This and that in native and learner English: From typology of use to tagset characterisation

Thomas Gaillat

This and *that* in native and learner English: From typology of use to tagset characterisation

Thomas Gaillat
Université Paris Diderot, Sorbonne Paris Cité, CLILLAC-ARP, (EA 3967) - France
Rennes 1 University - France

Abstract

Learner corpus research is now faced with a multiplicity of tagsets. It is therefore difficult to carry out cross-corpus analysis due to the variety of tags used for each part-of-speech (POS). In this paper, we envisage this issue through a specific linguistic point. We propose a typology of uses in both native and non-native corpora. Various tagsets are analysed so as to measure the relevance of the linguistic information provided for *this* and *that.* Overall, a comparative analysis of *this* and *that* in tagsets is proposed and the benefits and flaws of manual fine-grained annotation versus automatic annotation are assessed. This study comes as a first step towards automated annotation of *this* and *that* in various corpora as this process would pave the way to corpus interoperability at POS level.

Keywords: tagset, POS tagging, demonstratives, automatic annotation.

1. Introduction

Learner corpus annotation is a much debated issue as it is the base for linguistic analysis of learners' output. Considering the size of the data, it is crucial to orientate research towards automatic annotation and, as such, to focus on the type of annotation required. The following work is the preliminary phase of a project that aims to automatically annotate corpora including learner corpora. The first layer of the process includes part-of-speech (POS) tagging. As a comprehensive description of mainstream tagsets is beyond the scope of this article, we limit our analysis to the demonstratives *this* and *that*. A closer analysis of their use both in context and in tagsets will uncover linguistic details that cannot be appreciated in some tagsets, hence the need for modifications. Our goal is to automate the POS tagging process of corpora with a modified tagset which will help to provide large-scale finer-grained information on the uses of *this* and *that*. This paper gives an overview of the required refinements of existing tagsets in order to tag the two forms in both native and learner English corpora.

So, in this paper, we first focus on describing a typology of uses of the demonstratives. We review the literature to identify how natives use them, and we use a subset of the

LONGDALE[1] learner English corpus to examine learner output. It results in a classification of the forms that branches out in two directions. Not only are the correct uses of the forms classified but the unexpected uses of the forms also require full attention. We base our work on a referential and functional analysis of the forms so as to show different features of use according to various contexts. Much work has been done on the situations and the characteristics in which *this* and *that* are used, the novelty of our work lies in the combined examination of both learners and natives. The referential and functional features will give way to conclusions on the types of tags needed to perform a fine-grained annotation of the variety of uses both on syntagmatic and paradigmatic dimensions.
The outline of this paper consists of three parts. First, we focus on the uses of *this* and *that* by native speakers. Second, we look at learners' use of the forms, and suggest a classification of unexpected uses. Finally, various tagsets are analysed in the light of the characteristics highlighted in the aforementioned analysis.

2. The use of *this* and *that* in native language

In this part, we want to show how natives organise their syntax, and what they mean when they use the two demonstratives. We want to highlight the way natives operate their selection process of the two forms. First, we propose a functional approach in order to clearly present the position of the demonstratives in the sentence. Second, we focus on the meaning-effects of native output, and we explore the logic of the forms chosen by natives.

The function assigned to the forms is a first level of classification. The term 'demonstrative' is used when endophoric[2] or exophoric[3] reference is required in speech. Huddleston & Pullum (2002: 1504)[4] use the term 'demonstrative' and point two categories of use: dependent and independent. Indeed, 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.

This and *that* provide specific meaning to the contexts in which they can be found. With the proximate/distant paradigm the speaker is considered as the person located at the centre of the referential system, and every item referred to is in relation to the centre, hence the notion of distance. But it is argued that there is more than meets the eye (Biber *et al.* 1999: 349, Huddleston & Pullum 2002: 1505) as this is only an effect of a much deeper mechanism involving the speaker as a person. Fraser & Joly (1980) add the notion of sphere, one corresponding to the speaker's self and the other one corresponding to what is not the speaker's self. So *this* would correspond to the

[1] Cf. http://www.uclouvain.be/en-cecl-longdale.html (last accessed on 19 June 2012)
[2] Reference is endophoric when the referent is to be found within the speaker's discourse.
[3] Reference is exophoric when the referent is to be found in the situation in which the speaker is physically located.
[4] The chapter was written by Lesley Stirling and Rodney Huddleston.

speaker's sphere with notions such as interest, focus, foreground information, approach, present tense and speaker's responsibility in relation to referent. Conversely, *that* would correspond to notions involving rejection, conclusion, background information, distantiation, past tense and addressee's responsibility in relation to referent.

