference file. \n";
#print "Recall for occurrences of 'this RB' = ", $recallthisRB, " \n";
#print "Precision for occurrences of 'this RB' = ", $precisionthisRB,
"\n";
#print "F-score this RB is ", $fscorethisRB, " \n\n";

print "Confusion matrix for this \n";
print " \tthis\t|DT\t|TPRON\t|RB\t|\n";
print " tagged DT\t|", $countthisDT, "\t|",$counttagsthisDTbutRDTPRON,
"\t|",$counttagsthisDTbutRDRB, "\t|\n";
print " tagged TPRON\t|",$counttagsthisTPRONbutRDDT, "\t|",
$countthisTPRON, "\t|",$counttagsthisTPRONbutRDRB,"\t|\n";
print " tagged RB\t|",$counttagsthisRBbutRDDT,"\t|",
$counttagsthisRBbutRDDT,"\t|",$countthisRB,"\t|\n";

print "------------------------------THAT----------------------------\n";
print "Nc: There are ", $countthat, " correctly assigned tags to 'that'
forms. \n";
print "Na: There are ", $counttagsthat, " tags automatically assigned to
'that' forms in the data file. \n";
print "Nr: There are ",$counttagsRDthat, " correct tags of 'that' forms
according to the reference file. \n";
```

```
print "Recall for occurrences of 'that' = ", $recallthat, " \n";
print "Precision for occurrences of 'that' = ", $precisionthat, "\n";
print "F-score for that is ", $fscorethat, " \n \n";


print "Nc: There are ", $countthatDT, " correctly assigned tags to 'that
DT' forms. \n";
print "Na: There are ", $counttagsthatDT, " tags automatically assigned
to 'that DT' forms in the data file. \n";
print "Nr: There are ",$counttagsRDthatDT, " correct tags of 'that DT'
forms according to the reference file. \n";
print "Recall for occurrences of 'that DT' = ", $recallthatDT, " \n";
print "Precision for occurrences of 'that DT' = ", $precisionthatDT,
"\n";
print "F-score for that DT is ", $fscorethatDT, " \n \n";

print "Nc: There are ", $countthatTPRON, " correctly assigned tags to
'that TPRON' forms. \n";
print "Na: There are ", $counttagsthatTPRON, " tags automatically
assigned to 'that TPRON' forms in the data file. \n";
print "Nr: There are ",$counttagsRDthatTPRON, " correct tags of 'that
TPRON' forms according to the reference file. \n";
print "Recall for occurrences of 'that TPRON' = ", $recallthatTPRON, "
\n";
print "Precision for occurrences of 'that TPRON' = ",
$precisionthatTPRON, "\n";
print "F-score for that TPRON is ", $fscorethatTPRON, " \n \n";

print "Nc: There are ", $countthatRB, " correctly assigned tags to 'that
RB' forms. \n";
print "Na: There are ", $counttagsthatRB, " tags automatically assigned
to 'that RB' forms in the data file. \n";
print "Nr: There are ",$counttagsRDthatRB, " correct tags of 'that RB'
forms according to the reference file. \n";
#print "Recall for occurrences of 'that RB' = ", $recallthatRB, " \n";
#print "Precision for occurrences of 'that RB' = ", $precisionthatRB,
"\n";
#print "F-score for that RB is ", $fscorethatRB, " \n \n";

print "Nc: There are ", $countthatTCOM, " correctly assigned tags to
'that TCOM' forms. \n";
print "Na: There are ", $counttagsthatTCOM, " tags automatically assigned
to 'that TCOM' forms in the data file. \n";
print "Nr: There are ",$counttagsRDthatTCOM, " correct tags of 'that
TCOM' forms according to the reference file. \n";
print "Recall for occurrences of 'that TCOM' = ", $recallthatTCOM, " \n";
print "Precision for occurrences of 'that TCOM' = ", $precisionthatTCOM,
"\n";
print "F-score for that TCOM is ", $fscorethatTCOM, " \n \n";

print "Nc: There are ", $countthatTREL, " correctly assigned tags to
'that TREL' forms. \n";
print "Na: There are ", $counttagsthatTREL, " tags automatically assigned
to 'that TREL' forms in the data file. \n";
print "Nr: There are ",$counttagsRDthatTREL, " correct tags of 'that
```

```
TREL' forms according to the reference file. \n";
print "Recall for occurrences of 'that TREL' = ", $recallthatTREL, " \n";
print "Precision for occurrences of 'that TREL' = ", $precisionthatTREL,
"\n";
print "F-score for that TREL is ", $fscorethatTREL, " \n \n";


