31st International Conference on Lexis and Grammar

Jan Radimský

Università degli Studi di Salerno

Dipartimento di Scienze Politiche, Sociali e della Comunicazione

University of South Bohemia

Faculty of Philosophy

31e Colloque International sur le Lexique et la Grammaire

Avec la session thématique *Adverbes et compléments adverbiaux*

du 19 au 22 septembre 2012

Nové Hrady (République tchèque)

31st International Conference on Lexis and Grammar

With the thematic session *Adverbs and Adverbial complements*

19-22 September 2012

Nové Hrady (Czech Republic)

Indications bibliographiques

Titre : Actes du 31e Colloque International sur le Lexique et la Grammaire
Responsable éditorial : Jan Radimský
Publié par : Université de Bohême du Sud à České Budějovice (République tchèque)
ISBN (version électronique) : 978-80-7394-409-4

Sommaire / Contents

Processing Euskara Complex Postpositions in a Rule Based Approach in Order to Improve Shallow Syntactic Disambiguation

Jose Maria Arriola

Abstract: This paper presents a rule-based linguistically motivated grammar based on Constraint Grammar formalism (Karlsson, 2006) for assigning the correct syntactic function tags to complex postpositions. The hand-crafted rules are linguistically motivated and designed to improve the surface syntactic analysis. The complex postpositions are processed at surface syntactic level rather than considering these elements as postpositions categories in the lexicon. The developed grammar assigns the appropriate syntactic function to all the elements analyzed as complex postposition component in the texts we have experimented with. Besides, this approach reduces the overall syntactic ambiguity of syntactic function tags in 2, 6% and avoids erroneous syntactic analysis.

1. Introduction

This article describes the rule-based grammar for assigning shallow[1] syntactic tags to the components of complex postpositions. This approach is the continuation of the work in shallow syntactic parsing of Euskara. Euskara is an agglutinative language with a medium to large sized system of affixed case markers/postpositions[2].

Postpositions in Euskara play a role similar to that of prepositions in languages like English or Spanish, so that, postpositions suffixes are attached to the last element of the a phrase. They are defined as "forms that represent grammatical relations among phrases appearing in a sentence" (Euskaltzaindia, 1994). There are two main types of postpositions in Euskara: (1) a suffix appended to a lemma and, (2) a suffix followed by a lemma that can also be inflected.

```
(1)teilatu-tik
   teilatu-(from the)
   from the roof
(2)teilatu-aren gain-etik
   teilatu-(of the) top-(from the)
   from the top of the roof
```

The last type of elements has been termed as complex postposition. We will use this term to name the whole sequence of two words involved, and not just to refer to the second element. Complex postpositions can be described as:

(3) lemma1 + (suffix1 + lemma2 + suffix2)

In these constructions, the second lemma is fixed for each postposition, while the first lemma allows for much more variation, ranging from every noun to some specific semantic classes.

The above description (3) is intended to stress (with parentheses) the fact that the combination of both suffixes with the second lemma acts as a complex case-suffix that is "appended" to the first lemma. Both suffixes present different combinations of number and case, which can agree in several ways, depending on the lemma, case or contextual factors. Table 1 shows the different variants of the complex postposition, derived from the lemma *arte*. The lemma arte is polysemous ("means, holm oak, art, time, skill, among, until").

Lemma2	Suffix1	Suffix2	Examples
arte (noun)	-en (genitive)	-an/-ra/-tik/-ko (inessive/alative/ablative/genitive)	Gazteen artean (among young people)
	-0 (no case)	-an/-ra/-tik/-ko (inessive/alative/ablative/genitive)	Jende artean (among people)
	-absolutive	-0/ko (no case)/genitive	Bederatziak arte (until nine)
	-ra (alative)	-0/ko (no case)/genitive	Bilbora arteko trena (the train to Bilbao)
	-0 (no case)	-0 (no case)	Bihar arte (until tomorrow)

Table 1. Complex postpositions for *arte*.

These lemmas have not been included with the postposition category in the lexical database for Euskara (EDBL[3]) just to simplify our disambiguation process[4]. We have treated at surface syntactic level those postpositions that are formed by a suffix followed by a lemma (postposition) that can be also inflected:

[1] The shallow syntactic process is composed of a number of different grammars dealing with chunking, understood as recognition of phrase boundaries.

[2] The definition of the terms case/postposition is disputed. In the current terminology only ergative, absolutive and dative are considered cases, while the others are considered affixed postpositions.

```
(i) Bederatziak      arte
    Bederatzi + ak   arte
      (nine)(until)

(ii)  Zuhaitz  ederren            artean
      Zuhaitz  eder +en           arte +an
      (tree)   (beautiful + of)   (between + in)
```

The postposition element "arte" in the first example (i) takes as first component an NP in absolutive case –ak ("*Bederatziak*", stands for 'nine'), and in the second example (ii), "arte" is inflected in inessive case –n and takes as first component a noun group in genitive case –en ("zuhaitz ederren", stands for 'of beautiful tree'). If we look closely at the lexical items that fill in the postposition role, we notice that "arte" is a noun following the information provided by our morphological analyzer.

Most of the main elements of postposition structures are nouns. Literature on Euskara regarding these elements considers them as postposition category (Hualde, 2002).

We opted for annotating these structures as postpositions at shallow syntactic level rather than considering these elements as postposition categories. We have all lemmas and suffixes that take part in postpositions in EDBL. Most of the syntactic information is first introduced with all ambiguity regardless of the context and later select and remove rules take care of disambiguation. As a result, most of the main elements (lemmas) of these postpositions are included in EDBL with noun category and the corresponding syntactic function tags[5] for nouns:

```
(iii) Bederatziak                    arte
      DET  @OBJ @PRED @SUBJ      NON-FINITE VERB  @-NON-FINITE-VERB
      NOUN @OBJ @PRED @SUBJ      NOUN @OBJ @PRED @SUBJ
      NOUN @SUBJ                 NOUN @CM>
```

In example (iii) we can see that the postposition "arte" has four syntactic function tags corresponding to a noun and one for the non-finite verb interpretation. Besides, the first element of the postposition structure "bederatziak" has seven syntactic function tags taking into account the different morphological analysis. In Euskara, both morphological and syntactic ambiguity exist, i.e. one word receives multiple analyses. Syntactic ambiguity is added on top of morphological ambiguity.

However, our study is restricted to shallow syntax so we couldn't deal with constructions that are semantically ambiguous. Morphosyntactic information is insufficient to take care of this kind of ambiguity.

2. Previous work

This section outlines the previous work done in the surface syntactic processing. The shallow syntactic parser (Aduriz, 2003) is divided into several modules, each one dealing with a different task. First of all, the text is tokenized an analysed morphologically. After that, a tagger/lemmatiser obtains the lemma and the category corresponding to each word form, and another module disambiguates the proposed tags. Then, a rule-based chunker (Abney, 1991) identifies verb and noun phrases, and, a specific chunker for postpositions identifies complex postpositions, finally, a dependency based syntactic tree is obtained by means of a rule-based module.

We reuse the chunking grammar of complex postpositions. The complex postpositions the system recognizes in this phase consist of both a case suffix followed by an independent word. For example: *gizonaren aurrean* 'in front of the man'. This type of complex postposition is taken into account in the recognition of noun chains (these noun chains also represent a postpositional system even though the

postposition, in this case, consists of a single suffix). The function tags %INIT-POS and %FIN-POS define the beginning and the end of postpositional phrases.

Based on the analysis of the chunking grammar of complex postpositions the syntactic information extracted was a source of several syntactic errors, because the chunker works at phrase level and identifies accurately the postposition structures, but on the contrary the syntactic information attached to the postposition elements is erroneous and ambiguous. The set of rules of the grammar take into account the morphological features (the case demanded by the postposition) and the set of elements considered as postpositions. This way the chunker of complex postpositions recognizes by means of mapping rules the main complex postpositions tagging them with

---

[3] Currently, EDBL resides under the ORACLE DBMS, on UNIX, and it may be consulted via the Internet (http://ixa2.si.ehu.es/edbl).

[4] After morphosyntactic analysis (MORFEUS), the tagger/lemmatiser EUSTAGGER obtains the lemma and category of each form and also performs disambiguation using the part of speech (POS).

[5] Main syntactic function tags: subject (@SUBJ), object (@OBJ) and predicate (@PRED). Modifier function tags: @CM> stands for modifier of the element carrying case. Functions related with verbs: @-NON-FINITE-VERB.

the above mentioned function tags (%INIT-POS and %FIN-POS define the beginning and the end of postpositional phrases). For instance:

MAP (%INIT-POS56) TARGET IZE-DET-IOR-ADJ-ELI-SIG
IF (0 ABSOLUTIVE) (1 POSTPOSIZIOAK-56);

MAP (%FIN-POS56) TARGET POSTPOSIZIOAK-56
IF (-1 IZE-DET-IOR-ADJ-ELI-SIG+ABS);

Example of mapping rules for recognizing complex postpositions

The chunking grammar of complex postpositions after applying the above rules in the following example: "*Batzuetan ez dira etxera itzultzen iluntzeko bederatziak arte*" (stands for: 'Sometimes they don`t come back to home until nine') give the following output:

bederatziak %INIT-POS56 arte %FIN-POS56

The chunking grammar contains 192 mapping rules in order to detect the initial and the end of postposition structures. For each postposition the rule defines the category of the postposition according to EDBL and the case of the phrases that are selected by the postposition and the cases of the postposition. For instance, if the postposition "arte" selects between others a phrase in possessive genitive case –en, then the postposition will take the following cases: inessive -an; adlative –ra; ablative –tik and local genitive –ko.

3. The experiment

The grammar developed in this experiment assigns the appropriate syntactic function tag to all the elements of the postposition structure and reduces the ambiguity in the following way: one syntactic tag for "bederatziak" and one adverbial syntactic tag for "arte":

```
(iv)  Bederatziak          arte
      DET @CM>          NOUN @ADVERBIAL
```

As we can see in (iv) the analysis obtained after applying the grammar rules we have a postposition structure with adverbial function. For that purpose a set of replace rules was written in order to attach to the complex postposition components the corresponding surface syntactic tag. The rules make use of the morphosyntactic information and the chunk tags attached by the chunker. Besides, some disambiguation rules were implemented based on the output of the chunker.

For instance, the rules assigning the appropriate syntactic function tags to the complex postposition '*bederatziak arte'* have the following format:

REPLACE (NOUN ARR BIZ- ABSOLUTIVE MG @ADLG %FIN-POS56)
TARGET (NOUN %FIN-POS56)
IF (-1 (%INIT-POS56));

The above replace rule attaches to the postposition 'arte' the adverbial function tag (@ADLG and maintains its morphological analysis as well as its corresponding chunk marker tag (%FIN-POS56).

On the other hand, the ambiguity between the noun and the verb reading of 'arte' is resolved making use of the chunk marker tag and the noun reading is selected by means of the following rule:

REMOVE (VERB) IF (0 NOUN) (-1 ({POS-HAS56));

With regard the postposition element 'bederatziak' the following replace rule attaches case mark element modifier function tag (@CM>) and maintains its morphological analysis as well as its corresponding chunk marker tag (%INIT-POS56).

REPLACE (DET DZH NMGP ABS NUMP MUGM @CM> %INIT-POS56)
TARGET (DET %INIT-POS56))
IF (1 ((%FIN-POS56));

The grammar is composed by 74 replace rules and 14 disambiguation constraint rules. One set of replace rules assign to the initial part of the postposition the following syntactic function tags: @CM> that stands for modifier

of the element carrying case or @NOUN-COMPLEMENT that stands for the noun modifier. Another set of rules replaces the syntactic function tag of the final element of the postposition for the function tag @ADLG that stands for adverbial function tag.

The rules refers to the initial element of the complex postposition and therefore apply to any word that is tagged with %INIT-POS and the characteristics that appear in context specification of the rule in order to replace the syntactic function tag with @CM> or @NOUN-COMPLEMENT. In the same way other set of rules refer to the final element of the complex postposition and therefore apply to any word that is tagged with %FIN-POS and the characteristics that appear in context specification of the rule in order to replace the syntactic function tag with adverbial @ADLG.

Our working corpus to develop the grammar contains 53324 tokens and the ambiguity rate is the % 5, 46 analyses per token. After applying the grammar we reduce the ambiguity rate per token to 5, 3% we have removed 7,665 syntactic function tags that are inadequate. The new syntactic function tags attached to the complex postpositions and the reduction of the ambiguity reverts in the improvement in the quality of several applications, such as Part-Of-Speech (POS) taggers and parsers.