This approach shows the referents as shifting from one sphere to the other according to the speaker's interest for them. In doing so, *this* is used to refer to targets that belong to the sphere of the speaker and *that* to the non-self sphere (*non-moi* for Fraser & Joly 1979: 117; 1980: 31). So the explanations for the use of one form or the other are to be found in the referential procedure which can be divided into two parts. First, the speaker chooses the deixis type required to establish the link with the referent in terms of space and time. Once this process is complete, the speaker performs a selection to express how the referent relates to his/her sphere. Should the referent be part of the speaker's self for reasons such as interest, the choice will be *this.* In case it is necessary to locate the referent outside the speaker's sphere in order, for instance, to close the topic without further discussion, the choice will be *that*. The following example is taken from Fraser & Joly (1980: 35):

> I was often in liquor; and when in *that* condition, what gentleman is master of himself? Perhaps I did, in *this* state, use my lady rather roughly.

Here *this* and *that* are used with two synonymous substantives that refer to the same issue of being inebriated. There is an opposition between what is generic and what is the particular case of the narrator. With *that*, the referent is linked to *gentleman,* which locates it outside the speaker's sphere. With *this*, on the other hand, the narrator tells his own particular story.

3. A case study: the use of *this* and *that* in learner English

Having covered how *this* and *that* are used in native English, the focus is now placed on the issue of learner English, including unexpected uses. Agnieszka Lenko-Szymanska's study (Lenko-Szymanska 2004: 90) corroborates teachers' intuitions that referential processes constructed with the help of demonstratives are not always correct. Petch-Tyson (2000: 52) shows that variation exists in the uses of demonstratives depending on the learners' L1. Based on a subset of the Diderot-LONGDALE corpus[5] of 25 manually transcribed recordings of French learners of English, and totalling more than 125 minutes of speech, we analysed occurrences of *this* and *that* recorded in a free speech context where speakers explained one of their favourite experiences and answered questions from a native on their daily life. Our goal was then to address the following question: What kind of unexpected use can be found in learner language when referring to a known entity?

When looking at unexpected use from learners it appears that difficulties result, not

[5] http://www.clillac-arp.univ-paris-diderot.fr/projets/longdale (last accessed on 18 June 2012).

from the distribution patterns, but rather from the selection of a given form. In fact, learners operate substitutions. There are two main substitution branches. Either they substitute *this* with *that* or vice-versa within the deictic system itself (Section 3.1), or they substitute one of the demonstratives with another competing form corresponding to another system (Section 3.2).

3.1 Deictic system substitutions

First, there are endophoric/exophoric substitutions. Referring to Fraser & Joly (1980), some occurrences suggest that unexpected uses can be found at the endophoric/exophoric level since some forms suggest confusion in the referential procedure due to unclear distinction between endophoric and exophoric reference. For example, in (1) the user applies an exophoric deictic form in an endophoric context, resulting in a blurred referential process, as the referent becomes unclear due to the absence of a clear entity in the situation of communication. *This* is used with its collocate *like* in order to perform a situational reference process in spite of the fact that the addressee's expectation is an anaphoric procedure on the previously mentioned entities such as *church* and *shops*. In (2), the use of *this* is rather unexpected as it refers to a moment in which the speaker is not located. So in this example, temporal reference is biased by applying an exophoric use of *this* in an endophoric context.

> (1) […] 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)
> (2) […] we we waited for people to . to arrive because I got a . I got earlier cause I haven't class *this* day . (DID0162-S001)

Second, there are endophoric substitutions. What is meant here is the substitution between the two demonstratives within endophoric deixis. 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, especially in anaphoric procedures (Fraser & Joly 1980: 28). So in many cases, it appears that choices denote variations in meaning effects. However, some choices remain incompatible with their context as suggested by tests[6] on native speakers with the same contexts. In example (3), 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 would be expected.

> (3) […] 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 </B> (DID0167-S001)

[6] In order to predict the expected use in the same contexts as the learners', the same utterances were proposed to a panel of native speakers. These native speakers were asked to fill in gaps corresponding to the positions of *this* and *that*. Within unexpected uses, errors were identified when all natives did not select the form. When the selection was considered possible but not unique, competing forms such as *the* and *it* were suggested by native speakers.