print "Confusion matrix for that \n";
print " \tthat\t|DT\t|TPRON\t|TCOM\t|TREL\t|RB\t\n";
print " tagged DT\t|", $countthatDT, "\t|",$counttagsthatDTbutRDTPRON,
"\t|",$counttagsthatDTbutRDTCOM, "\t|",$counttagsthatDTbutRDTREL, "\t|",
$counttagsthatDTbutRDRB, "\t\n";
print " tagged TPRON\t|",$counttagsthatTPRONbutRDDT, "\t|",
$countthatTPRON, "\t|",$counttagsthatTPRONbutRDTCOM,"\t|",
$counttagsthatTPRONbutRDTREL,"\t|",$counttagsthatTPRONbutRDRB,"\t\n";
print " tagged TCOM\t|",$counttagsthatTCOMbutRDDT, "\t|",
$counttagsthatTCOMbutRDTPRON, "\t|",$countthatTCOM,"\t|",
$counttagsthatTCOMbutRDTREL,"\t|",$counttagsthatTCOMbutRDRB,"\t\n";
print " tagged TREL\t|",$counttagsthatTRELbutRDDT, "\t|",
$counttagsthatTRELbutRDTPRON, "\t|",$counttagsthatTRELbutRDTCOM,"\t|",
$countthatTREL,"\t|",$counttagsthatTRELbutRDTRB,"\t\n";
print " tagged RB\t|",$counttagsthatRBbutRDDT, "\t|",
$counttagsthatRBbutRDTPRON, "\t|",$counttagsthatRBbutRDTCOM,"\t|",
$counttagsthatRBbutRDTREL,"\t|",$countthatRB,"\t\n";
```

Annex H - NITE-XML metadata file

This is the example of the metadata XML file to be opened in NITE NXT Search tool to perform keyword searches (Kilgour, Carletta, and Evert 2003).

File: single-sentence-metadata.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- <!DOCTYPE corpus SYSTEM "meta-standoff.dtd"> -->

<!-- NXT METADATA EXAMPLE FOR A STANDOFF CORPUS JEAN CARLETTA AND JONATHAN KILGOUR
     ADAPTED FOR STEFAN EVERT'S EXAMPLE CORPUS 3/9/2
-->
<corpus description="Test Corpus" id="single-sentence" links="ltxml1" type="standoff">
    <!-- GENERIC CORPUS INFORMATION -->
    <reserved-attributes>
        <identifier name="nite:id"/>
        <starttime name="nite:start"/>
        <endtime name="nite:end"/>
        <agentname name="who"/>
    </reserved-attributes>
    <reserved-elements>
        <pointername name="nite:pointer"/>
        <child name="nite:child"/>
        <stream name="nite:stream"/>
    </reserved-elements>
    <observation-variables>
        <observation-variable name="eye-contact" type="enumerated">
            <value>no eye</value>
            <value>eye</value>
        </observation-variable>
        <observation-variable name="familiarity" type="enumerated">
            <value>familiar</value>
            <value>non-familiar</value>
        </observation-variable>
    </observation-variables>


    <!-- ONTOLOGIES -->
    <!-- ontologies are static hierarchies e.g. the gesture ontology in this
         example -->
    <ontologies path="../xml/SingleSentence">
        <ontology description="gesture ontology" name="gtypes" filename="" element-name="gtype"
attribute-name="type"/>
    </ontologies>

    <!-- CODINGS -->
    <codings path="../xml/SingleSentence">
        <interaction-codings>
            <coding-file name="gestures-right">
                <structural-layer name="gesture-layer" draws-children-from="phase-layer">
                    <code name="gest">
                        <attribute name="target" value-type="string"/>
                        <pointer number="1" role="TYPE" target="gtypes"/>
                    </code>
                </structural-layer>
                <time-aligned-layer name="phase-layer">
                    <code name="phase">
                        <attribute name="type" value-type="string"/>
                    </code>
                </time-aligned-layer>
            </coding-file>
            <coding-file name="gestures-left">
                <structural-layer name="gesture-left-layer" draws-children-from="phase-left-layer">
                    <code name="lgest">
                        <attribute name="target" value-type="string"/>
                        <pointer number="1" role="TYPE" target="gtypes"/>
                    </code>
                </structural-layer>
                <time-aligned-layer name="phase-left-layer">
                    <code name="lphase">
                        <attribute name="type" value-type="string"/>
                    </code>
                </time-aligned-layer>
            </coding-file>
            <coding-file name="prosody">
                <structural-layer name="prosody-layer" draws-children-from="words-layer">
                    <code name="accent">
                        <attribute name="tobi" value-type="string"/>
```

```
                </code>
            </structural-layer>
        </coding-file>

            <coding-file name="turns">
                        <structural-layer name="turn-layer" draws-children-from="syntax-layer">
                          <code name="turn"/>
                        </structural-layer>
            </coding-file>

        <coding-file name="syntax">
            <structural-layer name="syntax-layer" draws-children-from="phrase-layer">
                <code name="s"/>
            </structural-layer>
            <structural-layer name="phrase-layer" recursive-draws-children-from="words-layer">
                <code name="vp">
                    <attribute name="hlem" value-type="string"/>
                </code>
                <code name="np">
                    <attribute name="hlem" value-type="string"/>
                </code>
                <code name="pp">
                    <attribute name="hlem" value-type="string"/>
                    <attribute name="prep" value-type="string"/>
                </code>
            </structural-layer>
        </coding-file>
        <coding-file name="words">
            <time-aligned-layer name="words-layer">
                <code name="word">
                    <attribute name="orth" value-type="string"/>
                    <attribute name="pos" value-type="enumerated">
                        <value>CC</value>
                        <value>CD</value>
                        <value>DT</value>
                        <value>EX</value>
                        <value>FW</value>
                        <value>IN</value>
                        <value>JJ</value>
                        <value>JJR</value>
                        <value>JJS</value>
                        <value>LS</value>
                        <value>MD</value>
                        <value>NN</value>
                        <value>NNS</value>
                        <value>NNP</value>
                        <value>NNPS</value>
                        <value>PDT</value>
                        <value>POS</value>
                        <value>PRP</value>
                        <value>PRP$</value>
                        <value>RB</value>
                        <value>RBR</value>
                        <value>RBS</value>
                        <value>RP</value>
                        <value>TO</value>
                        <value>UH</value>
                        <value>VB</value>
                        <value>VBD</value>
                        <value>VBG</value>
                        <value>VBN</value>
                        <value>VBP</value>
                        <value>VBZ</value>
                        <value>WDT</value>
                        <value>WP</value>
                        <value>WP$</value>
                        <value>WRB</value>
                    </attribute>
                    <pointer number="1" role="ANTECEDENT" target="phrase-layer"/>
                </code>
            </time-aligned-layer>
        </coding-file>
    </interaction-codings>
</codings>

<callable-programs>
  <callable-program name="SimpleSaveExample" description="load and save example">
    <required-argument name="corpus" type="corpus"/>
  </callable-program>
```

```
    </callable-programs>

    <observations>
        <observation name="o1"/>
    </observations>
</corpus>
```