4. Evaluation

The test corpus (1680 tokens) is running text and contains 70 postposition cases. The postpositions were annotated by the chunker on the analysis of the morphological analyzer. This leads to high ambiguity rate on the one hand, but takes into account all the morphological analyses that are necessary for chunking on the other hand.

These postpositions were annotated with the appropriate syntactic function tag by means of 74 replace rules. The syntactic ambiguity rate per token before applying the rules was 5, 97 syntactic analysis. After applying the rules the ambiguity rate is 5, 71 analysis per token. In 5 of the 70 postposition cases the existing chunking rules do not manage to annotate the postposition chunk correctly. Wrong applications of rules are mainly due to the high ambiguity of some words and scope mistakes of the rules. The precision[6] and recall[7] for the replace rules involved in the correct assignment of syntactic functions and disambiguation are 94, 2% and 100% respectively. In 4 cases the replace rules need to be refined in order to apply the correct syntactic function tag.

Taking into account the general impact of the annotation of complex postpositions on the overall analysis with respect to the qualitatively complexity of postpositions, as they are distributed across two words, and they also show different kinds of syntactic agreement, we think that this approach is necessary in order to improve syntactic disambiguation.

5. Conclusions and future work

The experiment has shown that linguistically motivated grammar resources encoding linguistic analysis such as complex postpositions can be reused. Grammar rules based on the complex postposition structures provide an efficient way to reduce syntactic ambiguity as they manage to select the correct syntactic function in cases where the syntactic context itself remains ambiguous, because the identification of initial and the ending of the postposition can be used to resolve this ambiguity. That's why chunking process has a key function in picking out and assigning the correct syntactic function tag.

With regard to the syntactic tags we gain a fine surface syntactic analysis in which the first element of the postposition is analyzed as modifier of the element carrying case ((@CM>) or noun modifier function (@NOUN-COMPLEMENT>) and the ending element of the postposition is analyzed as adverbial (@ADLG). As a result of this process the ambiguity rate is reduced in % 2, 6. On the other hand, qualitatively, the analysis is appropriate and it will improve the disambiguation process as well as other applications such as subcategorization pattern studies.

Future plans involve extending both the complex postposition grammar and the grammar for assigning the syntactic functions. The idea is to integrate those grammars in the syntactic analysis process and evaluate the results and improvements.

References

Abney, Steven. (1991). Parsing by chunks. In Principle-Based Parsing. Kluwer Academic Publishers.

Aduriz I. and Díaz de Ilarraza A. (2003). Morphosyntactic disamgiguation and shallow parsing in Computational Processing of Basque. Inquiries into the lexicon-syntax relations in Basque. Bernarrd Oyharçabal (Ed.). University of the Basque Country. Bilbao.

Euskaltzaindia. (1994). *Basque Grammar: First Steps* (in Basque). Euskaltzaindia.

[6] precision = correctly detected postpositions/(correctly detected postpositions + wrong postpositions)
[7] recall = correctly detected postpositions/all postpositions

Hualde J. I. (2002). Regarding Basque postpositions and related matters. In Erramu Boneta: A Festschrift for Rudolf P.G. de Rijk, ed. by Xabier Artiagoitia, Patxi Goenaga & Joseba Lakarra, p. 325-339. Bilbao: Univ. del PaísVasco/Euskal Herriko Unib. Supplements of ASJU 44.

Karlsson, Fred (2006), Constraint Grammar – A Language-Independent System for Parsing Unrestricted Text, Mouton de Gruyter, Berlin.

ViPEr: A Lexicon-Grammar of European Portuguese Verbs [1]

Jorge Baptista

Abstract: This paper presents the current state of ViPEr, the Lexicon-Grammar of European Portuguese verbs, a database with distributional, syntactic and semantic properties of the most frequently occurring verbs. The classification follows the theoretical framework of the Lexicon-Grammar. The paper presents the main linguistic criteria that were adopted in the classification of 5052 frequently occurring verbs, which yield 6,059 different constructions or word senses. The paper concludes with some preliminary results on the application of ViPEr to texts and plans for future work.

1. Introduction

The use of reliable, large-coverage language resources is key to the performance of many Natural Language Processing (NLP) systems. To our knowledge, while some syntactic descriptions of European Portuguese verbs exist, most have not been made publicly available to the NLP community or just consist in human-oriented dictionaries, not having been built for computational processing (FERNANDES 2008, BORBA 1991; BUSSE 1994). On the other hand, partial linguistic studies have produced throughout the last three decades, with major efforts in the late 90s (OLIVEIRA 1981; NASCIMENTO 1997; RODRIGUES 1997), but little, if any, use was made by the NLP community of such data, and very little effort has been addressed to validate or test those, mostly introspective-sourced, and theoretically-oriented, linguistic descriptions. A recent attempt in that direction is that of GOMES (2011), that highlighted the difficulties of the task.

For this project, a practical approach to the lexicon was adopted. For many NLP tasks, but specially for any task where a fine-grained semantic distinction in required of ambiguous lexical forms, being able to identify the meaning of the verb (and of the surrounding elements as well) can be facilitated by the knowledge of the syntactic and semantic constraints the verb imposes on the lexical fulfillment of its argument positions. In particular, the number of verb arguments; their structural and distributional type; the prepositions the verb selects to introduce its essential complements; the main shape-changes that these structures can undergo; and other relevant linguistic information; besides its intrinsic linguistic interest, all this data can be put to use to improve parsing strategies, word-sense disambiguation, question-answer systems, computer-assisted language learning systems, among other applications. Above all, an inventory of basic word senses and their corresponding structures is necessary, and this is the aim of the *ViPEr* project.

This paper presents the current state of the Lexicon-Grammar of European Portuguese verbs. The classification of European Portuguese verb constructions is largely based on the methodology presented by M. GROSS (1975, 1981, 1996) and his collaborators under the Lexicon-Grammar theoretical framework (see LAMIROY (1998) for an overview). The classification proper is directly inspired in the synthesis of LECLÈRE (2002). In the following, the main linguistic criteria that were adopted in the classification are presented. For lack of space, only the most salient classes and properties are presented. The paper concludes with preliminary results from applying this new resource to real texts, and prepares future work.

2. General classification principles

This section presents and briefly discusses the main principles applied to the classification verbal constructions. The classification excludes auxiliary verbs (BAPTISTA *et al.* 2010), support-verbs (BAPTISTA 2005a, RANCHHOD 1990), and operator verbs (M. GROSS 1981). For lack of space, the definition of these verbal constructions is omitted.

The main taxonomical principle is the notion of *simple* (or *elementary*) sentence, which can be defined as the syntactic expression of a semantic predicate. In this sense, adverbial and adjectival (i.e. relative-restrictive) subordinate clauses, as well as coordinated clauses, forming complex sentences, are out of the scope of the classification procedure. The term *simple* (or *elementary*) is to be taken here in the more precise sense of a sentence resulting from the most basic of constraints forming the kernel of a language (HARRIS 1991). Hence, verbs with completive arguments are considered second-order operators and are included in the base set of simple sentences.

Each simple sentence is described following the principle of *maximal projection of a predicate*, *i.e.* including all essential arguments of the predicate. Each semantic predicate is defined by a fixed number of arguments, usually no less than one and no larger than three. Exceptions to this general rule, by virtue of their specialized meaning, are: (i) *meteorological predicates*, forming impersonal constructions e.g. *chover* 'to rain'; part-of-the-day verbs, *e.g. amanhecer* 'to dawn' (class **31I**); and (ii) *object transference* predicates, e.g. *importar* 'import' (**38LT**); a few other verbs, such as verb *apostar* 'to bet' (**10**), with more than two complements. These are, in a way, exceptional predicates and, as they are not very productive, they may all be defined extensionally.

Obligatory complements are always essential arguments. However, even essential arguments can often be reduced in discourse; therefore, it is difficult to ascertain if a given complement should be considered as essential

[1] This paper is dedicated to Christian Leclère.

or circumstantial to that particular predicate. Furthermore, not only some complements expressing so-called circumstances of the predicate should be considered as their essential arguments, *e.g.* the locative complement of *viver* 'to live, reside' (**35LS**) as in *O Pedro vive em Lisboa* 'Peter lives in Lisbon' (place); or the manner complement of *portar-se* (to behave' (**33MV**) in *O Pedro portou-se bem* 'Peter behaved well' (manner).

However, some complements are also transformationally derived from the splitting and reanalysis of larger constituents (BAPTISTA 1997, 2000; GUILLET & LECLÈRE 1992; LECLÈRE 1995), *e.g. O cão mordeu as canelas do carteiro* 'The dog bites the postman heels' = *O cão mordeu as canelas ao carteiro* 'The dog bite the heels to-the postman' = *O cão mordeu o carteiro nas canelas* 'The dog bite the postman on the heels' (**32CL**). In these examples, the part-whole (metonymic) relation between the body-part noun (*Nbp*) *canelas* (legs) and the human noun (*Nhum*) *carteiro* (postman) allows for two other alternative syntactic configurations, where the human noun becomes the dative complement or the direct object, while the body-part noun is the direct object or a locative, respectively. In this case, all sentence forms are considered transformationally equivalent, and a single lexicon-syntactic entry has been construed.

In *neutral* constructions, that is, verbs that are diathetically neuter regarding transitivity and intransitivity, *e.g. engordar* 'to become fat/to put some weight': (1) *O Pedro engordou* vs. (2) *Os doces engordaram o Pedro* 'Peter get-fat-ed/the candies get fat-ed Peter'; preference was given in the classification to the transitive structure (class **32C**), for this being the longest structure. Nevertheless, the transitive construction (2) is considered to be derived from a complex structure with causative operator-verb (GROSS 1981), operating on the intransitive sentence form (1), after this being reduced to an infinitive object sub-clause: (1a) *Os doces fizeram o Pedro engordar* 'The candies made Peter get fat'; these embedding is followed by a subject reversion in the infinitive sub-clause, that moves to a post-verbal position, and its reanalysis as an object of the infinitive verb: (1b) *Os doces fizeram engordar o Pedro* 'The candies made get fat Peter'; and, finally, by *fusioning* the causative operator-verb with the main, intransitive verb, yielding the (superficial) direct transitive structure of (2): *Os doces engordaram o Pedro* 'The candies get fat-ed Peter'. In the lexical entry of these verbs, the N_1 *V* transformational property yielding the intransitive form (1) was then explicitly encoded.

Finally, for intrinsic reflexive constructions, like *suicidar-se* 'to commit suicide' (**31H**), e.g. *O Pedro suicidou-se* 'Peter suicide himself = commit suicide', the reflex pronoun is treated as part of the verb and not as an autonomous constituent, as it can not be zeroed: **O Pedro suicidou* 'Peter suicide', nor replaced by a distributionally free constituent: *O Pedro suicidou o João* 'Peter suicide John'.

This last example could be said ironically, but such discursive phenomena, that is, productive, figurative uses, were disregarded in the classification, since they correspond to the expression of the speakers' creativity, and not to the language basic structure from which they are drawn. On the other hand, conventional figurative uses may give rise to splitting a verb into two lexical entries, as with *ralar* 'grate/worry' *e.g. O Pedro ralou o queijo* 'Peter grated the cheese' (**32C**) *vs. Isso ralava o Pedro* 'That worried Peter' (**4**).

Since there is no general rule to define the maximal projection of a given predicate, and because the semantic and syntactic complexity of the language structure is such, often only through a case-by-case decision had one to proceed. The observation of extensive data from corpora was key to help deciding the most adequate solution for each case.

The classification of verbal constructions may be structured in two layers, based on the structural complexity and semantic compositionality of the sentences: (i) *completive constructions*, i.e. simple sentences whose verb selects at least one sentential argument (a completive, subordinate clause; these have precedence over (ii) *non-completive constructions*, i.e. simple sentences whose verb selects only noun phrases as their arguments. As far as subordinate clauses are concerned, only completive clauses are here considered. All adverbial and adjectival (relative) subordinate clauses are excluded. To make matters more complex, though, completive subordinate clauses are often replaced by complex noun phrases built around predicative nouns, functioning in as much the same way as a completive: *Que o Pedro tenha vindo/a vinda do Pedro entristeceu o João* 'That Peter has come/the coming of Peter has saddened John' (**4**). In some cases, the completive is rarely observed in corpora, but if a propositional content can be semantically defined for a given argument position, which is usually reflected in the predicative nature of the nouns selected for that position, that particular construction has been included in completive constructions, even if only these complex noun phrases are observed in that position. Non-completive constructions, therefore, exclusively present nominal-arguments, whose head noun cannot be a predicative noun.