3.2 Interactions: two micro-systems

Having explored the substitutions within the deictic system, the focus is now put on the second type of substitution, namely substitution between either a demonstrative and a pro-form or a determiner.

Endophoric use: substitution it vs. that/this

The first type of interaction is found within the endophoric category, *i.e.*, substitutions between *it* and *that/this.* In this case, the pro-form use of the demonstrative 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 unexpected uses which show that learners understand the referential procedure but do not understand the conditions in which they can use the demonstratives. Table 1 lists a series of occurrences classified according to the function of the form (e.g. subject vs. object). By native preference we refer to the fact that the occurrences were also submitted to native speakers[7] in order to see what choice they would have made in the same contexts. A classification is proposed according to natives' preferences.

As can be seen in Table 1, unexpected use can be found in all four cases of uses. As appears in this micro-system the demonstratives act as competitors with the pronoun *it*. The opposition is placed on a paradigmatic axis as all forms compete for the same functional position of pro-form. Their paradigmatic opposition is at the origin of the discrepancies between native and learner preferences in their selection process. As this is a qualitative approach, statistics about each case have not been compiled. This remains a task to be completed as such information would allow us to have a better view of the major sources of error. In order to achieve such a task it seems relevant to develop a search on data annotated in tree structures (parsing), and also to tag the determiners differently from the pro-forms as both types of information would help identify the function of each form and allow accurate calculations.

Native preference	**Function**	**Learner use**
It	Pro-form: Subject	(4) […] 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	(5) <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)
This/that	Pro-form: Subject	(6) […] 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

[7] Cf. footnote above.

		(eh) try to (em) to apologize to them through the book but (er) </B> (DID0164-S001)
This/that	Pro-form: Object	(7) […] 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)

Table 1. Interactions between the demonstratives and it

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, Huddleston & Pullum (2002: 1507)[8] 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 the following succession of items is to be expected: [Reference via Noun Phrase] > [Reference via *this/that*] > [Reference via *it*].

At sentence 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 Fraser and Joly about foreground vs. background information (Fraser & Joly 1979: 138). In some cases the information carried by the form clearly needs to be placed in the background as focus is given to a new element. In example (5) the new information item is 'grandma' so 'pizza' should not receive renewed focus. The fact of choosing a demonstrative instead of the pronoun introduces a logical contradiction, hence the unexpected use. Conversely, the learner may only repeat the referent with *it* where, instead, the demonstrative should bring a specific meaning expected to be found in the context as in sentences (6) and (7). Still at sentence level, example (4) can be looked at in the light of Huddleston and Pullum's definition (2002: 226) of the dummy pronoun that "makes no independent contribution to the meaning".

Degrees of specificity: substitution the/ this/ that

Determination is the second micro-system in interaction with the demonstratives, with *the* as determiner. Table 2 lists a certain number of occurrences where the determiner *the* is not chosen by learners even though tests[9] of the phrases on native speakers show its selection. As in the previous micro-system of interactions between *this, that* and *it*, the demonstratives are in direct paradigmatic opposition to the article *the*. When the choice of the determiner is to be operated, *the* appears as a competitor with *this* and *that*, as all three forms may function as determiners. Learners' choices show discrepancies with natives' preferences.

[8] The chapter was written by Lesley Stirling and Rodney Huddleston.
[9] Cf. footnote 6.

As previously done with *this/that* and *it*, it is relevant to analyse the variations between native and learner English 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. Biber *et al.* (1999: 272) give the first stage as the fact of "marking an entity as known". Guillemin-Flescher (1993: 181-208) adds an extra step in the construction progression by indicating that the demonstratives single out the referent. So there is a scale for the use of the article *the* or the demonstratives. If the reference procedure is only limited to marking an entity as known, then users avail themselves of the article *the.* If the procedure requires the need to single out a referent in relation to another one, both of them linked to the same class entity, then the users may avail themselves of the demonstratives. Our sample data suggest that the distinction is not always performed correctly. The examples in the corpus indicate a tendency to over-determine referents that only require marking with *the* as in (8).