Annex I - PERL script for NXT word coding conversion

PERL script used to convert three-column-layout format of corpus files into a XML structure. This script creates the coding file dedicated to word units. It creates three attributes for word elements: orthography, PoS for all forms and position for the *it*, *this* and *that* forms.

```
# Copyright Thomas Gaillat – Université de Paris-Diderot
# script adapted from Hathout and Tanguy's script found in "Perl pour les
linguistes" http://perl.linguistes.free.fr/ script 9.5
# split utterances at . or ,. Check line 56
#construction of word layer for NXT
use strict;
use locale;
use XML::Writer;

#multi-file processing in one folder

if ( $#ARGV != 0 ) {
    die "Usage : ", $0, " enter directory name please\n";
}

my $repertoire = $ARGV[0];


opendir ( REP, $repertoire ) or
    die "Impossible d'ouvrir ", $repertoire, " : ", $!, "\n";
my @fichiers = readdir( REP );
closedir ( REP );

my @fichierstt = grep (/\.tt$/, @fichiers);

foreach my $fichier ( sort @fichierstt ) {
          my $r = $fichier ;
          $r =~ s/^(.*)\.tt$/$1/;
          print STDERR $r, "\n";

          open ( ENTREE, "<", $repertoire."/".$fichier) or
                    warn "Erreur d'ouverture de ",$fichier, " :" ,$!,
"\n";
          open(SORTIE, ">", $repertoire."/".$r.".words.xml") or die
"impossible d'ouvrir ", $r;

#throw each token tag pair in arrays.
my (@tokens, @tags, @context, @discourse);

while (my $line = <ENTREE>) {
chomp $line;
        my @array = split ( /\t/, $line );
          push ( @tokens, $array[0] );
          push ( @tags, $array[1] );
          push ( @context, $array[2] );
        push ( @discourse, $array[3] );
}
```

```
# traitement du fichier

my $ut_id = 0;



my $scripteur = new XML::Writer ( DATA_MODE=>1, DATA_INDENT=>1,
OUTPUT=>\*SORTIE);

#$scripteur->startTag( "nite:stream", "nite:id"=>"word_str_1",
"xmlns:nite"=>"http://nite.sourceforge.net/");
$scripteur->startTag( "txt", "nite:id"=>$1);

for (my $i=0; $i <= $#tokens; $i++){

#capture any token except speakers' tags.
if (( $tokens[$i]=~ /^(.*)$/ ) and ( $tokens[$i] !~/<A>|<B>|<\/A>|
<\/B>/ )) {
          #under such a condition, capture the deictic forms and pronoun
it
          if (($tokens[$i] =~/^[T|t]his$/ ) or ($tokens[$i] =~/^[T|
t]hat$/) or ($tokens[$i] =~/^[I|i]t$/) or ($tokens[$i] =~/^[T|t]hese$/)
or /^[T|t]hose$/) {
                      my $position = "NA";
                      # Check for interrogative form & check if it is
followed by a modal or a verb or by an adverb or repeated proform and a
verb or modal.
                   if ($tags[$i] =~ /TPRON|PRP/) {
                            if (
                                     ($tags[$i-1] =~ /MD/)
                                     or (($tags[$i-1] =~ /^do$|^does$|
^am$|^are$|^is$|^was$|^were$/) and (
                                                                ($tags[$i+1]
=~ /^JJ$|^VB$|^VBG$/)
                                                                or (($tags[$i+1]
=~ /^RB$/) and ($tags[$i+2] =~ /^|^VB$|^VBG$/)))
                                     or ($tags[$i+1] =~ /V.*|MD|NNS|
POS/)
                                     or (($tags[$i+1] =~ /RB|TPRON|PRP/)
and ($tags[$i+2] =~ /V.*|MD|NNS|POS/)
                                                          )
                                     )
                            ) {
                            $position = "NOMI";
                            }
                            else {
                            $position="OBLI";
                            }
                   }
                   #Check the form is a DT
                   elsif (($tags[$i] =~ /DT/) and (($tags[$i+1] =~
/NN.*/)

          or (($tags[$i+1] =~ /JJ|DT/) and ($tags[$i+2] =~ /NN.*/)))) {
                            if (
```

```
                                        ($tags[$i-1] =~ /MD/)
                                        or (($tags[$i-1] =~ /^do$|^does$|
^am$|^are$|^is$|^was$|^were$/) and (
                                                                ($tags[$i+1]
=~ /^JJ$|^VB$|^VBG$/)
                                                                or (($tags[$i+2]
=~ /^RB$/) and ($tags[$i+3] =~ /^|^VB$|^VBG$/)))
                                        or ($tags[$i+2] =~ /V.*|MD|NNS|
POS/)
                                        or (($tags[$i+2] =~ /RB|TPRON|PRP/)
and ($tags[$i+3] =~ /V.*|MD|NNS|POS/)
                                                        )
                                        )
                                ) {
                                $position = "NOMI";
                                }
                                else {
                                $position="OBLI";
                                }
                        }

                        $scripteur->startTag( "word", "nite:id"=>$r."-".$i ,
"orth"=>$tokens[$i], "pos"=>$tags[$i], "position"=>$position);
                        $scripteur->dataElement( "nite:child", "",
"href"=>"o1.context.xml#id(".$r."-ctxt-".$i.")");
                        $scripteur->endTag("word");
                }

                else {
                        $scripteur->startTag( "word", "nite:id"=>$r."-".$i ,
"orth"=>$tokens[$i], "pos"=>$tags[$i]);
                        $scripteur->dataElement( "nite:child", "",
"href"=>"o1.context.xml#id(".$r."-ctxt-".$i.")");
                        $scripteur->endTag("word");
                        }
}

if ($tokens[$i] =~/<A>|<B>/) {
                                $ut_id++;
                        $scripteur ->startTag ("speaker", "nite:id"=>$r."ut".
$ut_id, "agent"=>$tokens[$i]);
                        }




if ( $tokens[$i] =~/<\/A>|<\/B>/ ){
                        $scripteur->endTag( "speaker" );
                        }

                }
$scripteur->endTag("txt" );
#$scripteur->endTag( "nite:stream");
close ( ENTREE );
                close ( SORTIE );}
```

Annex J - PERL script for NXT context coding conversion

PERL script used to convert the context layer of the TreeTagger formatted files into the context coding file of the NITE XML format.

```
#Author Gaillat Thomas. University of Paris Diderot
#Inspired from Tangy and Hathout http://perl.linguistes.free.fr/
#usage perl TT2NITE.pl directory
# directory is at the same level as the perl programme. It includes texts
that have been tokenized an tagged. One token and POS tag per line.
#Construction of context layer for NITE NXT
use strict;
use locale;
use XML::Writer;


#multi-file processing in one folder

if ( $#ARGV != 0 ) {
    die "Usage : ", $0, " enter directory name please\n";
}

my $repertoire = $ARGV[0];


opendir ( REP, $repertoire ) or
    die "Impossible d'ouvrir ", $repertoire, " : ", $!, "\n";
my @fichiers = readdir( REP );
closedir ( REP );

my @fichierstt = grep (/\.tt$/, @fichiers);

foreach my $fichier ( sort @fichierstt ) {
          my $r = $fichier ;
          $r =~ s/^(.*)\.tt$/$1/;
          print STDERR $r, "\n";

          open ( ENTREE, "<", $repertoire."/".$fichier) or
                    warn "Erreur d'ouverture de ",$fichier, " :" ,$!,
"\n";
          open(SORTIE, ">", $repertoire."/".$r.".context.xml") or die
"impossible d'ouvrir ", $r;

#throw each token tag pair in arrays.
my (@tokens, @tags, @context, @discourse);

while (my $line = <ENTREE>) {
chomp $line;
        my @array = split ( /\t/, $line );
          push ( @tokens, $array[0] );
          push ( @tags, $array[1] );
          push ( @context, $array[2] );
        push ( @discourse, $array[3] );
}

# traitement du fichier XML writer (named scripteur in French) applies
tags to array elements
```

```
my $id = 0;
my $ut_id = 0;

my $scripteur = new XML::Writer ( DATA_MODE=>1, DATA_INDENT=>1,
OUTPUT=>\*SORTIE);
#$scripteur->xmlDecl("UTF-8");

#$scripteur->startTag( "nite:stream", "nite:id"=>"context_str_1",
"xmlns:nite"=>"http://nite.sourceforge.net/");

$scripteur->startTag( "txt", "nite:id"=>$1);


for (my $i=0; $i <= $#tokens; $i++){

          if ( $context[$i]=~ /^(.*)$/ ){
          $scripteur->startTag( "context", "nite:id"=>$r."-ctxt-".$i,
"context"=>$context[$i] );

          $scripteur->endTag( "context");
                    }
          }
$scripteur->endTag( "txt" );
#$scripteur->endTag( "nite:stream");
close ( ENTREE );
        close ( SORTIE );}
```