The preposition introducing the complement(s) is an important taxonomical criterion, since each prepositional construction usually presents only one, invariable, preposition for a given argument position, *e.g. telefonar a* 'phone to' (**33**; RODRIGUES 1997), *gostar de* 'like of' (**8**), *confiar em* 'trust in' (**35R**), *casar com* 'marry with' (**35S**), *ansiar por* 'expect' (**8**). Basic prepositions in European Portuguese are *a* 'to, *com* 'with', de 'of/from' and *em* 'in/at'; locative prepositions have been collapsed under the notation *Loc*; all other prepositions, *v.g. para* 'to/towards', *por* 'by/for', *sobre* 'about/on/over/upon', *etc.*, are collapsed under the same taxon, though they have always been explicitly encoded in the lexical entry. Some constructions present, however, more than one preposition introducing its complement positions, e.g. *lutar com/contra* 'fight with/against' (**35R**): *O Pedro luta contra/com a injustiça* 'Peter fights against/with the injustice'. If the meaning of the sentence cannot be clearly distinguished, nor the distributional constraints differ significantly depending on this preposition variation, a single

lexical-entry has been construed, using the basic preposition (in this case, *com* 'with') for its classification and explicitly registering the other variants in the entry's syntactic properties.

A special case of prepositional construction consists of *symmetric* constructions (BAPTISTA 2005), usually presenting a complement introduced by *com* 'with', *e.g. O Pedro conversou com o João* 'Peter talked with John' (**35S**); *O Pedro confundiu o João com o Paulo* 'Peter mistook John for Paul' (**36S1**); *O Pedro debateu esse assunto com o João* 'Peter debated this topic with John' (**36S2**); and *O Pedro concordou com o João em fazerem isso* 'Peter agreed with John in doing that' (**42S**). Because this set of constructions is also defined by particular syntactic-semantic and transformational properties, it had to be singled out from other prepositional constructions.

A rarer case of preposition alternation consist of constructions where the verb presents not only a direct object structure but also a prepositional one, while the meaning and the distribution do not change: *O Pedro namora a Sara* 'Peter is dating Sara' (**32H**) = *O Pedro namora com a Sara* 'Peter is dating with Sara' (**35S**). In this case, the longer, prepositional construction is given precedence over the direct-transitive construction. The transitive construction is noted by the property [pcz], corresponding to the zeroing of the complement preposition.

The distributional constraints imposed by the verb on its arguments positions (subject and essential complements) are the next major criterion for verb classification. The main distributional classes here considered are:

– *Nhum/Nnhum*: Human/non-human noun opposition. This property is only tested with proper names; animal denoting nouns are excluded, except for specialized verbs such as *animal voices*, e.g. *bramir* 'roar' (exclusively applied to elephants; **31R**). Verb constructions with strictly human arguments (**31H**, **32H**) have priority over those that allow both human and non human (**31R**, **32x**).

– *Npl*: plural noun. This abstract notion describes the distributional constraints of verbs imposing a conceptual plural in a given argument position (NASCIMENTO 1997), e.g. the subject of *abundar* 'abound' (**31PL**), or the object of *dispersar* 'disperse' and *colecionar* 'collect' (**32PL**).

– *Nbp*: body-part nouns. Nouns designating a body-part, that is, nouns that intrinsically imply a metonymic relation, usually with a human noun in the sentence, and hence allow for several types of sentence restructuring (see examples above; BAPTISTA 1997, 2000).

– *Nloc*: locative nouns, i.e. nouns designating places or locations. Locative verbs form a complex syntactic-semantic system and a wide range of syntactic constructions can express the predicates they denote, depending on whether the verb can be defined as a *stative* or *dynamic* predicate, or the particular orientation that can be accorded in the case of *dynamic* verbs. Hence, one considers *source*- and *destination*-oriented dynamic verbs; more rarely, a *traject*-oriented predicate can be considered; *e.g. O Pedro vive em Lisboa* 'Peter lives in Lisbon'(stative; **35LS**); *O Pedro vai a/para Lisboa* 'Peter goes to Lisbon' (dynamic, destination-oriented; **35LD**); *O Pedro veio de Lisboa* 'Peter came from Lisbon (source-oriented; **35LD**); and *O Pedro passou por Lisboa* 'Peter passed through Lisbon' (traject-oriented; **35LD**). Locatives can appear not only in prepositional phrases, as in the examples above, but they can also appear as noun phrases, usually in object position, *e.g. O Pedro atravessou o pátio* 'Peter crossed the yard' (dynamic, traject-oriented; **38L1**). *Locative-Fusioned* verbs can be considered cases of *Fusion* (M. GROSS 1981): for example, an object is fusioned with verbs like *pôr* 'put', *deitar* 'throw' or *meter* 'insert' and its locative prepositional complement becomes the direct object of the resulting verb in *O Pedro envenenou o vinho* 'Peter poisoned the wine'= *O Pedro deitou/meteu/pôs veneno no vinho* 'Peter put poison in/on the wine' (destination-oriented; **38L4**); in another class, the verb is fusioned with the locative: *O Pedro enjaulou o leão* 'Peter caged the lion' = *O Pedro meteu/pôs o leão na jaula* 'Peter put the lion in the cage' (destination-oriented; **38L2**).

– *R*: Constraint direct object constructions: Some verb constructions, without being frozen or idiomatic (BAPTISTA *et al.* 2004; VALE 2008), present such narrow distributional constraints that the general distributional properties stated above fail to capture the precise choice of words for a given syntactic position. For the most part, these are direct transitive constructions where the object must be selected from a very limited word set: *O Pedro estrelou uns ovos* 'Peter fried some eggs' (**32R**). In this example, only *ovo* can fill the object distributional constraints; the overall meaning is clearly compositional; otherwise, number and determiner variation is free. These type of distributional constraint is harder to discover and to describe, so they constitute residual, though large, lexical classes (DIAS *et al.* 2006, LECLÈRE 2002).

Transformational properties are considered, in order to distinguish structurally similar constructions that correspond to different types, semantically homogenous, sets of predicates. Another use of these properties is to allow for the constitution of amenable-sized verb classes, so that if a set becomes too large (usually over 200) it is then advisable to split it in two, more treatable subsets. Here, only some of these operations are presented:

– *Passive*: For example, the large set of direct transitive verbs with measurement nouns allow the constitution of a highly homogenous subclass of *measurement* verbs, since none of the elements of this set allows the *Passive* transformation: *As batatas pesam três quilos* 'The potatoes weight 3 Kg' *vs.* **Três quilos são pesados pelas batatas* '3Kg are weighed by the potatoes' (**32NM**).

– *Fusion* of causative operator verb and adjectival sentences. In another example, the de-adjectival causative verbs consist of a set of direct-transitive verbal constructions that is distinguished by systematically having paraphrases with adjectival basic sentences under a causative-operator verb *e.g. clarificar = tornar* (*mais*) *claro*

'clarify, make clear(er)': *O Pedro clarificou a sua posição* 'Peter clarified his position' = *O Pedro tornou* (*mais*) *clara a sua posição* 'Peter made his position clear(er)' (**32TA**).

Finally, and as it is clear from the above remarks, the meaning of the verb in the simple sentence that constitutes its construction is always present during its classification, and it is a *datum* that the operations the sentence may undergo do not alter. Therefore, one cannot speak of syntactic classification alone, since meaning is always implied. The use of semantic concepts as declarative propositions in formal classification, however, has been avoided, and whenever needed, only as a last resource; these propositions can only be adopted when the intuitions about the meaning involved are highly reproducible. Usually, these intuitions should be backed by clear syntactic properties, as is the case of symmetric constructions (BAPTISTA 2005b), where the special relation between the verb and two of its arguments determines a set of syntactic properties, which are unique to that particular class of predicates.

Only seldom was any semantic concept used so far in the classification here proposed. The concept of **apparition** serves to distinguish the direct-transitive constructions with concrete objects where the object appears after/during the process, *e.g. construir* 'build', class **32A**; as opposed to the predicates where the object preexists the process *e.g. sublinhar* 'underline', class **32C**. (in this later case, the object usually it undergoes some transformation/manipulation). Other properties may have to be devised, especially for the large and very heterogeneous **32C** class.

3. Results and future work

A selection of verbs taken from the CETEMPúblico, a large-sized (approx. 190 million words), journalistic text, European Portuguese corpus (ROCHA & SANTOS 2000) underwent the classification process. The corpus was processed through the STRING NLP chain (MAMEDE *et al.* 2012)[2] for part-of-speech tagging and disambiguation and for verbal chains parsing. The most frequent verbs (frequency above 5 instances in the corpus, in descending order) were studied to determine their basic constructions using the criteria briefly explained above. From an initial list of about 7,000 different verb lemmas, 5,052 have already been classified, yielding 6,059 different lexicon-syntactic entries (or clear-cut verb senses). The classification of the remaining verbs is still on going. While not considered in the classification, 259 support and operator-verbs' constructions (M. Gross 1981) have also been identified. Table 1 presented the breakdown of the ambiguous verbs (ws=word senses):

Table 1. Number of different word senses per lemma

ws	lemmas	ws	lemmas	ws	lemmas	ws	lemmas
1	4293	3	145	5	9	7	3
2	565	4	41	6	8	8	2

Fifty-five lexicon-syntactic classes were established so far (see Appendix). Most of these classes correspond to the French Lexicon-Grammar tables (LECLÈRE 2002) and the same conventional code was adopted for easier comparison. For lack of space these cannot be presented here in full (see BAPTISTA 2012 for a complete overview). As one can see, most verbs have been encoded for only one sense. A very conservative approach was adopted here, describing, at this stage, only the most common word senses, thus the still small number of duplicates. For example, the verb *apontar* 'to point/aim/take note/indicate/signal' yields 4 different lexicon-grammar entries: (**36DT**) *O bandido apontou uma faca ao polícia* 'The bandit aimed a knife at the policeman', (**38LD**) *O Pedro apontou os números premiados num papel* 'Peter took note of the lotary numbers in a piece of paper', (**39**) *O Pedro apontou o João como sério candidato ao prémio* 'Peter indicated John as a serious candidate ' and (**9**) *O Pedro apontou ao João quais os defeitos que devia corrigir* 'Peter signaled to John which issues he should correct'. For each entry, the structural, distributional, semantic and transformational properties were encoded in a single binary matrix, and an illustrative example was provided.

In order to do a preliminary assessment of ViPEr, two small articles were retrieved form the online daily edition of *Público*, with about 1.000 words. From the 85 verb forms, 41 correspond to auxiliaries (modal, temporal and aspectual verb auxiliaries; support and operator verbs; and copula verbs). The output of the STRING system looks like the following sentences:

"A Europa *deve***{dever(VMOD)}** *cumprir***{cumprir(05,35R,32R#)}** os acordos com a maior celeridade possível. Espero**{esperar(06,35R#)}** que a Europa *esteja***{estar(VSUP#)}** a_a altura de_as circunstâncias" , *afirmou***{afirmar(31H,09#)}**.

'Europe must fulfill the agreements as fast as possible. I hope that Europe is up to the job, he said.'

Notice the correct parsing of the modal auxiliary *dever* 'must' (BAPTISTA *et al.* 2010) and the support verb *estar à altura de* (RANCHHOD 1990). In most cases, at least one the verb classes encoded in ViPEr match the actual uses found in these texts. Thus, *cumprir* 'fulfil' corresponds to the **32R** entry, *esperar* 'hope' is an instance of the **06** use, and afirmar is the **09** verbum dicendi (Baptista . However, for the remaining 43 verbs, even in such a small sample, it was possible to find lexical lacuna and many other problems:

[2] https://string.l2f.inesc-id.pt [22/07/2012].

As médias totais *têm***{ter(VOPL#)}** *em conta* os resultados de os alunos internos - aqueles que frequentam**{frequentar(38L1#)}** as aulas até a o final de o ano lectivo e *vão***{ir(35LD#)}** *a exame* com uma classificação interna igual ou superior a 10 - e as de os externos , que *anularam*{anular**(32TA#)}** a *matrícula* e se autopropuseram**{autopropuserir(No viper data)}** a exame .

'The overall averages take into account the results of internal students - those who attend classes until the end of the school year and take a final exam with an internal grade equal to or greater than 10 - and the results of the external students, which cancelled their registration and presented themselves to exam.'

Thus, *ter em conta* and *ir a exame* may be considered frozen sentences (classes **CNP2** and **CP1**, respectively; BAPTISTA *et al.* 2004); this type of structure has not been integrated in the system yet. The expression *anular matrícula* can be considered a support verb construction, but this use had not been identified yet. The only verb for which there is no entry in ViPEr is also an unknown, derived word, formed on the base verb *propor* 'propose' with prefix *auto-* 'self'. An extension of the morphologic module LexMan (Diniz & Mamede 2011) of the STRING system will enable in the near future (FREITAS 2012).