Native preference	**Learner use**
The	(8) <A> okay how old were you when you visited for the first time [Morocco] </A> […] <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	(9) […] I remember there was this little comic book store just just . like one block away from the . the: the Empire State Building [...] 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 2. Interactions between the demonstratives and article the

4. Tagsets

As mentioned in the introduction of this paper our ultimate goal is to be able to compare the uses of the demonstratives in several corpora. In order to do so, we have described a typology of use of the forms. Parallel to that, comparing large corpora implies the ability to annotate them in a uniform and consistent manner, so the choice for automatic processing (Leech 2005: 17-29) was made and POS-tagging will be the first level of annotation that will be undertaken to annotate corpora. Consequently, the choice of a tagset that provides accurate information on the uses of the demonstratives

becomes central. Thus, it is important to analyse several tagsets, commonly and historically used for POS-tagging and error-tagging, in order to see what tags the demonstratives are assigned. This second part of our study reveals that tagsets have versatile and non-uniform methods for the characterisation of the demonstratives and that there are two types of tagsets: POS tagsets and error tagsets that do not necessarily mention the POS (Granger 2008: 346-347). POS tagsets initially developed to annotate native corpora include the Penn Treebank (Marcus *et al.* 1993), CLAWS7 (Garside 1987) and TOSCA-ICE (Aarts *et al.* 1998). There are also those used to annotate learner errors like the Louvain tagset (Dagneaux *et al.* 1996) or the fine-grained tagset NOCE (Díaz-Negrillo 2007).

The annotation process of deictic forms also leads to the side exploration of other realisations of *that* and Table 3 reflects 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 Treebank and CLAWS7, but the latter leaves room for the pronominal function in its description. TOSCA/ICE clearly identifies *that*, and so do NOCE and the Louvain tagsets.

If we observe the data in Table 3 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 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).

As it appears, functional distinction is an issue. Without a clearer approach of this linguistic aspect, automated tagging would yield a blurred vision of the uses of the demonstratives. It appears necessary to distinguish the forms according to their functional role in the utterances. When learners perform the selection process, there are competitors that interact for the same syntactic position. The alternate candidate to a demonstrative is either its counterpart or *it* or *the*. However, the alternate candidate varies depending on the syntactic position of the demonstrative. Subsequently, error analysis in learner corpora requires that such a distinction be taken into account.

	Functions	**Penn Treebank Native POS tags**	**CLAWS7 native POS tags**	**ICE-GB native POS tags**	**NOCE error tags**	**Louvain error tags*****
this	Determiner	DT*	DD1/2	PRON(dem,sing)	DT	GP
this	Pro-form	DT**	DD1/2	PRON(dem,sing)	WG.PO.DM.DL/PX.	GP
that	Determiner	DT*	DD1/2	PRON(dem,sing)	DT	GP
that	Pro-form	DT**	DD1/2	PRON(dem,sing)	WG.PO.DM.DL/PX.	GP

that	Complementizer	IN	CST	CONJUNC(subord)	CJ	LCS OR XCONJCO
that	Relative pronoun	WDT	CJT	PRON(rel)	PO.RL	GP
that	Adverbial	RB	RG	ADV(inten)	AV	GADV OR GADVO OR LS

* 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.
*** 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 3: This *and* that *in tagsets*

5. Conclusion

In this article we have explored how learners make use of the demonstratives. After establishing an analytical grid of native use, a classification of learner errors was attempted. Two levels of unexpected use have been highlighted. First, it may appear within deixis itself due to confusion between endophoric and exophoric contexts, or between the expected meanings of a demonstrative in a given context and the actual selection operated by learners. The second level of use is found within the proximity that exists between the deictic system and two other systems: endophoric reference with *it*, degrees of specificity with *the.* As it appears, interactions exist with one system or the other, depending on the syntactic position of the demonstratives. Alternate candidates to the demonstrative appear as competitors in the learners' selection process.

This work on learner use comes as a first step for the POS annotation of the Diderot-LONGDALE corpus with the Penn Treebank. Even though the copyrighted CLAWS system implements a tag distinction between the relative and complementizer functions, the distinction between determiner and pro-form is not made. The open-source Treetagger program (Schmid 1994) implements the Penn Treebank tagset and does not include any distinction between the determiner and pro-form functions either. However, the program makes it possible to modify the tagset with finer-grained tags for *this* and *that*. We see the POS layer as a base for learner error analysis. Our study has shown the need for an annotation scheme to encompass all types of uses of the demonstratives, including unexpected uses. In parallel, several tagsets have been detailed in relation to *this* and *that*. A comparative view of the relevant tags has shown characteristics that would prove to be limitations for the differentiation of the various forms depending on the syntactic position or their meanings. These limitations need to be lifted by enriching the Penn Treebank with the functional distinction between determiner/pro-form and by enriching the POS layer with encapsulated features (in a similar way to ICE-GB) such as the semantic distinction between *this* and *that* pro-forms, or the paradigmatic interaction between pronouns or between determiners.