Annex K - PERL script for feature sequencing for R

PERL script used to convert 3-column format of corpus files into sequences of features. This script is used to prepare the data for R by finding all occurrences of *it, this* and *that* in the corpus files and by placing them in tables whose lines are made of the forms and the values of specific features such as the 3, 2 and 1gram PoS, the 3-, 2- and 1-gram words, the position and the context.

```
#Author Gaillat Thomas.University of Paris Diderot
#Inspired from Tangy and Hathout http://perl.linguistes.free.fr/
#usage perl TT2seq_feat_3grams.pl directory
# directory is at the same level as the perl programme. It includes texts
that have been tokenized an tagged. One token and POS tag per line.
#_3grams_context_position
use strict;
use locale;

#multi-file processing in one folder

if ( $#ARGV != 0 ) {
    die "Usage : ", $0, " enter directory name please\n";
}

my $repertoire = $ARGV[0];


opendir ( REP, $repertoire ) or
    die "Impossible d'ouvrir ", $repertoire, " : ", $!, "\n";
my @fichiers = readdir( REP );
closedir ( REP );

my @fichierstt = grep (/\.tt$/, @fichiers);

foreach my $fichier ( sort @fichierstt ) {
         my $r = $fichier ;
         $r =~ s/^(.*)\.tt$/$1.seq4R/;
         print STDERR $r, "\n";

         open ( ENTREE, "<", $repertoire."/".$fichier) or
                  warn "Erreur d'ouverture de ",$fichier, " :" ,$!,
"\n";
         open(SORTIE, ">", $repertoire."/".$r) or die "impossible
d'ouvrir ", $r;

#throw each token tag pair in arrays.
my (@tokens, @tags, @context, @discourse);
#print SORTIE  "DIDID","\t","TOKENS","\t" ,"TAGS","\t",
"TOKENS3BEFORE","\t", "TAGS3BEFORE","\t", "TOKENS2BEFORE","\t",
"TAGS2BEFORE","\t", "TOKENS1BEFORE","\t", "TAGS1BEFORE","\t",
"TOKENS1AFTER","\t", "TAGS1AFTER","\t", "TOKENS2AFTER","\t",
"TAGS2AFTER","\t",  "TOKENS3AFTER","\t", "TAGS3AFTER","\t",
"CONTEXT","\t","POSITION","\t","\n";
while (my $line = <ENTREE>) {
chomp $line;
        my @array = split ( /\t/, $line );
```

```
          push ( @tokens, $array[0] );
          push ( @tags, $array[1] );
          push ( @context, $array[2] );
         push ( @discourse, $array[3] );
}

#place features as n-3 and n+3 around each it this and that and also the
ENDO/EXO feat and the positional feature OBLI or NOMI
for (my $i=0; $i <= $#tokens; $i++){
          if (!defined($context[$i])) {
          $context[$i]=0;
                    }
          if (!defined($discourse[$i])) {
          $discourse[$i]=0;
                    }
          #if token is this, that or it and if it is a pro-form then
          if (($tokens[$i] =~/^[T|t]his$/ ) or ($tokens[$i] =~/^[T|
t]hat$/) or ($tokens[$i] =~/^[I|i]t$/) or ($tokens[$i] =~/^[T|t]hese$/)
or /^[T|t]hose$/) {
                      my $position = "NA";
                      # check if it is followed by a modal or a verb or by
an adverb or repeated proform and a verb or modal.
                    if ($tags[$i] =~ /TPRON|PRP/) {
                              if ( #Question construction
                                        ($tags[$i-1] =~ /MD/)
                                        or (($tags[$i-1] =~ /^do$|^does$|
^am$|^are$|^is$|^was$|^were$/) and (
                                                                  ($tags[$i+1]
=~ /^JJ$|^VB$|^VBG$/)
                                                                  or (($tags[$i+1]
=~ /^RB$/) and ($tags[$i+2] =~ /^|^VB$|^VBG$/)))
                                        or ($tags[$i+1] =~ /V.*|MD|NNS|
POS/) #Affirmative or negative construction - NNS and POS is added to
correct Treetagger's tagging errors on 's and third person singular
verbs.
                                        or (($tags[$i+1] =~ /RB|TPRON|PRP/)
and ($tags[$i+2] =~ /V.*|MD|NNS|POS/) #TPRON and PRP are added to
encompass repetitions in the case of oral expression
                                                            )
                                        )
                              ) {
                              $position = "NOMI";
                              }
                              else {
                              $position="OBLI";
                              }
                    }
      #Check the form is a DT
                    elsif (($tags[$i] =~ /DT/) and (($tags[$i+1] =~
/NN.*/)

          or (($tags[$i+1] =~ /JJ|DT/) and ($tags[$i+2] =~ /NN.*/)))) {
                              if (#Question construction
                                        ($tags[$i-1] =~ /MD/)
                                        or (($tags[$i-1] =~ /^do$|^does$|
```

```
^am$|^are$|^is$|^was$|^were$/) and (
                                                    ($tags[$i+3]
=~ /^JJ$|^VB$|^VBG$/)
                                                    or (($tags[$i+3]
=~ /^RB$/) and ($tags[$i+4] =~ /^|^VB$|^VBG$/)))
                                        or ($tags[$i+2] =~ /V.*|MD|NNS|
POS/) #Affirmative or negative construction
                                        or (($tags[$i+2] =~ /RB|TPRON|PRP/)
and ($tags[$i+3] =~ /V.*|MD|NNS|POS/)
                                                )
                                        )
                                ) {
                                $position = "NOMI";
                                }
                                else {
                                $position="OBLI";
                                }
                        }
                print SORTIE  $r,"\t", $tokens[$i], "\t",$tags[$i],"\t",
$tokens[$i-3],"\t", $tags[$i-3], "\t", $tokens[$i-2],"\t", $tags[$i-2],
"\t",$tokens[$i-1],"\t", $tags[$i-1], "\t", $tokens[$i+1],"\t",
$tags[$i+1], "\t", $tokens[$i+2],"\t", $tags[$i+2], "\t ",
$tokens[$i+3],"\t", $tags[$i+3],"\t", $context[$i],"\t",$position,"\n";
                }
}
close ( ENTREE );
        close ( SORTIE );
}
```