In future work, the lexical coverage of ViPEr needs to be further assessed. All the verbs of a 290K words, POS-disambiguated corpus (RIBEIRO 2004, DINIZ & MAMEDE 2011) has been annotated with the ViPEr classes and is currently being manually reviewed. Verb forms associated to 136 lemmas had not a ViPEr tag. Many of them correspond to lemmas that were changed by the new Portuguese orthographic reform, but only 46 had to be added to the database. This corpus, manually annotated with the verb senses, will enable the use and evaluation of rule-based and machine-learning techniques in word-sense disambiguation tasks (Travanca 2012).

Acknowledgments

Research for this paper was partially funded by project REAP.PT (CMU-Portugal/HuMach/0053/2008). The author wishes to thanks, in chronological order, the students who, in different ways, have contributed to this work: David Monteiro (UNL), Rui Santos (UALG) to Tiago Travanca (IST) and Cláudio Diniz (INESC-ID Lisboa) for their help in the tagging of the corpus; and, finally, to Nuno Mamede (IST/INESC-ID Lisboa), for his constant support and friendship.

References

BAPTISTA, Jorge. 1997. Conversão, nomes parte-do-corpo e restruturação dativa. In Castro, Ivo (ed.) *Actas do XII Encontro da APL*, Lisboa, 1996, p. 51-59. Lisboa: APL/Colibri.

BAPTISTA, Jorge. 2000. Boby-part nouns and local grammars. *Revue Informatique et Statistique dans les Sciences Humaines*, 36, p. 53-66. Liège: Univ. Liège.

BAPTISTA, Jorge; CORREIA, Anabela; FERNANDES, Graça. 2004. Frozen Sentences of Portuguese: Formal Descriptions for NLP. Workshop on Multiword Expressions: Integrating Processing, Barcelona, Spain, July 26, 2004, p.72-79. Barcelona: ACL.

BAPTISTA, Jorge. 2005a. *Sintaxe dos Predicados Nominais com* SER DE. Lisboa: FCT/FCG.

BAPTISTA, Jorge. 2005b. Construções simétricas: argumentos e complementos. in Figueiredo, O; and Rio-Torto, Graça; and Silva, F. (eds.). *Estudos de Homenagem a Mário Vilela*, p. 353-367. Porto: Campo das Letras.

BAPTISTA, Jorge. 2010. *Verba dicendi*: a structure looking for verbs. in Nakamura, Takuya; Laporte, Éric; Dister, Anne; Fairon, Cédrick (eds.). *Les Tables. La grammaire du français par le menu. Mélanges en hommage à Christian Leclère*. Cahiers du CENTAL 6, p. 11-20. Louvain-la-Neuve: CENTAL.

BAPTISTA, Jorge; MAMEDE, Nuno; GOMES, Fernando. 2010. Auxiliary verbs and verbal chains in European Portuguese. *Computational Processing of the Portuguese Language*. LNAI/LNCS 6001, p.

BAPTISTA, Jorge. 2012. *Verb Classification Guidelines*. Technical Report. Lisboa: L2F-Spoken Language Lab/INESC-ID Lisboa.

BOONS, Jean-Paul; GUILLET, Alain & LECLÈRE, Christian. 1976a. *La structure des phrases simples en français: Les contructions intransitives*. Paris: Droz.

BOONS, Jean-Paul; GUILLET, Alain & LECLÈRE, Christian. 1976b. *La structure des phrases simples en français: Les contructions transitives*. Rapport de Recherches 10. Paris: LADL.

BORBA, Francisco. 1991. *Dicionário Gramatical de Verbos do Português Contemporâneo do Brasil*. São Paulo, Brasil: UNESP.

BUSSE, Winfried. 1994. *Dicionário Sintáctico de Verbos*. Coimbra: Livraria Almedina.

CHABY, Teresa. 1997. *Construções de segmentação. Propriedades léxico-sintácticas*. MA Thesis). Universidade de Lisboa - Faculdade de Letras.

DIAS, Maria Carmelita P.; LAPORTE, Éric & LECLERE, Christian. 2006. Verbs with very strictly selected complements. In: *Collocations and Idioms: The First Nordic Conference on Syntactic Freezes*. Finland: University of Joensuu.

DINIZ, Cláudio; MAMEDE, Nuno. 2011. *LexMan - Lexical Morphological Analyser*. Technical Report. Lisboa: L2F/INESC-ID Lisboa.

FERNANDES, Francisco. 2008. *Dicionário de Verbos e Regimes* (45th ed). São Paulo, Brasil: Globo.

GOMES, Fernando. 2011. *Validation of Lexical-Syntactical Matrices*. MSc Dissertation. Instituto Superior Técnico, Universidade Técnica de Lisboa.

GROSS, Gaston. 1996a. *Les expressions figées en français*. Paris: Ophrys.

GROSS, Maurice. 1975. *Méthodes en Syntaxe*. Paris: Hermann.

GROSS, Maurice. 1981. Les bases empiriques de la notion de prédicat sémantique. *Langages*, 63, 7–52.

GROSS, Maurice. 1982. Une classification des phrases « figées » du français. *Revue Québécoise de Linguistique*, 12-2, p. 16

GROSS, Maurice. 1996b. Lexicon-Grammar. in Brown, K. and Miller, J. (eds.), *Concise Encyclopedia of Syntactic Theories*. Cambridge: Pergamon, p. 244–259.

GUILLET, Alain & LECLÈRE, Christian. 1981. Restructuration deu group nominal. *Langages* 63, p. 99–125.

GUILLET, Alain & LECLÈRE, Christian. 1992. *La structure des phrases simples en français: 2 - Les contructions transitives locatives*. Paris: Droz.

HARRIS, Zellig S. 1991. *A Theory of Language and Information. A Mathematical Approach*. Oxford: Clarendon.

LAMIROY, Béatrice (ed.) 1998. *Le Lexique-Grammaire*. *Travaux de Linguistique,*Vol. 37. Duculot.

LECLÈRE, Christian. 1995. Sur une restructuration dative. *Language Research*, p. 179–198.

LECLÈRE, Christian. 2002. Organization of the Lexicon-Grammar of French Verbs. *Linguisticae Investigationes*, 25-1, p. 29–48.

MAMEDE, Nuno; BAPTISTA, Jorge; DINIZ, Cláudio. 2012. STRING - An Hybrid Statistical and Rule-Based Natural Language Processing Chain for Portuguese. Proceedings of PROPOR'2012, Coimbra, Portugal. http://www.propor2012.org/demos/DemoSTRING.pdf (on-line)

OLIVEIRA, Maria Elisa. 1981. *Syntaxe des verbes psychologiques du Portugais*. Lisboa: INIC.

RIBEIRO, Ricardo. 2004. *Anotação Morfossintáctica Desambiguada do Português* (MSc Thesis). Lisboa: IST/UTL.

RANCHHOD, Elisabete. 1990. *Sintaxe dos predicados nominais com* ESTAR. Lisboa: INIC.

ROCHA, Paulo & SANTOS, Diana. 2000 (November). CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. *Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada, PROPOR'2000*, p. 131–140.

RODRIGUES, Rosinda. 1997. *Determinação das propriedades sintácticas das frases simples com a forma* N0 V a N1. Master thesis, Universidade de Lisboa - Faculdade de Letras.

TRAVANCA, Tiago. 2012. *Disambiguation of Verb Senses*. MA Project. Lisboa: IST.

VALE, Oto. 2001. *Expressões Cristalizadas do Português do Brasil: Uma Proposta de Tipologia*. Araraquara: UEJMF(PhD Thesis).

Appendix. ViPEr : Verb classes of European Portuguese.

Class	Count	Structure	verb	example
01I	13	$QueF_0$ *V*	*adiantar* 'matter'	*Não adianta fazer isso* 'doesn't matter to do that'
01T	61	$QueF_0$ *V* $QueF_1$	*evitar* 'avoid'	*Fazer isso evitava ter de fazer aquilo* 'To do this avoids having to do that'
02	15	$QueF_0$ *V* $Prep_1$ $QueF_1$	*obrigar* 'force'	*Isto obriga a fazer aquilo* 'This forces to do that'
03	1	N_0 *V* N_1 (Loc_1 $Nloc_1$) $Vinf^d$	*mandar* 'send'	*O Pedro mandou o João à loja comprar café* 'Peter sent John to the shop to buy some coffeee'
04	330	Nnr_0 *V* $Nhum_1$	*irritar* 'irritate'	*Isso irrita o Pedro 'That irritates Peter'*
05	24	Nnr_0 *V a* $Nhum_1$	*agradar* 'please'	*Isso agrada ao Pedro* 'That pleases Peter'
06	221	$Nhum_0$ *V* $QueF_1$	*pensar* 'think'	*O Pedro pensa que o João é inteligente* 'Peter thinks that John is intelligent'
07	46	$Nhum_0$ *V a* $(Vinf^0)_1$	*aprender* 'learn'	*O Pedro aprendeu a fazer isso* 'Peter learn to do that
08	96	N_0 *V* $Prep_1$ $QueF_1$	*depender* 'depend'	*O Pedro dependia da autorização do João* 'Peter depended on John's authorization'
09	162	$Nhum_0$ *V* $QueF_1$ *a* $Nhum_2$	*dizer* 'say	*O Pedro disse ao João que está feliz* 'Peter said to John that [he] is happy'
10	4	$Nhum_0$ *V* $QueF_1$ $Prep_2$ $Nhum_2$	*apostar* 'bet'	*O Pedro apostou com o João que ganhava* a corrida 'Peter bet with John that [he] would win the race'