With a view to comparing data from various learner and native corpora, it is essential to have a more standardising approach on the issue (Wynne 2005). We propose to develop a tool to import any given corpus of raw data with embedded automated tagging. Stochastic methods are used in order to improve data characterisation and learner error prediction. The TreeTagger software will first be used to POS-tag the corpus and the classifier TiMBL will be used on the POS-token pairs to extract finer-grained information in the form of features. Our methodology, if successfully implemented, would be a first step towards a standardised tagset since the existence of many tagsets hinders cross-corpus data comparison. This preliminary approach is part of the broader question of corpus interoperability. It is necessary to develop tools for tagging automation and to simultaneously send queries to several corpora. As such, query results, retrieved from several corpora, would help linguists to better understand the learning process of a language. They would also help English teachers to anticipate specific language difficulties, and e-learning course authors to enrich applications by customising feedback to learners' output.

References

Aarts, J., Van Halteren, H. & Oostdijkn, N. (1998). The linguistic annotation of corpora: The TOSCA analysis system. *International Journal of Corpus Linguistics* 3(2), 189-210.

Biber, D., Johanson, S., Leech, G., Conrad S. & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Cotte, P. (1993). De l'étymologie à l'énonciation. *Travaux de Linguistique et de Philologie* XXXI, 43-89.

Dagneaux, E., Denness, S., Granger, S., & Meunier F. (1996). *Error Tagging Manual version 1.1*. Louvain-la-Neuve: Centre for English Corpus Linguistics, Université catholique de Louvain.

Díaz-Negrillo, A. (2007). *A Fine-Grained Error Tagger for Learner Corpora*. Unpublished PhD dissertation, Jaén: University of Jaén.

Fraser, T. & Joly A. (1979). Le système de la deixis : Esquisse d'une théorie d'expression en anglais. *Modèles linguistiques* 1, 97-157.

Fraser, T. & Joly A. (1980). Le système de la deixis (2) : Esquisse d'une théorie d'expression en anglais. *Modèles Linguistiques* 2, 22-49.

Garside, R. (1987). The CLAWS word-tagging system. In R. Garside, G. Leech, & G. Sampson (eds) *The Computational Analysis of English: A Corpus-based Approac*h. London: Longman.

Granger, S. (2008). Learner corpora in foreign language education. In N. Van Deusen-Scholl & N.H. Hornberger (eds) *Encyclopedia of Language and Education, vol 4,* Second and Foreign Language Education. US: Springer, 337-351.

Guillemin-Flescher, J. (1993). Étude contrastive de la deixis en anglais et en français. In L. Danon-Boileau & J.L. Duchet (eds) *Opérations énonciatives et interprétations de l'énoncé.* Paris: Ophrys, 181-208.

Huddleston, R. & Pullum, G.K. (eds). (2002). *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Leech, G. (2005). Adding linguistic annotation. In M. Wynne (ed.) *Developing Linguistic Corpora: A Guide to Good Practice.* Oxford: Oxbow Books, 17-29.

Lenko-Szymanska, A. (2004). Demonstratives as anaphora markers in advanced learners' English. In G. Aston, S. Bernardini & D. Stewart (eds) *Corpora and Language Learners* (Studies in Corpus Linguistics 17). Amsterdam & Philadelphia: John Benjamins, 84-108.

Marcus, P.M., Marcinkiewicz, M.A. & Santorini, B. (1993). Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics* 19(2), 313-330.

Petch-Tyson, S. (2000). Demonstrative expressions in argumentative discourse. A computer-based comparison of non-native and native English. In S. Botley & A. M. McEnery (eds) Corpus-based and Computational Approaches to Discourse Anaphora. Amsterdam & Philadelphia: John Benjamins, 43-64.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*. Manchester: UK.

Wynne, M. (2005). *Developing Linguistic Corpora: A Guide to Good Practice*. Oxford: Oxbow Books. http://ahds.ac.uk/linguistic-corpora/ (last accessed on 12 December, 2010).