Annex L - PERL script for feature sequencing for TiMBL

PERL script used to convert 3 column format of corpus files into sequences of features followed by the class assigned to each sequence. This script is used to prepare the data for TiMBL.

```
#Author Gaillat Thomas. University of Paris Diderot
#Inspired from Tangy and Hathout http://perl.linguistes.free.fr/
#usage perl TT2seq_feat_3grams.pl directory
# directory is at the same level as the perl programme. It includes texts
that have been tokenized an tagged. One token and POS tag per line.
#_3grams_context_position
use strict;
use locale;

#multi-file processing in one folder

if ( $#ARGV != 0 ) {
    die "Usage : ", $0, " enter directory name please\n";
}

my $repertoire = $ARGV[0];


opendir ( REP, $repertoire ) or
    die "Impossible d'ouvrir ", $repertoire, " : ", $!, "\n";
my @fichiers = readdir( REP );
closedir ( REP );

my @fichierstt = grep (/\.tt$/, @fichiers);

foreach my $fichier ( sort @fichierstt ) {
         my $r = $fichier ;
         $r =~ s/^(.*)\.tt$/$1.seq4tbl/;
         print STDERR $r, "\n";

         open ( ENTREE, "<", $repertoire."/".$fichier) or
                  warn "Erreur d'ouverture de ",$fichier, " :" ,$!,
"\n";
         open(SORTIE, ">", $repertoire."/".$r) or die "impossible
d'ouvrir ", $r;

#throw each token tag pair in arrays.
my (@tokens, @tags, @context, @discourse);

while (my $line = <ENTREE>) {
chomp $line;
        my @array = split ( /\t/, $line );
         push ( @tokens, $array[0] );
         push ( @tags, $array[1] );
         push ( @context, $array[2] );
        push ( @discourse, $array[3] );
}

#place features as n-3 and n+3 around each it this and that and also the
ENDO/EXO feat and the positional feature OBLI or NOMI
```

```
for (my $i=0; $i <= $#tokens; $i++){
          my $position = "-";
          my $featVBZ="-";
          my $featED = "-";
          my $featNOT = "-";
          my $featCC = "-";
          my $featPUNC = "-";
          my $featCAP = "-";
          my $featREFNN="-";
          my $featREFPRON="-";
          my $featREFWH="-";
          my $featPREPINT = "-";
          my $featPREPPOST = "-";

          if (!defined($context[$i])) {
          $context[$i]="-";
                    }
          if (!defined($discourse[$i])) {
          $discourse[$i]=0;
                    }

          if (($tokens[$i] =~/^[T|t]his$/ ) or ($tokens[$i] =~/^[T|
t]hat$/) or ($tokens[$i] =~/^[I|i]t$/) or ($tokens[$i] =~/^[T|t]hese$/)
or  ($tokens[$i] =~/^[T|t]hose$/)) {



                    #If the form is a proform or a pronoun
                    if ($tags[$i] =~ /TPRON|PRP/) {

                              #Propriétés énonciatives. Rupture avec le
plan de l'énonciation en utilisant le preterit. Non rupture avec l'usage
du présent. Rejet émanant de l'énonciateur en utilisant la négation.
Focus change/contrasting with "and"

                              # if there is VB or VBZ
                              if (($tags[$i-3] =~ /^VB$|^VBZ$/) or
($tags[$i-2] =~  /^VB$|^VBZ$/) or ($tags[$i-1] =~ /^VB$|^VBZ$/) or
($tags[$i+3] =~ /^VB$|^VBZ$/) or ($tags[$i+2] =~  /^VB$|^VBZ$/) or
($tags[$i+1] =~ /^VB$|^VBZ$/)) {
                              $featVBZ = "VBZ";
                              }

                              #if there is an ED form in the close context

                              if (($tags[$i-3] =~ /VBD/) or ($tags[$i-2]
=~  /VBD/) or ($tags[$i-1] =~ /VBD/) or ($tags[$i+3] =~ /VBD/) or
($tags[$i+2] =~  /VBD/) or ($tags[$i+1] =~ /VBD/)) {
                              $featED = "ED";
                              }

                              # Negation in close context

                              if (($tokens[$i-3] =~ /n\'t|[N|n]ot/) or
($tokens[$i-2] =~  /n\'t|[N|n]ot/) or ($tokens[$i-1] =~ /n\'t|[N|n]ot/)
```

```
or ($tokens[$i+3] =~ /n't|[N|n]ot/) or ($tokens[$i+2] =~  /n\'t|[N|n]ot/)
or ($tokens[$i+1] =~ /n\'t|[N|n]ot/)) {
                            $featNOT = "NOT";
                            }

                            if (($tags[$i-3] =~ /CC/) or ($tags[$i-2] =~
/CC/) or ($tags[$i-1] =~ /CC/) or ($tags[$i+3] =~ /CC/) or ($tags[$i+2]
=~  /CC/) or ($tags[$i+1] =~ /CC/)) {
                            $featCC = "CC";
                            }

                            #Text properties. The start of a new
utterance.
                            # Capital letter on the token
                            if  ($tokens[$i]=~/[A-Z]/){
                            $featCAP= "CAP";
                            }

                            # Close to start or end of
sentence/utterance
                            if (($tags[$i-3] =~ /"|\.|\.\.|\.\.\.|\?/)
or ($tags[$i-2] =~  /"|\.|\.\.|\.\.\.|\?/) or ($tags[$i-1]
=~ /"|\.|\.\.|\.\.\.|\?/) or ($tags[$i+3] =~ /"|\.|\.\.|\.\.\.|\?/) or
($tags[$i+2] =~  /"|\.|\.\.|\.\.\.|\?/) or ($tags[$i+1]
=~ /"|\.|\.\.|\.\.\.|\?/)) {
                            $featPUNC = "PUNC";
                            }


                            #Endophora related properties.
                            #The presence of other referential or
endophoric items: Potential co-referentiality.
                            #NN NNS NNP TREL PRP TPRON
                            if (($tags[$i-3] =~ /^NN[P|S]?$/) or
($tags[$i-2] =~  /^NN[P|S]?$/) or ($tags[$i-1] =~ /^NN[P|S]?$/) or
($tags[$i+3] =~ /^NN[P|S]?$/) or ($tags[$i+2] =~  /^NN[P|S]?$/) or
($tags[$i+1] =~ /^NN[P|S]?$/)) {
                            $featREFNN = "REFNN";
                            }

                            if (($tags[$i-3] =~ /^PRP$|^TPRON$/) or
($tags[$i-2] =~  /^PRP$|^TPRON$/) or ($tags[$i-1] =~ /^PRP$|^TPRON$/) or
($tags[$i+3] =~ /^PRP$|^TPRON$/) or ($tags[$i+2] =~  /^PRP$|^TPRON$/) or
($tags[$i+1] =~ /^PRP$|^TPRON$/)) {
                            $featREFPRON = "REFPRON";
                            }

                            if (($tags[$i-3] =~ /^TREL$|^WDT$|^WP.?$/)
or ($tags[$i-2] =~  /^TREL$|^WDT$|^WP.?$/) or ($tags[$i-1] =~ /^TREL$|
^WDT$|^WP.?$/) or ($tags[$i+3] =~ /^TREL$|^WDT$|^WP.?$/) or ($tags[$i+2]
=~  /^TREL$|^WDT$|^WP.?$/) or ($tags[$i+1] =~ /^TREL$|^WDT$|^WP.?$/)) {
                            $featREFWH = "REFWH";
                            }

                            #Positional properties around the form
```

```
                                #Introductory preposition
                                if ($tags[$i-1] =~ /^TO$|^IN$/) {
                                $featPREPINT = "PREPINT";
                                }
                                #Prep after the form
                                if ($tags[$i+1] =~ /^TO$|^IN$/) {
                                $featPREPPOST = "PREPPOST";
                                }

                                        #Nominative or oblique case ,
check if it is followed by a modal or a verb or by an adverb or repeated
proform and a verb or modal.
                                if (
                                        ($tags[$i-1] =~ /MD/) #Question
construction
                                        or (($tags[$i-1] =~ /^do$|^does$|