Class	Count	Structure	verb	example
11	43	N_0 *V* N_1 *a* $QueF_2$	*obrigar* 'force'	*O Pedro obrigou o João a fazer isso* 'Peter forced John to do that'
12	34	N_0 *V* N_1 *de*$_2$ $(Vinf^d)_2$	*impedir* 'prevent'	*O Pedro impediu o João de fazer isso* 'Peter prevented John from doing that'
13	21	N_0 *V* N_1 *de*$_2$ $QueF_2$	*informar* 'inform'	*O Pedro informou o João de que ia fazer isso* 'Peter informed John that [he] was going to do that
14	11	N_0 *V* N_1 $Prep_2$ $N_{2,1}$ (*a* $Nhum_3$)	*pagar* 'pay'	*O Pedro pagou 20€ por isso ao João* 'Peter payed 20€ to John for that'
16	9	N_0 *V* $QueF_1$ $Prep_2$ $QueF_2$	*deduzir* 'deduce'	*O Pedro deduziu isso daquilo* 'Peter deduced this from that'
31CL	20	Nbp_0 *V*	*suar* 'sweat	*Os pés do Pedro suam* 'Peter's feet sweat'
31H	287	$Nhum_0$ *V*	*espirrar* 'sneeze	*O Pedro espirrou* 'Peter's sneezed'
31I	24	*0 V*	*chover* 'rain'	*Chove* '[it] rains'
31PL	11	Npl_0 *V*	*proliferar* 'proliferate'	*As bactérias proliferam* 'Bacteria proliferate'
31R	275	(*Nhum*+*Nnhum*)$_0$ *V*	*morrer* 'die'	*O Pedro morreu* 'Peter died'
32A	35	N_0 *V* $Nnhum_1$ {apparition}	*preparar* 'prepare'	*O Pedro preparou o almoço* 'Peter prepared lunch'
32C	1,148	N_0 *V* $Nnhum_1$ *{apparition}	*ler* 'read'	*O Pedro leu um livro* 'Peter read a book'
32CL	193	N_0 *V* Nbp_1	*partir* 'break'	*O Pedro partiu um braço* 'Peter broke an arm'
32CV	13	N_0 *N-v* N_1 [= *converter* N_1 *em N*]	*cristalizar* 'crystalize	*Isso cristalizou o açúcar* 'That crystalized the sugar'
32H	440	$Nhum_0$ *V* $Nhum_1$	*amar* 'love'	*O Pedro ama a Ana* 'Peter loves Ana'
32NM	29	N_0 *V* $Nmeas_1$	*medir* 'measure'	*O Pedro mede 1,80 m* 'Peter measures 1.80 m'
32PL	55	N_0 *V* Npl_1	*ordenar* 'order'	*O Pedro ordenou os alunos* 'Peter ordered the students'
32R	243	N_0 *V* Nc_1	*estrelar* 'fry'	*O Pedro estrelou um ovo* 'Peter fried an egg'
32TA	289	N_0 *Adj-v* N_1 [V=*tornar Adj* N_1]	*amaciar* 'soften'	*O sabonete amacia a pele* 'The soap softens the skin'
33	77	N_0 *V a* N_1	*telefonar* 'phone'	*O Pedro telefonou ao João* 'Peter phoned to John'
33MV	4	N_0 *V Advmanner*	*portar-se* 'behave'	*O Pedro portou-se mal* 'Peter behaved badly'
33NM	1	N_0 *V* $Prep_1$ $Nmeas_1$	*ascender* 'ascend'	*O PIB ascende a 1B$* 'example'
34	53	N_0 *V* $Prep_1$ N_1 $Prep_2$ N_2	*saber* 'know'	*O Pedro sabe muito de futebol* 'Peter knows a lot about football'
35LD	273	N_0 *V-dyn* Loc_1 $Nloc_1$	*entrar* 'enter'	*O Pedro entrou na sala* 'Peter entered into the room'
35LS	40	N_0 *V-stat* Loc_1 $Nloc_1$	*viver* 'live'	*O Pedro vive em Lisboa* 'Peter lives in Lisbon'
35R	173	N_0 *V* $Prep_1$ N_1	*confiar* 'trust'	*O Pedro confia no João* 'Peter trusts in John
35S	112	N_0 *V com* N_1	*conversar* 'talk'	*O Pedro conversou com o João* 'Peter talked with John'
36DT	108	$Nhum_0$ *V* $Nobj_1$ *a* $Nhum_2$	*dar* 'give'	*O Pedro deu um livro ao João* 'Peter gave a book to John'
36R	89	N_0 *V* N_1 $Prep_2$ N_2	*transformar* 'transform'	*O Pedro transforma barro em arte* 'Peter transforms clay in art'
36S1	82	N_0 *V* $Nobj_1$ *com*$_2$ $Nobj_2$	*misturar* 'mix'	*O Pedro mistura o açúcar com a farinha* 'Peter mixes sugar with flour'
36S2	15	$Nhum_0$ *V* $Nobj_1$ *com*$_2$ $Nhum_2$	*combinar* 'arrange'	*O Pedro combinou com o João uma ida ao cinema* 'Peter arranged with John to go to the movies'
36TA	6	N_0 *Adj-v* N_1 $Prep_2$ N_2 [V=*tornar Adj* N_1]	*adequar* 'adjust'	*O Pedro adequa o discurso ao público* 'Peter adjusts his talk to the public'
38L1	193	N_0 *V* $Nloc_1$	*invadir* 'invade'	*O Pedro invadiu a sala* 'Peter invaded the room'
38L2	35	N_0 *Nloc-v* $Nobj_1$ [V=*put in Nloc*]	*enjaular* 'cage'	*O Pedro enjaulou o leão* 'Peter caged the lion'
38L3	9	$Nloc_0$ *V* $Nobj_1$	*encerrar* 'enclose'	*A jaula encerrava a fera* 'The cage enclosed the beast'
38L4	86	N_0 *Nobj-v* $Nloc\text{-}d_1$ [V=*pôr Nobj*]	*envenenar* 'poison'	*O Pedro envenenou a bebida* 'Peter poisoned the drink'
38L5	10	N_0 *Nobj-v* $Nloc\text{-}s_1$ [V=*tirar Nobj*]	*desengordurar* 'ungrease'	*O Pedro desengordurou o prato* 'Peter ungreased the dish'
38LD	73	N_0 *Vdyn* N_1 $Loc\text{-}d_2$ $Nloc_2$	*pousar* 'put'	*O Pedro pousou o livro na mesa* 'Peter put the book on the table'
38LS	73	N_0 *Vdyn* N_1 $Loc\text{-}s_2$ $Nloc_2$	*retirar* 'remove'	*O Pedro retirou o livro da mesa* 'Peter removed the book from the table'
38LT	41	N_0 *Vdyn* N_1 $Loc\text{-}s_2$ $Nloc_2$ $Loc\text{-}d_3$ $Nloc_3$	*transferir* 'transfer'	*O Pedro transferiu o livro daqui para ali* 'Peter transferred the book from here to there'
38LR	5	N_0 *Vstat* N_1 $Prep_2$ N_2	*situar* 'place'	*O Pedro situou a casa no mapa* 'Peter placed the house in the map'
38PL	57	N_0 *V* N_1 $Prep_2$ Npl_2	*dividir* 'divide'	*O Pedro dividiu o bolo em fatias* 'Peter divided the cake in tranches'
38R	7	N_0 *V* N_1 $Loc\text{-}d_2$ N_2	*remeter* 'send'	*O Pedro remeteu o João para a Ana* 'Peter sent John to Anna'
38TD	9	N_0 *V* N_1 $Loc\text{-}s_2$ N_2	*receber* 'receive'	*O Pedro recebeu uma prenda do João* 'Peter received a gift from John'
39	58	N_0 *V* N_1 ($Prep_2$) N_2	*nomear* 'appoint'	*O Pedro nomeou o João (como) seu representante* 'Peter appointed John (as) his representative'
40	11	N_0 *V* $Prep_1$ N_1 $Prep_2$ N_2	*dar* 'hit'	*O Pedro deu com um livro na cabeça do João* 'Peter hit with a book on John's head'
41	11	N_0 *V* $Prep_1$ N_1 $Prep_2$ $QueF_2$	*apelar* 'appeal'	*O Pedro apelou ao João para que fizesse isso* 'Peter appealed to John for [him] to do that'
42S	5	N_0 *V com* N_1 $Prep_2$ N_2	*comungar* 'commune'	*O Pedro comungava com o João dos mesmos ideais* 'Peter communed with John from the same ideals'
Total:	6,059			

Notations

N_0, N_1, N_2, N_3: subject and complements; *Prep*: preposition; *Adj*: adjective; *Adv*: adverb; *Nhum*: human noun; *Nnhum*: non-human nouns; *Nbp*: body-part noun; *Npl*: plural noun; *Nloc*: locative noun; *Nnr*: non-constraint noun; *Nobj*: "object" noun (semantic role);*QueF*: completive sub-clause; *Loc*: locative preposition; *V*: verb; *Vdyn*: dynamic locative verb; *Vstat*: static locative verb; *-v*: verb ending.

Brazilian Portuguese nominal predicates with 'fazer' (make/do): sports

Cláudia Dias de Barros, Oto Araújo Vale

Abstract: This paper presents a study on the nominal predicates formed with the support verb (SV) 'fazer' (make/do) and a sport noun, and shows some syntactic-semantic properties of these structures. The nominal predicates (SV 'fazer' and sport) and their syntactic-semantic properties are extracted from a Brazilian Portuguese corpus called PLN.Br FULL (BRUCKSCHEIN et al., 2008). These study results can be used in future work for creating the lexical resource NomBank.Br, a database with Brazilian Portuguese predicative nouns and their argument structures, based on the NomBank project (MEYERS et al., 2004)

1. Introduction

Many studies have been carried out on nominal predicates with some support verbs, like: dar (give), estar/ser (be), ter (have) and fazer (make/do) for French and European Portuguese.

For Brazilian Portuguese (BP), there are some studies about the verb as the predicator in a sentence, such as CANÇADO (1995; 2010), CAMACHO (1996) and BORBA (2002). However, the exhaustive study of nominal predicates in BP is not a very common field of study.

Example (1) presents a sentence with a nominal predicate in which the noun is the predicator instead of the verb:

(1) "Ana fez uma doação de sangue."
(Ana made a blood donation.)[1]

In this case, the verb 'fez' (made) isn't the predicator of the sentence. It only provides, through inflection, some details on the event semantics, such as aspect, mood or tense, i. e. it is a support verb (SV). The main semantic content is provided by the noun 'doação' (donation).

The correct identification of the predicator and its arguments is very useful for the correct Semantic Role Labeling (SRL) (GILDEA and JURAFSKY, 2002), a task that has been increasingly developing in Natural Language Processing (NLP) area, since it is necessary to correctly identify a sentence predicator (a verb or a noun) to properly label the semantic roles, as in (2):

(2) "Ele$_{\text{Agente}}$ **faz$_{\textbf{Vsup}}$** coleção$_{\text{Predicador}}$ de chaveiros$_{\text{Tema}}$"
(He$_{\text{Agent}}$ makes$_{\text{SV}}$ collection$_{\text{Predicator}}$ of keychains$_{\text{Theme}}$)

Thus, in this context, this paper presents a study on the nominal predicates formed with the SV 'fazer' (make/do) and a sport noun, and shows some syntactic-semantic properties of these structures.

All the analysis are based on the Lexicon Grammar Theory (GROSS, 1975), because this theory kind of data formalization can be used in researches on automatic semantic role labeling (SRL), for example.

The nominal predicates (SV 'fazer' and sport) and their syntactic-semantic properties are extracted from a Brazilian Portuguese corpus called PLN.Br FULL (BRUCKSCHEIN et al., 2008) through the tool Unitex (PAUMIER, 2002).

The corpus has been analyzed and it was noticed that some nouns could be grouped in a semantic class (the sport nouns), because they share some features and syntactic regularities, presented on section 4.

This paper is structured as follow: section 2 presents some works related to this study; in section 3, there is the methodology used in this study; section 4 presents the data analysis and in section 5, the conclusion.

2. Related works

There are some researches on nominal predicates for European Portuguese, as RANCHHOD (1990), which worked with the SV 'estar' (be); FREIXO (1992), ATHAYDE (2001) and CHACOTO (2005), which worked with the SV 'fazer' (make/do); BAPTISTA (2005), which studied the nominal predicates with 'ser de' (be).

GIRY-SCHNEIDER (1978) and LABELLE (1984) have worked with nominal predicates with SV 'faire' (make/do) and 'avoir' (have) for French.

LA FAUCI and MIRTO (2003) have studied the four syntactic types of the Italian verb 'fare' (make/do) and the differences between them. RASSI (2008) have done a similar study with the verb 'fazer' (make/do) for Brazilian Portuguese. In this research, the author presents an extensive study of the verb 'fazer' (make/do) acting and uses, finding that this verb can occur in BP as a full verb, hiperverbo, operator-causative, support verb, vicarious verb and a multiword expression.

The difference between these researches and the study presented in this paper is that this one focuses on predicate nouns, especially sport and not on the SV 'fazer' (make/do). This verb acts like a filter to find the predicate nouns in the corpus.

[1] The English translations are just approximate from Brazilian Portuguese.

3. Methodology

This paper addresses the following research questions:

1) How is it possible to identify the nominal predicates with the SV 'fazer' (make/do)?;
2) What is the syntactic structure of these constructions like?;
3) Is it possible to consider sport predicative nouns a distinctive class of predicates (for its regularity, syntactic properties) within the general set of nouns that select this SV?

In the first step of the research the predicate nouns (sport) that co-occurred with SV 'fazer' (make/do) were extracted from a BP corpus, called PLN.Br FULL (BRUCKSCHEIN et al., 2008), which contains 103.080 texts, with 29.014.089 *tokens*, from Folha de São Paulo, a Brazilian newspaper, from 1994 to 2005. The predicate nouns for this study have been extracted from the files of 2003 and 2004 (11.086.271 *tokens*, 15.189 <fazer>) with Unitex, a corpus processing tool. The predicate nouns extraction was done through a manual identification, based on Unitex concordances.

In order to recognize the SV 'fazer' (make/do) and therefore the predicate nouns from the corpus, some rules have been established, like:

a) The construction 'SV + predicate noun' can be replaced by a full verb (fazer uma doação – doar – make a donation – to donate);
b) The SV 'fazer' can be replaced by another SV without changing the meaning (fazer um trabalho – realizar um trabalho – do a work – perform a work);
c) The predicate noun is an abstract noun (fazer sentido – make sense).

The second research step was the 'SV + predicate noun' syntactic-semantic properties identification.

In the third step, the binary table (the lexical entries and their properties) has been elaborated.

4. Data analysis

In order to extract the predicate nouns from the corpus, a search for '<fazer>' (the canonical form) and a word on its left and three words on its right was carried out. Then, all the identified predicate nouns extracted from the corpus (815) were transferred to an Excel file and only sports (26) were selected for the study in this paper.