^am$|^are$|^is$|^was$|^were$/) and (
                                                                ($tags[$i+1]
=~ /^JJ$|^VB$|^VBG$/)
                                                                or (($tags[$i+1]
=~ /^RB$/) and ($tags[$i+2] =~ /^|^VB$|^VBG$/)))
                                        or ($tags[$i+1] =~ /V.*|MD|NNS|
POS/) #Affirmative or negative construction - NNS and POS is added to
correct Treetagger's tagging errors on 's and third person singular
verbs.
                                        or (($tags[$i+1] =~ /RB|TPRON|PRP/)
and ($tags[$i+2] =~ /V.*|MD|NNS|POS/) #TPRON and PRP are added to
encompass repetitions in the case of oral expression
                                                        )
                                        )
                                ) {
                                $position = "NOMI";
                                }
                                else {
                                $position="OBLI";
                                }
                        }


                        #oTHERWISE IF THE the form is a DT
                        elsif (($tags[$i] =~ /DT/) and (($tags[$i+1] =~
/NN.*/)

                or (($tags[$i+1] =~ /JJ|DT/) and ($tags[$i+2] =~ /NN.*/)))) {


                                #Propriétés énonciatives. Rupture avec le
plan de l'énonciation en utilisant le preterit. Non rupture avec l'usage
du présent. Rejet émanant de l'énonciateur en utilisant la négation.

                                # if there is VB or VBZ
                                if (($tags[$i-3] =~ /^VB$|^VBZ$/) or
($tags[$i-2] =~  /^VB$|^VBZ$/) or ($tags[$i-1] =~ /^VB$|^VBZ$/) or
($tags[$i+4] =~ /^VB$|^VBZ$/) or ($tags[$i+3] =~  /^VB$|^VBZ$/) or
($tags[$i+2] =~ /^VB$|^VBZ$/)) {
```

```
                            $featVBZ = "VBZ";
                            }

                            #if there is an ED form in the close context

                            if (($tags[$i-3] =~ /VBD/) or ($tags[$i-2]
=~  /VBD/) or ($tags[$i-1] =~ /VBD/) or ($tags[$i+4] =~ /VBD/) or
($tags[$i+3] =~  /VBD/) or ($tags[$i+2] =~ /VBD/)) {
                            $featED = "ED";
                            }

                            # Negation in close context

                            if (($tokens[$i-3] =~ /n\'t|[N|n]ot/) or
($tokens[$i-2] =~  /n\'t|[N|n]ot/) or ($tokens[$i-1] =~ /n\'t|[N|n]ot/)
or ($tokens[$i+4] =~ /n't|[N|n]ot/) or ($tokens[$i+3] =~  /n\'t|[N|n]ot/)
or ($tokens[$i+2] =~ /n\'t|[N|n]ot/)) {
                            $featNOT = "NOT";
                            }

                            if (($tags[$i-3] =~ /CC/) or ($tags[$i-2] =~
/CC/) or ($tags[$i-1] =~ /CC/) or ($tags[$i+4] =~ /CC/) or ($tags[$i+3]
=~  /CC/) or ($tags[$i+2] =~ /CC/)) {
                            $featCC = "CC";
                            }

                            #Text properties. The start of a new
utterance.
                            # Capital letter on the token
                            if  ($tokens[$i]=~/[A-Z]/){
                            $featCAP= "CAP";
                            }

                            # Close to start or end of
sentence/utterance
                            if (($tags[$i-3] =~ /"|\.|\.\.|\.\.\.|\?/)
or ($tags[$i-2] =~  /"|\.|\.\.|\.\.\.|\?/) or ($tags[$i-1]
=~ /"|\.|\.\.|\.\.\.|\?/) or ($tags[$i+4] =~ /"|\.|\.\.|\.\.\.|\?/) or
($tags[$i+3] =~  /"|\.|\.\.|\.\.\.|\?/) or ($tags[$i+2]
=~ /"|\.|\.\.|\.\.\.|\?/)) {
                            $featPUNC = "PUNC";
                            }


                            #Endophora related properties.
                            #The presence of other referential or
endophoric items: Potential co-referentiality.
                            #NN NNS NNP TREL PRP TPRON
                            if (($tags[$i-3] =~ /^NN[P|S]?$/) or
($tags[$i-2] =~  /^NN[P|S]?$/) or ($tags[$i-1] =~ /^NN[P|S]?$/) or
($tags[$i+4] =~ /^NN[P|S]?$/) or ($tags[$i+3] =~  /^NN[P|S]?$/) or
($tags[$i+2] =~ /^NN[P|S]?$/)) {
                            $featREFNN = "REFNN";
                            }
```

```
                            if (($tags[$i-3] =~ /^PRP$|^TPRON$/) or
($tags[$i-2] =~  /^PRP$|^TPRON$/) or ($tags[$i-1] =~ /^PRP$|^TPRON$/) or
($tags[$i+4] =~ /^PRP$|^TPRON$/) or ($tags[$i+3] =~  /^PRP$|^TPRON$/) or
($tags[$i+2] =~ /^PRP$|^TPRON$/)) {
                            $featREFPRON = "REFPRON";
                            }

                            if (($tags[$i-3] =~ /^TREL$|^WDT$|^WP.?$/)
or ($tags[$i-2] =~  /^TREL$|^WDT$|^WP.?$/) or ($tags[$i-1] =~ /^TREL$|
^WDT$|^WP.?$/) or ($tags[$i+4] =~ /^TREL$|^WDT$|^WP.?$/) or ($tags[$i+3]
=~  /^TREL$|^WDT$|^WP.?$/) or ($tags[$i+2] =~ /^TREL$|^WDT$|^WP.?$/)) {
                            $featREFWH = "REFWH";
                            }

                            #Positional properties  of the form
                            #Introductory preposition
                            if ($tags[$i-1] =~ /^TO$|^IN$/) {
                            $featPREPINT = "PREPINT";
                            }
                            # Prep after the use of the NP
                            if (($tags[$i+1] =~ /^TO$|^IN$/) or
($tags[$i+2] =~/^TO$|^IN$/)){
                            $featPREPPOST = "PREPPOST";
                            }


                            #Nominative or oblique case

                            if (
                                      ($tags[$i-1] =~ /MD/) #Question
construction
                                      or (($tags[$i-1] =~ /^do$|^does$|
^am$|^are$|^is$|^was$|^were$/) and (
                                                            ($tags[$i+3]
=~ /^JJ$|^VB$|^VBG$/)
                                                            or (($tags[$i+3]
=~ /^RB$/) and ($tags[$i+4] =~ /^|^VB$|^VBG$/)))
                                      or ($tags[$i+2] =~ /V.*|MD|NNS|
POS/) #Affirmative or negative construction
                                      or (($tags[$i+2] =~ /RB|TPRON|PRP/)
and ($tags[$i+3] =~ /V.*|MD|NNS|POS/)
                                                      )
                                      )
                            ) {
                            $position = "NOMI";
                            }
                            else {
                            $position="OBLI";
                            }
                  }
         print SORTIE  $tags[$i],
         "\t", $tokens[$i-3],"\t", $tags[$i-3],
         "\t", $tokens[$i-2],"\t", $tags[$i-2],
         "\t",$tokens[$i-1],"\t", $tags[$i-1],
         "\t", $tokens[$i+1],"\t", $tags[$i+1],
```

```
          "\t", $tokens[$i+2],"\t", $tags[$i+2],
          "\t", $tokens[$i+3],"\t", $tags[$i+3],
          "\t", $context[$i],"\t",$position,
          "\t", $featVBZ,"\t", $featED,"\t", $featNOT, "\t", $featCC,
"\t", $featCAP, "\t", $featPUNC,
          "\t", $featREFNN, "\t",$featREFPRON, "\t",$featREFWH, "\t",
$featPREPINT, "\t", $featPREPPOST,
          "\t",$tokens[$i],"\n";
          }
}
close ( ENTREE );
        close ( SORTIE );
}
```

Annex M - R scripts

List of R scripts used for different purposes in the Chapter 6 and 7.

Script 1

This script is used in Section 6.2.2 page 259 to compute the dfbetas of model.02.

Testing the dfbeta assumption with R:

```
> summary(dfbetas <-abs(dfbeta(model.02)))
```

Script 2

In Section 6.2.2 page 259, this script is dedicated to computing the predictions of the model. In order to see how well the model classifies, the number of misclassified forms is computed according to the following code lines in R (Gries 2009, 305):

```
predictions.num <- fitted(model.02,
data=wsj.noce.did.proforms.determiners, type="response");
predictions.num
predictions.cat <- ifelse(predictions.num >=0.5, "this", "that");
predictions.cat
predictions.cat <-factor(predictions.cat)
MC.contingency<-table(wsj.noce.did.proforms.determiners$TOKENS,
predictions.cat); MC.contingency
MC1 <- sum(predictions.cat!
=wsj.noce.did.proforms.determiners$TOKENS)/nrow(wsj.noce.did.proforms.de
terminers); MC1
```

Script 3

In Section 6.2.2 page 259, the effect of the predictors are computed. To do so, R's *effect* function is used. For instance, the CONTEXT predictor's effect is computed with the following lines of code:

```
sot <- effect("CONTEXT", model.02)
preds<-data.frame(sot$x,
PREDICTIONS=ilogit(sot$fit),LOWER=ilogit(sot$lower),
UPPER=ilogit(sot$upper)); preds
```

Predictions are specified as probabilities computed with the *ilogit* function.

Script 4

In Section 7.1.2.1, the datasets are manipulated in order to prepare the forms for automated processing. An R script is used to sample *this* and *that* pro-forms and *it* pronoun from sequenced corpus files. The grammatical number is standardised. The singular form is applied to singular and plural forms (lines 14-15). Previously, all forms are set to lowercase (line 11).

```
6 #Data manipulation of corpus.
7 #manip sqe4tbl file to extract forms in their proforms functions.
8 a <-wsj.tokentags.it_PNR.this.that.40000.forms.manual.rev.training
9
10 #convert all this.that.it to lower case
11 a$V27 <-tolower(a$V27)
12 #replace plural forms by singular forms
13 #change all "these" to "this" and "those" to "that"
14 a$V27 <- replace(a$V27, a$V27=="these", "this"); a
15 a$V27 <- replace(a$V27, a$V27=="those", "that"); a
16
17
18 #create data frame composed of only pro-forms
19 #create data frame composed of only TPRON
20 a.this.TPRON <- subset(a, V1=="TPRON" & V27=="this")
21
22 #create data frame composed of only TPRON
23 a.that.TPRON <- subset(a, V1=="TPRON" & V27=="that")
24
25 #create data frame composed of only PRP
26 a.it.PRP <- subset(a, V1=="PRP")
27
28 #bind all three data frames
29 a.TPRON.PRP<- rbind(a.this.TPRON, a.that.TPRON,a.it.PRP)
30
31 #Delete V1 column and contextual feature (ENDO/EXO V14)
32 a.TPRON.PRP <- subset(a.TPRON.PRP , select = -c(V14))
```

Annex N - Concordances of *that* pro-form in a sample of the Diderot-LONGDALE corpus

	TAGS1BEFORE	TOKENS3BEFORE	TOKENS2 BEFORE	TOKENS1 BEFORE	TOKENS	TOKENS1 AFTER	TOKENS2 AFTER	TOKENS3 AFTER
29	RB	</A>	<B>	about	that	I	think	well
43	RB	movies	too	so	that	's	a	lot
92	RB	's	all	about	that	it	's	only
94	RB	felt	upset	about	that	</B>	<A>	</A>
119	RB	alone	mostly	so	that	's	an	impression
227	RB	<B>	and	so	that	was	very	good
239	RB	wow	and	so	that	was	good	I
260	RB	to	be	well	that	's	my	dream
280	RB	<clears throat	.	so	that	was	-LRB-	er
282	RB	music	.	so	that	was	interesting	that
287	RB	up	everything	so	that	was	very	interesting
367	RB	-RRB-	.	well	that	's	hard	-LRB-
445	RB	it	and	so	that	's	why	I
529	RB	er	-RRB-	so	that	's	contribute	it
541	RB	er	-RRB-	so	that	's	why	I
594	RB	things	<overlap	so	that	's	quite	interest
659	RB	yeah	ok	so	that	's	a	bit
675	RB	playing	and	so	that	's	really	a
682	RB	think	.	so	that	's	the	two
730	RB	er	-RRB-	so	that	's	a	litle
781	RB	they	talk	about	that	I	do	n't
785	RB	of	that	so	that	's	the	reason
790	RB	n't	think	about	that	I	just	.
877	RB	say	<laughs>	about	that	<laughs>	-LRB-	er
908	RB	-RRB-	and	after	that	I	go	to
909	RB	year	and	after	that	I	have	to
951	RB	er	-RRB-	so	that	's	why	it
964	RB	and	funny	so	that	was	cool	</B>
991	RB	America	.	so	that	was	cool	because
993	RB	to	see	so	that	was	really	really
1150	RB	not	serious	about	that	there	there	a
1165	RB	who	come	so	that	seemed	to	be
1258	RB	frightening	things	so	that	's	why	I
1334	RB	's	all	about	that	it	's	like
1420	RB	er	-RRB-	so	that	might	look	like
1426	RB	er	-RRB-	so	that	's	my	.
1477	RB	er	-RRB-	so	that	that	was	really
1893	RB	er	-RRB-	yes	that	's	all	I
1895	RB	English	.	so	that	's	why	I
1905	RB	.	and	so	that	's	why	I
1913	RB	.	.	so	that	's	why	that
1924	RB	er	-RRB-	so	that	's	-LRB-	er
1972	RB	this	picture	so	that	's	why	I
2327	RB	</A>	<B>	so	that	's	really	fine
2362	RB	you	had	all	that	it	did	n't
2375	RB	er	-RRB-	all	that	stuff	but	-LRB-
2458	RB	teachers	<overla>	so	that	's	great	<overlap
2464	RB	saying	.	so	that	's	another	problem
2545	RB	one	is	really	that	-LRB-	er	-RRB-
2595	RB	yeah	because	yes	that	's	was	it

Annex O - Questions on social background

Before recording learners in the Diderot-LONGDALE project, a number of “social” questions were asked. The answers provide metadata for the corpus.