Table 1 presents these nouns and the syntactic elements that can occur with them in a sentence with SV 'fazer' (make/do). The plus signal (+) indicates the presence of a feature and the minus signal (-) indicates its absence. At this point of the study, the properties signaled in this table were obtained introspectively, but in a second moment they will be validated in the *corpus.*

	N0=Nhum	ArtDef	ArtIndef	Absent Article	Modif	N1 Singular	N1 Plural	Preposition	N1=N-hum	Passive	Examples
Academia (gym)	+	-	-	+	-	+	-	-	-	-	João faz academia (John makes gym)
Acrobacia (acrobatics)	+	+	+	+	+	+	+	de	+	+	João faz acrobacia (John does acrobatics)
Arvorismo (tree climbing)	+	-	-	+	-	+	-	-	-	-	João faz arvorismo (John does tree climbing)
Atletismo (athletics)	+	-	-	+	-	+	-	-	-	-	João faz atletismo (John makes athletics)
Caminhada (hiking)	+	+	+	+	+	+	+	em/ de	+	+	João faz caminhada (John is walking)
Capoeira	+	-	-	+	-	+	-	-	-	-	João faz capoeira
Contorcionismo (contortion)	+	-	-	+	-	+	+	-	-	-	João faz contorcionismo (John does contortion)
Corrida (jogging)	+	+	+	+	+	+	-	-	-	+	João faz corrida (John is running)
Escalada (climbing)	+	-	-	+	-	+	+	-	-	+	João faz escalada (John is climbing)
Esporte (sport)	+	+	+	+	+	+	+	em/ de	+	+	João faz esporte (John makes sport)
Exercício (exercise)	+	+	+	+	+	+	+	em/ de	+	+	João faz exercício (John exercises)
Flexão (push up)	+	-	-	+	-	+	+	em/ de	+	-	João faz flexão (John pushes up)
Ginástica (gymnastics)	+	-	-	+	-	+	-	-	-	-	João faz ginástica (John does gymnastics)
Hidroginástica (hydrogymnastics)	+	-	-	+	-	+	-	-	-	-	João faz hidroginástica (John does hydrogymnastics)
Malabarismo (juggling)	+	-	-	+	-	+	+	em/ de	+	+	João faz malabarismo (John juggles)
Mergulho (scuba diving)	+	+	+	+	+	+	+	-	-	+	João faz mergulho (John makes diving)
Musculação (weight lifting)	+	-	-	+	-	+	-	-	-	-	João faz musculação (John does weight lifting)
Pedalada (cycling)	+	+	+	+	+	+	-	em	+	-	João faz pedalada (John makes cycling)
Pesca (fishing)	+	+	+	+	+	+	+	em/ de	+	+	João faz pesca esportiva (John makes sport fishing)
Rapel (rappelling)	+	-	-	+	-	+	-	-	-	-	João faz rapel (John makes rapelling)
Rachão (training)	+	-	-	+	-	+	+	em	+	-	João faz rachão (John trains)
Salto (jumping)	+	-	-	+	-	+	+	-	-	-	João faz salto (John is jumping)
Snowboarding	+	-	-	+	-	+	-	-	-	-	João faz snowboarding (John makes snowboarding)
Tiro (shooting)	+	-	-	+	-	+	+	-	-	-	João faz tiro) (John makes shooting)
Trilha (tracking)	+	+	+	+	+	+	+	em	+	+	João faz uma trilha na floresta (John makes a forest trail)
Voo (flight)	+	-	-	+	-	+	+	em/ de/a	+	-	João faz voo livre (John does a hang glider flight)

Table 1: The Sport table

Through the data analysis, it is possible to interpret that:

a) There are two semantic subclasses of sport nouns: the nouns of more or less conventional sports proper as 'arvorismo' (tree climbing), 'atletismo' (athletics), 'capoeira', 'corrida' (jogging), 'escalada' (climbing), 'mergulho' (scuba diving), 'fishing' (pesca), 'rapel' (rappelling), 'salto' (jumping), 'snowboarding', 'tiro' (shooting) and 'voo' (flight). These nouns represent activities practiced in a regular/habitual basis and some of them have also a parallel use as a non-conventional activity, as corrida (jogging). The second class is formed by the sport activities as 'academia' (gym), 'acrobacia' (acrobatics), 'caminhada' (hiking), 'contorcionismo' (contortion), 'esporte' (sport), 'exercício' (exercise), 'flexão' (push up), 'ginástica' (gymnastics), 'hidroginástica (hydrogymnastics), 'malabarismo' (juggling), 'musculação' (weight lifting), 'pedalada' (cycling), 'rachão' (training), 'trilha' (tracking).

b) All the sport nouns have a human subject (N0=Nhum):
(1) "A menina quer fazer contorcionismo"
(The girl wants to do contortion);

c) There are some nouns that admit neither a definite nor an indefinite determiner. These nouns aren't followed by a modifier, e.g. an adjective, and are only used in the singular, except for 'contorcionismo' (contortion), 'escalada' (climbing), 'malabarismo' (juggling), 'musculação' (weight lifting), which can be used in the plural as well. They don't admit a complement with a preposition. Some examples of these nouns are: 'academia' (gym), 'atletismo' (athletics), 'arvorismo' (tree climbing), 'capoeira', ginástica (gymnastics), 'hidroginástica' (hydrogymnastics), 'rapel' (rapelling), and snowboarding:
(2) "Os meninos fazem atletismo"
(The boys do athletics);

d) The predicate nouns that admit a definite determiner also admit a modifier and can be used in the singular or in the plural: 'acrobacia' (acrobatics), 'caminhada' (hiking), 'esporte' (sport), 'exercício' (exercise), 'mergulho' (scuba diving), 'pesca' (fishing), 'trilha' (track). The exceptions are 'corrida' (jogging) and 'pedalada' (cycling), which can just be used in the singular:
(3) "Fizemos a caminhada mais longa de nossas vidas."
(We went on the longest hiking trip of our lifes.);

e) Some nouns can have a complement with a preposition and another noun (N1). The prepositions used can be: em (in), de (of), com (with), a (to), like 'acrobacia' (acrobatics), 'caminhada' (hiking),'esporte' (sport), 'exercício' (exercise), 'flexão' (push up), 'malabarismo' (juggling), 'pedalada' (cycling), 'pesca' (fishing), 'rachão' (training), 'trilha' (track), 'voo' (flight):
(4) "João faz exercício de impacto."
(John makes impact exercices.)

f) There are some nouns that can only be used without a determiner or a modifier, but they admit a complement with a preposition and a N1, like: 'flexão' (bend), 'rachão' (training), 'salto' (jumping), 'tiro' (shooting), and 'voo' (flight).
(5) "Ele faz voo de asa-delta" (He does a gliding flight).

g) Some nouns can be used in passive like: 'acrobacia' (acrobatics), 'caminhada' (hiking), 'corrida' (jogging), 'esporte' (sport), 'exercício' (exercise), 'malabarismo' (juggling), 'pesca' (fishing), 'trilha' (tracking). These nouns also can be followed by a modifier:
(6) "Corrida com obstáculos é feita por João"
(Obstacle course is made by John)
Real sport activities cannot undergo passive.

h) The SV 'fazer' (make/do) can be replaced by the verb 'praticar' (practice) in constructions with sport nouns.
(7) "John makes sports" → "John practices sports".

i) No sport nouns allows for the reduction of the support verb and the formation of a complex noun phrase ("O Pedro faz/pratica judo" "Peter does judo"). In fact, the relative ("O judô que o Pedro faz/pratica" "The judo that Peter does") has a different interpretation, no longer associated to a regular activity interpretation and closer to a manner-pivoting relative (that is, similar to the interpretation of "A maneira como o Pedro executa as técnicas de judô" "The manner as Peter executes the judo techniques").

5. Conclusion

This paper presented a study on a specific predicate noun semantic class: sport. Some syntactic features of the sentences with SV 'fazer' (make/do) have been presented, as: the type of subject, the determiners presence or absence, a modifier presence or absence, if the predicate nouns can be used in the singular or in the plural, the possibility to have a complement (preposition + N1) and the possibility to have a passive.

The data has been formalized in a binary table as the Lexicon Grammar Theory (GROSS, 1975) suggests.

This study aims at contributing to the Brazilian Portuguese description, through the predicate noun study under the Lexicon Grammar Theory perspective. This description can be used in future work on the automatic identification of predicate nouns with SV 'fazer' (make/do).

The correct predicate noun identification can contribute to the improvement of Semantic Role Labeling (SRL), which is a growing NLP research area.

References

ATHAYDE, Maria Francisca (2001), *Construções com verbo-suporte (funktionsverbgefüge) do português e do alemão.* Number 1 in Cadernos do CIEG Centro Interuniversitário de Estudos Germanísticos. Universidade de Coimbra, Coimbra, Portugal.

BAPTISTA, Jorge (2005), *Sintaxe dos predicados nominais com ser de*. Lisboa, F.Calouste Gulbenkian/FCT.

BORBA, Francisco da Silva (1996), *Uma gramática de valências para o português*, São Paulo, Ática, 199 p.

BRUCKSCHEN Mírian et al. (2008), *Anotação linguística em XML do corpus PLN-BR*. Série de Relatórios do NILC (NILC-TR-09-08), São Carlos - SP, Junho 2008, 39 p.

CAMACHO, Roberto Gomes (1996), O papel da estrutura argumental na variação de perspectiva. In: KOCH, Ingedore Vilaça (Org.). *Gramática do Português Falado.* Campinas: Editora da UNICAMP/FAPESP, v. 6, p. 253-274.

CANÇADO, Márcia (1995), *Verbos psicológicos: a relevância dos papéis temáticos vistos sob a ótica de uma semântica representacional.* 1995. Tese (Doutorado) – Instituto de Estudos da Linguagem, Universidade de Campinas, Campinas.

CANÇADO, Márcia (2010), Verbal alternations in Brazilian Portuguese: a lexical semantic approach. *Studies in Hispanic and Lusophone Linguistics*, V.3:1.

CHACOTO, Lucília (2005), *O verbo fazer em construções nominais predicativas*, Dissertação de Doutoramento em Linguística (Especialidade: Sintaxe), Universidade do Algarve.

GILDEA, Daniel, JURAFSKY, Dan (2002), Automatic labeling of semantic roles. *Computational Linguistics*, 28 (3): 245-288.

GROSS, Maurice (1975), *Méthodes en syntaxe.* Paris: Hermann.

LA FAUCI, Nunzio; MIRTO, Ignazio M. (2003), *Fare: elementi di sintassi*, Pisa, Itália, Edizioni, ETS.

MARCUS, Mitchell; SANTORINI, Beatrice; MARCINKIEWICZ, Mary Ann (1993), Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, v. 19, n. 2, p. 313-330.

NEVES, Maria Helena de Moura (2000), *Gramática de usos do português*, São Paulo, Editora UNESP.

PAUMIER Sébastian (2002), *Unitex: manuel d'utilisation*, research report. França, University of Marne-la-Vallée, 200 p.

RANCHHOD, Elisabete Marques (1990), *Sintaxe dos predicados nominais com estar*, Lisboa, INIC - Instituto Nacional de Investigação Científica.

RASSI, Amanda Pontes (2008), *Estatuto sintático-semântico do verbo fazer no português escrito do Brasil*, Goiânia, 123 f. Dissertação (Mestrado em Linguística e Língua Portuguesa), Faculdade de Letras, Universidade Federal de Goiás.

A Qualitative and Quantitative Description of Syntactic Development in Two Textual Sequences*

Ricardo Benitez, Nina Crespo

Abstract: This study is part of more comprehensive research that seeks to correlate syntactic complexity with written discourse produced by various age groups. Syntactic complexity is defined according to the development of interclausular relationships within clause packages. These clause packages are constructed by syntactic, thematic, or discursive criteria. The interclausular categories used in this study are isotaxis, symmetrical parataxis, asymmetrical parataxis, hypotaxis, endotaxis, and nominalizations (Nir-Sagiv and Berman, 2010). The participants were Chilean Spanish-speaking school students who were asked to retell a story that had been previously recorded and to explain why schools exist. These two tasks were designed to elicit a narrative sequence and an explicative sequence (Adam, 1996, 2001). Their responses were transcribed according to guidelines proposed by Tusón (1995). The analysis of the transcriptions reveal that, among other things, the older the participant (i.e., the higher the school grade), the greater the complexity; that the explicative sequence develops more slowly than the narrative sequence; that there are early forms of hypotaxis; and that the use of endotaxis is rare.

Introduction

The objective of this paper is to describe the syntactic complexity of narrative and explicative texts orally produced by elementary and high school Spanish-speaking students.

Late language development has been linked to the acquisition of literacy and to schooling. This development impacts various aspects of language not only on quantitative, but also on qualitative terms, and it has also been related to the increase in syntactic complexity.

Many definitions of 'complexity' have arisen from different paradigms of human language; however, this paper presents only two of the most important notions of syntactic complexity: one is related to the number of both, transformations and words per T-Unit (HUNT, 1970); another suggests a functional and a discursive perspective on this linguistic phenomenon (BERMAN, 2004).

In this study, these two perspectives are considered in the quantitative approach to analyzing the syntactic complexity. First, this complexity is measured in terms of the number of words and clauses per unit. Second, a unit delimited by discursive criteria—called 'clause package'— is used to measure it. Berman's studies identify relationships between clauses within clause packages (CP).

A CP is defined as a unit embedded in the text made up of one or more clauses connected by a syntactic, thematic or discursive criterion (KATZENBERGER and CAHANA-AMITAY, 2002; KATZENBERGER, 2003; NIR-SAGIV and BERMAN, 2010).

In the qualitative approach, a preliminary study accounts for the types of relationships within all CPs: isotaxis, symmetric parataxis, asymmetric parataxis, hypotaxis, endotaxis, and nominalization.

Isotactic clauses are clauses that establish no relationships with any other, and they may be isolated or main clauses. Symmetric Parataxis establishes a relationship of informative equivalence with a link of clauses to a main clause by means of asyndeton (adjacent clauses) or coordinating connectors (coordinated clauses); cohesive devices are considered in this category. Asymmetric Parataxis is the relationship established by clauses that partially depend on a main clause; the dependent clause can only be interpreted in relation to the preceding clause; Berman (2004) includes in this group coordinated and adjacent clauses, with an elliptical subject or with an elliptical verb in the dependent clause, as well as subordinated (completive) clauses; that is, those that have a dependent relation to the main clause. Given the nature of Spanish syntax, dependent clauses with an omitted subject were not considered asymmetric but symmetric paratactic. Due to the great informativity of verbal endings, the omission of subjects is a frequent feature of Spanish. In this category, coordinated, adjacent and subordinated clauses are included and show a greater structural dependence than that of the symmetric parataxis. Hypotaxis is the category where a subordinated clause establishes a dependent relationship on a main clause; the clause can function as a noun, for instance, in clauses that act as subjects. It can also function as a restrictive adjective that delimits the scope of meaning of the noun to which it applies. This category includes those clauses whose function is adverbial, indicating time, place or manner, and those clauses that the Royal Spanish Academy (RAE, 2009) calls 'improper', such as causal, consecutive, final, concessive, conditional and comparative clauses. Endotaxis occurs when clauses are embedded in other clauses; it includes clauses that have been incorporated into the main clause to add information as adjectival explanatory clauses (non restrictive / non-defining / nonessential) as well as appositions and parenthetical clauses.

This study considers a narrative sequence (NS) and an explicative sequence (ES) (ADAM, 1996; ADAM, 2001; ADAM and LORDA, 1999). An NS is a textual unit that contains a network of hierarchical relations that can be divided into parts (sentences) linked to one another (propositions) and joined to the whole (sequence); besides, an NS can be a relatively autonomous entity that is provided with an internal organization and maintains a dependence/independence relationship with a larger unit: the text (ADAM and LORDA, 1999). It contains two

* This is part of more comprehensive research funded by Chile's National Fund for the Development of Science and Technology [FONDECYT Project Nº 1100600] and has been designed to establish correlations between the syntactic complexity and written discourse produced by the same individuals who participated in this study.

major foci: an Initial Situation and a Final Situation; to arrive at the Final Situation the producer of the sequence must go through the stages of Plot, Action, Evaluation and Denouement.

An ES asks a question that constitutes a problem; it responds to an explanation and concludes with an evaluation of what has been set forth. Its structure includes a first operator (the question "why?"; "how?" in the case of an expository sequence) and a second operator that leads to the solution-explanation with 'because'. The purpose of an ES is to contribute to the acquisition of knowledge. Berman (2004) adds that this type of text shows the author's detachment from the subject matter, thus distancing him/herself from it.

Method

Participants

The sample was made up of Chilean students, native Spanish-speakers, who study at three different types of schools: state, subsidized and private. These three types of schools both represented the universe of Chilean education and helped avoid bias but have not been considered intervening variables in the present analysis.

The participants were recruited from five different school grades: kindergarten, third and sixth elementary grades, and first and last (fourth) years of high school. Table 1 shows the exact number of participants per grade per type of school:

Grade	State	Subsidized	Private	Total
Kinder	26	32	33	91
Third grade	24	26	37	87
Sixth grade	31	35	41	107
First year HS	19	28	36	83
Fourth year HS	20	35	39	94
Total	**120**	**156**	**186**	**462**

Table 1. Number of participants per grade per type of school

Tasks

The tasks consisted of (a) the oral retelling of an NS, and (b) the oral answer of a question in order to produce the ES. For the NS task, the participants watched a video that told the story of an extraterrestrial who lands his spacecraft in the middle of a schoolyard and tells students and teachers that his dream was to be a galactic explorer; thanks to his mega-brain, he can speak Spanish fluently; however, his energy starts to sap over time, and he has to abandon the Earth for good. Then, the participants were asked to retell the story.

To carry out the ES task, the participants were asked to answer the question: "Why do schools exist?" The answers to this question were prompted by some visual material.

Analysis and findings

The participants' responses to both tasks were recorded and then transcribed for the analysis, following Tusón's (1995) guidelines. The analysis was carried out by means of a word index per CP as well as a clause index per CP.

One of the first findings about the NS is that the higher the school grade, the higher the number of words and CPs. Table 2 shows that the kindergarten participants employed 14 words and 2 clauses per CP and those participants in the 4th year of high school used 31 words and 4 clauses per CP.

Grade	W	CP	CL	W/CP	CL/CP
Kinder	39	3	6	14	2
3rd grade	94	5	15	20	3
6th grade	130	5	19	25	4
1st year HS	166	6	25	29	4
4th year HS	191	6	28	31	4

Table 2. Means and ratios between units (words and clauses) in the NS

Table 3 shows the inferential analysis performed for the NS:

Grade	Kinder	3rd grade	6th grade	1st year HS	4th year HS
Kínder	-	0,000838	0,000000	0,000000	0,000000
3rd grade	0,000838	-	0,000380	0,000000	0,000000
6th grade	0,000000	0,000380	-	0,029267	0,000528
1st year HS	0,000000	0,000000	0,029267	-	1,000000
4th year HS	0,000000	0,000000	0,000528	1,000000	-

Table 3. Significance of the differences of words per CP among school levels (NS)

As can be seen in Table 3, the differences per number of words per CP is significant only between Kindergarten and the elementary grades, since the p-value[**] is smaller than type I error at the 5% level. Between the first and fourth grades of high school, the differences are not significant, because the p-value is greater than 5%.

Grade	Kinder	3rd grade	6th grade	1st year HS	4th year HS
Kinder	-	0,000160	0,000000	0,000000	0,000000
3rd grade	0,000160	-	0,027123	0,000000	0,000000
6th grade	0,000000	0,027123	-	0,004898	0,005241
1st year HS	0,000000	0,000000	0,004898	-	1,000000
4th year HS	0,000000	0,000000	0,005241	1,000000	-

Table 4. Significance of the differences of clauses per CP (NS)

Table 4 shows the differences per number of clauses per CP; they are significant only between Kindergarten and the elementary grades, since the p-value is smaller than type I error at the 5% level. There are no significant differences between the first and the fourth grades of high school, because the p-value is greater than 5%.

As can been seen in Tables 3 and 4 in both the Word per CP and the Clause per CP significant differences were found between the elementary school grades and between these and the high school levels.

Table 5 shows the findings yielded by the analysis performed on the ES:

Grade	W	CP	CL	W/CP	CL/CP
Kinder	49	2	7	24	4
3rd grade	88	3	13	28	4
6th grade	112	4	17	30	4
1st year HS	143	4	21	33	5
4th year HS	213	6	28	38	5

Table 5. Means and ratios among the units (words and clauses) in the ES.

The analysis of the ES reveals that the higher the school grade, the higher the number of both words and CPs. The kindergarten participants employed 24 words and 4 clauses per CP, whereas those in 4th year of high school used 38 words and 5 clauses per CP.

In contrast to the inferential analysis of the NS, the inferential analysis performed on the ES indicates that there are no significant differences between neighboring school grades, yet there are significant differences between not so neighboring school grades, as Table 6 shows:

[**] The Kruskal-Wallis non-parametric test—used to compare means—was applied to obtain the p-value.

Grade	Kinder	3rd grade	6th grade	1st year HS	4th year HS
Kinder	-	0,447917	0,000150	0,000000	0,000000
3rd grade	0,447917	-	0,314977	0,000102	0,000000
6th grade	0,000150	0,314977	-	0,122036	0,000079
1st year HS	0,000000	0,000102	0,122036	-	0,773980
4th year HS	0,000000	0,000000	0,000079	0,773980	-

Table 6. Significance of the differences of words per CP among school levels (ES)

Table 6 reveals that there are differences in the number of words per CP. Because the p-value is smaller than the type I error at the 5% level, the above is significant only for kindergarten in relation to the sixth grade, first and fourth years of high school; for the third grade in relation to the first and fourth years of high school; and for sixth grade in relation to the fourth year of high school. Apparently, statistical differences per school grade significantly occur only every six years.

Grade	Kinder	3rd grade	6th grade	1st year HS	4th year HS
Kinder	-	1,000000	0,001037	0,000003	0,00000
3rd grade	1,000000	-	0,299978	0,004892	0,000176
6th grade	0,001037	0,299978	-	1,000000	0,212628
1st year HS	0,000003	0,004892	1,000000	-	1,000000
4th year HS	0,00000	0,000176	0,212628	1,000000	-

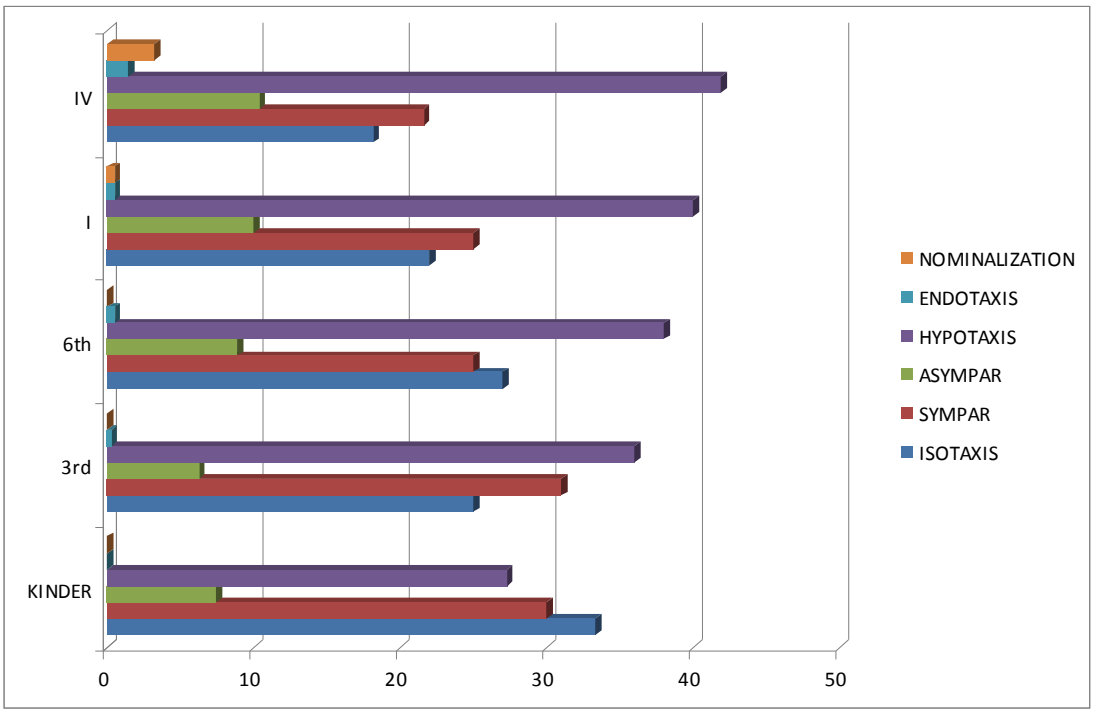
Table 7. Significance of the differences of clauses per CP among school levels (ES)

Table 7 indicates that there are differences in the number of clauses per CP. This is significant only for kindergarten in relation to the sixth grade, first and fourth year of high school; and for the third grade in relation to the first and fourth years of high school. It seems that there are no significant differences in the increase of this value from the sixth grade on.