1. Please enter your ID number here:
2. Please enter your ID number again for verification:
3. What is your year of birth?
4. Are you male or female?
5. What is your home country?
6. What is your native language?
7. What language(s) do you speak at home?
8. Apart from English, do you speak any (other) foreign languages?
9. Apart from English, which foreign language do you know best?
10. Do you know any more foreign languages? Please list them in decreasing order of proficiency, separated by commas.
11. What was your dominant language of instruction at school (before university)? [Language of instruction in primary school]
12. What was your dominant language of instruction at school (before university)? [Language of instruction in secondary school (high school)]
13. What subject are you currently studying?
14. How many years have you been studying at university?

15. For what degree or diploma are you currently studying?

16. What is the name of the institution where you are currently studying?

17. In which country are you currently studying?

18. Proportion of subject courses (other than English) that are taught in English in the institution where you are studying

19. How many years of English language classes did you have at school?

20. How many years of English language classes have you had at university? Have you taken any international exams or tests that assess your level of English (e.g. Cambridge, TOEFL, IELTS, etc.)?

21. Which exam(s) or test(s)?

22. Which exam(s) or test(s)? - comment

23. Have you spent some time in an English-speaking country?

24. Please enter any comments or additional information which you think would be useful.

Annex P - Personal publications related to this PhD

This is the list of publications related to this PhD project.

Gaillat, Thomas. 2013a. *This* and *That* in Native and Learner English: From Typology of Use to Tagset Characterisation. In Sylviane Granger, Gaëtanelle Gilquin & Fanny Meunier (eds.), *Twenty Years of Learner Research: Looking Back, Moving Ahead*: 167–177. (Corpora and Language in Use). Louvain-la-Neuve: Presses Universitaires de Louvain.

—. 2013b. *This* et *that* dans les domaines spécialisés du corpus ICE-GB : quelles caractéristiques distributionnelles ?. *ASp. la revue du GERAS* (64) : 161–183.

—. 2013c. Annotation automatique d'un corpus d'apprenants d'anglais avec un jeu d'étiquettes modifié du Penn Treebank. *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)* : 271–284. Les Sables d'Olonne : Association pour le Traitement Automatique des Langues.

Gaillat, Thomas, Pascale Sébillot & Nicolas Ballier. 2014. Automated Classification of Unexpected Uses of *This* and *That* in a Learner Corpus of English. In Lieven Vandelanotte, Kristin Davidse, Caroline Gentens & Ditte Kimps (eds.), *Recent Advances in Corpus Linguistics: Developing and Exploiting Corpora*: 309–324. Amsterdam: Rodopi.

Gaillat, Thomas, Martine Schuwer & Sophie Belan. 2014. Combiner les scénarios linguistique et actionnel au sein d'un parcours en ligne d'apprentissage de l'anglais. *Recherche et pratiques pédagogiques en langues de spécialité. Cahiers de l'Apliut* (Vol. XXXIII N° 3): 107–120.

Gaillat, Thomas. Les emplois adverbiaux de *this* et *that* dans les corpus d'apprenants. *Au coeur de la langue – Hommages à Martine Schuwer - Journée d'études organisée par LIDILE*. Université de Rennes 2: LIDILE. (Forthcoming)

Gaillat, Thomas, Nicolas Ballier & Pascale Sébillot. Exploring a Learner-specific Microsystem of Reference: the Case of *This*, *That* and *It*. A Multi-corpus Approach. *International Journal of Learner Corpus Research.* (Submitted).

Index

Résumé

Cette thèse s'attache à décrire les constructions inattendues en THIS et en THAT des apprenants francophones et hispanophones de l'anglais. Le chapitre 1 pose la problématique de l'étude des marqueurs THIS et THAT au sein des deux microsystèmes des déictiques et des proformes. Le chapitre 2 présente les différentes analyses de la référence de THIS et de THAT en anglais natif, dans les différents cadres théoriques (Cornish, Cotte, Halliday & Hasan, Kleiber, Fraser & Joly, Lapaire & Rotgé) et croise les problématiques de représentations (anaphore/deixis ; endophoricité/exophoricité) avec l'analyse des réalisations fonctionnelles. Le chapitre 3 dresse un état des lieux rapide de l'analyse de l'interlangue et montre la nécessité d'une approche dynamique des systèmes fondée sur la nécessité de la distinction fonctionnelle. Le chapitre 4 détaille les jeux d'étiquettes existants dans les corpus de l'anglais (Penn Treebank, Claws7, ICE-GB) et montre la nécessité d'une ré-annotation plus fine fondée sur des étiquettes fonctionnelles et d'une sémantique des positions (sujets vs. oblique). Le chapitre 5 décrit l'architecture de l'annotation multi-niveaux mise en œuvre pour l'analyse de corpus différents, les méthodes de ré-annotation automatique des catégories fonctionnelles (ainsi que leur évaluation) et expose les choix retenus pour l'interopérabilité de ces corpus. Le chapitre 6 propose une analyse statistique fondée sur des modèles de régression qui mettent au jour les tendances des variables opérationnalisées dans l'analyse (la L1, le mode écrit ou oral du corpus, le type de référence). Le chapitre 7 examine, à partir du recours aux classifieurs, le rôle respectif des propriétés linguistiques codées dans l'analyse et simule un système d'analyse automatique des erreurs. Le chapitre 8 tire les conséquences pour l'analyse linguistique des méthodologies mobilisées dans la thèse.

Mots-clés : Référence, deixis, anaphore, annotation, interopérabilité

Abstract

This thesis describes unexpected constructions based on THIS and THAT by French and Spanish learners of English. Chapter 1 raises the issue of the study of THIS and THAT as markers in the two microsystems of pro-forms and deictics. Chapter 2 covers different types of analyses of reference with THIS and THAT in native English and refers to different theoretical frameworks (Cornish, Cotte, Halliday & Hasan, Kleiber, Fraser & Joly, Lapaire & Rotgé). It cross-references representations (anaphora/deixis; endophoricity/exophoricity) with an analysis of functional realisations. Chapter 3 broaches the issue of interlanguage analysis, and it shows that a dynamic systemic approach grounded in the functional distinction of the forms is necessary. Chapter 4 gives details about existing annotation tagsets for English corpora (Penn Treebank, Claws7, ICE-GB). It shows the need for a finer-grained annotation relying on functional tags and for semantic information on the positions (subject v. oblique). Chapter 5 describes the multilayer annotation structure which is implemented for the analysis of different corpora. It also covers the methods used to automatically annotate functional categories (as well as their evaluation), and it justifies the choices made to support corpus interoperability. Chapter 6 offers a regression analysis which provides evidence on the tendencies of the operationalised variables (the L1, the written or spoken mode of the corpora and the type of reference). Chapter 7 examines the role of the previously coded linguistic properties of the analysis. With the use of classifiers, it describes a system for automatic error analysis. Chapter 8 concludes on the methodologies used in the thesis and their implications in linguistic analysis.

Keywords: Reference, deixis, anaphora, annotation, interoperability