Qualitative analysis

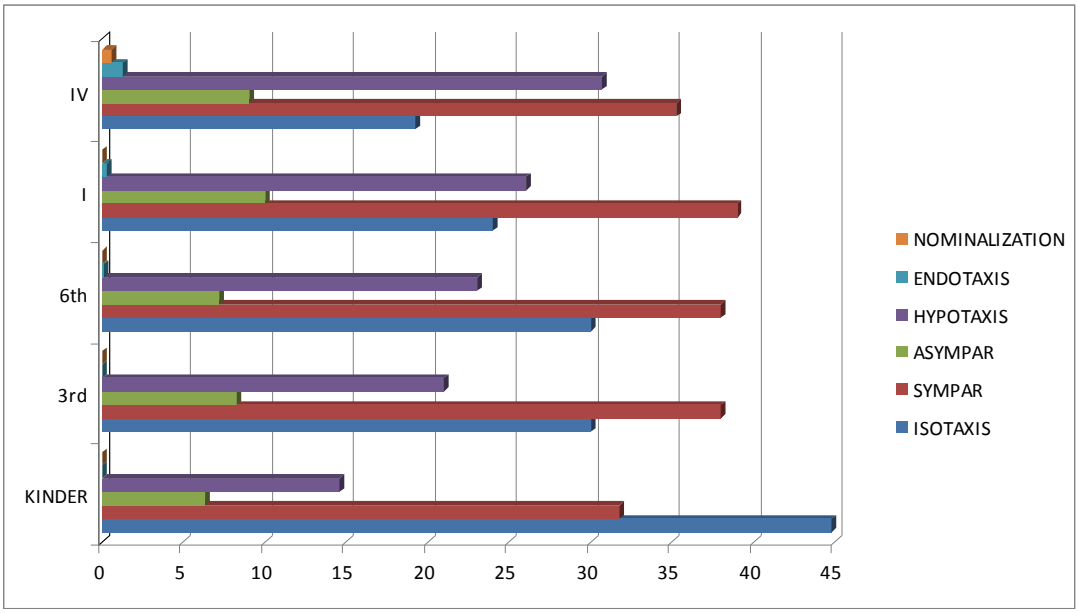
This analysis indicates that in both types of texts, the symmetric parataxis decreases, yet the asymmetric, hypotaxis and endotaxis increase. Early forms of hypotaxis, such as the causal (*El extraterrestre se volvió azul porque le faltaban los minerales de su planeta* [The extraterrestial turned blue because it lacked its planet's minerals]*)* and the final (*Las escuelas existen para que los niños aprendan cosas nuevas* [Schools exist so children can learn new things]*)* were observed. Besides, there is prevalence of a type of interclausular relationship per type of text and CP.

Graph 1 shows the interclausular relationships identified within the explicative sequence according to school grade. As can be observed, isotaxis (*Un marciano llegó a la escuela* [A Martian arrived at the school]*)* is often used by the kindergarten participants, and hypotaxis (*El extraterrestre que se llamaba Iofe se hizo amigo de los niños* [The extraterrestrial whose name was Iofe made friends with the children]) is the most employed interclausular structure as age increases. The frequency of use of asymmetrical parataxis (*Los niños tenían frío y los profesores también* [The children were cold and so were the teachers]*)* is almost the same across school levels as well as the symmetrical parataxis (*El marciano dijo hola y todos lo saludaron* [The Martian said hi, and everybody greeted it]*).* Both, endotaxis (*Los profesores, que saben mucho, enseñan a los niños* [Teachers, who know a lot, teach children]) and nominalizations (*El aprendizaje de cosas nuevas se da en la escuela* [Learning new things occurs at school]*)* are seldom used by all participants.



Graph 1. Inteclausular relationships in explicative sequence

Graph 2 shows the interclausular relationships within the NS according to school grade. Once again, isotaxis is the relationship often used in the ES produced by the kindergarten participants, whereas in all other school levels it is the symmetrical parataxis the structure that is most used. It can also be observed that the use of hypotactic structures increases with age and that the frequency of use of the asymmetrical parataxis is almost steady across school levels.



Graph 2. Interclausular relationships in NS

Conclusions

The findings can be analyzed by considering different perspectives. One of them concerns the relationship between type of sequence and syntactic complexity. From the Hunt's studies on, the findings seem to indicate that expository and argumentative texts often reveal a greater syntactic complexity than that of the narrative text. The above shows that somehow the pragmatic purpose that orients all texts has a decisive impact on the grammatical resources used to construct them. The results of this study indicate the same: the syntax used to construct the ES is more complex than the syntax to construct the NS.

In addition, this study shows how this complexity in school ages develops. Thus, the increase in syntactic complexity seems to be more evident and fast since it is significantly produced during the elementary levels of

education, whereas the development of syntactic complexity in the ES seems to be slower and gradual because the differences in the indices that mark this development are significant only when there is a six-year gap between age groups.

As far as the qualitative perspective of the analysis, there exists some increase in the use of hypotaxis, of endotaxis and, in the case of the ES, of nominalizations. However, there are early forms of hypotaxis emerging in kindergarten; these are adverbial causal and final forms in the ES, and adjectival restrictive forms in the NS. Nominalizations were found to emerge only from the first year of high school on and are considered a way of packing the information according to a morphological criterion, not according to a syntactic one.

References

ADAM Jean-Michel (1996), "(Proto)Tipos: la estructura de composición en los textos", *Textos nº 10,¿Textos? ¿Qué textos?*, Barcelona, Graó, p. 9-22.

ADAM Jean-Michel (2001), Les textes types et prototypes. Récit, description, argumentation, explication et dialogue, Paris, Nathan Université.

ADAM Jean-Michel ; LORDA Clara-Ubaldina (1999), *Lingüística de los textos narrativos*, Barcelona, Editorial Ariel, S.A.

BERMAN Ruth (2004), Introduction: Developing discourse stance in different text types and languages, *Journal of Pragmatics 37*, p. 105-124.

HUNT Kellogs (1970), Recent measures in syntactic development, in : LESTER Mark (éd.), *Readings in applied transformational grammar*, New York, Holt-Rinehart and Wiston, p. 187-200.

KATZENBERGER Irit (2003) The development of clause package in spoken and written texts, *Journal of Pragmatics 36*, p. 1921-1948.

KATZENBERGER Irit ; CAHANA-AMITAY Dalia (2002), Segmentation marking in text production, *Linguistics 40*, p. 1161-1184

NIR-SAGIV Bracha ; BERMAN Ruth (2010), Complex syntax as a window on contrastive rhetoric, *Journal of Pragmatics* 42, p. 744-765.

RAVID Dorit ; TOLCHINSKY Liliana (2002), Developing linguistic literacy: a comprehensive model, *Journal of Child Language* 29, p. 417-447.

TUSÓN Amparo (1995), *Anàlisi de la conversa*, Barcelona, Empuries.

Le figement à l'épreuve des corpus : les expressions « V. *dire* Det ∅ N_1 *propos* » de la table italienne CZER1

Catherine Camugli

Résumé: Ce texte est une contribution partielle à la mise à jour en cours de la table italienne des expressions figées à verbes transitifs CZER1 « V Det ∅ N_1 » : N_0 *cambiare bandiera*, « tourner casaque » (D'AGOSTINO et alii 2004). Une vérification systématique en corpus du micro-secteur « V. *dire* Det ∅ N_1 *propos* » précise certaines notations et enrichit partiellement la liste de variantes sur le verbe. L'analyse comparée des occurrences révèle en outre que c'est moins la nature particulière de N_1 (ANSCOMBRE 1986) que la nature [± hum] de N_0 qui inscrit les expressions dans le figement.

Abstract: This paper aims to contribute to the current updating of the CZER1 lexicon grammar table of Italian frozen verbal phrases (« V Det ∅ N_1 » /N_0 *cambiare bandiera*, change sides). The analysis of a contemporary corpus of written utterances with verbs of speech whose objects are exclamations « V. *say* Det ∅ N_1 *exclamation* » has helped me to add further detail to some definitions in the original table and to offer other possible verbs for these frozen phrases. Comparing these utterances has enabled me to suggest that what makes these phrases frozen utterances is less the specific quality of N_1 than a so called [± hum] factor of N_0.

La table italienne CZER1 des expressions figées à verbes transitifs et à Det zéro (structure « V Det ∅ N_1 » : N_0 *cambiare bandiera*, « tourner casaque », cf. D'AGOSTINO et alii 2004) a été créée par S. VIETRI (1985). Une réactualisation de cette table est en cours. La présente analyse porte sur un secteur dont on pourrait dire à l'intuition que V appartient aux verbes dits de paroles et que N_1 correspond à des propos proférés :

(1) N_o dire mea culpa

(2) N_o cantare vittoria

(3) N_o gridare (pietà + vendetta)

(4) N_o piangere miseria

(5) N_o non dire (a + beo)

Si l'entière table CZER1 questionne le lien entre transitivité et détermination zéro, ce secteur le pose de façon spécifique. Citant des expressions françaises analogues, J-C ANSCOMBRE (1986 :6) écrit de ces structures qu'elles sont « formées non sur la valeur substantivale de N mais sur sa valeur formulaire ». En effet, on peut leur associer une exclamation correspondant à N_1 si bien que l'on peut se demander quelle sera la part de cette caractéristique dans les facteurs portant à Det ∅.

Notre démarche sera d'observer en corpus le comportement de ces expressions pour a) en vérifier la vivacité en langue, b) observer dans le continuum des constructions connexes ce qui fait la spécificité des expressions relevant du figement et c) leur relations avec les structures avoisinantes.

1. Le corpus et les variations qu'il suggère

1.1. Délimitations du corpus

L'enjeu d'un corpus est de relativiser les degrés d'usage que les listes des dictionnaires aplatissent quelquefois et de mettre à disposition des contextes significatifs pour vérifier les propriétés linguistiques au-delà de la simple intuition du chercheur. Sans ignorer les préventions de M. GROSS quant à l'exhaustivité des corpus, le choix a été fait ici d'appuyer systématiquement l'analyse des variations sur des occurrences vérifiables dans des états de langue actuels et d'un niveau représentatif de la communauté des parlants. Pour ce faire, nous avons croisé les données d'un corpus oral (LIP, DE MAURO 1993), d'un autre de romans contemporains ayant obtenu le prix littéraire *Strega* de 1947 à 2006 (DE MAURO 2007) et la presse quotidienne à partir de la base de données *Factiva*[1]. Ont été délibérément écartées les requêtes tout-venant sur *Google* ou sur des blogs parce que susceptibles de recueillir également des idiolectes. C'est donc sur un italien moyen, proche de la norme sans y être inféodé, que l'observation est faite. L'observation s'appuie sur plusieurs centaines d'occurrences dont nous ne citerons ici que quelques-unes pour illustration.

[1] Base de données de presse internationale englobant plus de 20 000 sources d'information, issues de 159 pays et dans 23 langues, dont les principaux quotidiens et hebdomadaires, depuis janvier 1984.
Une autre réticence (ou simple limite) que l'on peut ajouter est la tendance des journalistes à reprendre textuellement les formulations des communiqués de presse et ce, quel que soit le journal et son bord politique, modifiant ainsi de façon non négligeable d'éventuelles statistiques.

1.2. Ce qu'apporte le repérage sur corpus et les questions qu'il pose

L'expression (1) *N_o dire mea culpa* qui n'apparaît ni dans le corpus oral (DE MAURO 1993) ni dans le corpus littéraire (DE MAURO 2007)[2], ne s'observe que de façon épisodique dans le corpus journalistique[3] ; on lui préfère les formes (1b) qui ne sont pas sans rappeler l'expression analogue française (1c).

(1b) fare (un+il+∅) mea culpa (1c) faire son mea culpa (C1G)

Pour les expressions (2) et (3), la fouille propose des variantes sur le verbe :

(2b) N_o (cantare + gridare) vittoria[4]

(3b) N_o (gridare + chiedere + invocare) (pietà + vendetta)[5]

rapprochant encore une fois des expressions françaises recensées dans l'index élaboré par M. GROSS:

(2c) (chanter + crier) victoire (C1R)

(3c) (crier + demander) vengeance (C1R)

Enfin, le corpus fournit sans difficulté des occurrences de (4) *N_o piangere miseria*[6] et, de façon moindre, de *non dire beo* (5b)[7]. La variante *non dire a* (5a) n'a pas été repérée sans doute à cause de son caractère oral.

Les expressions figées se caractérisent par un faisceau de traits linguistiques (cf. §. 1.3) dont l'impossibilité de modifications lexicales de V ou de N_1 et que le corpus semble ici contredire. L'éventuel élargissement lexical suggéré pour les expressions (1-3) correspond-il véritablement à des expressions figées ?

1.3. Les tests

A la fixité lexicale de V et de N_1 à peine évoquée, il convient d'ajouter des contraintes sur N_0 (±hum) (a), sur le déterminant de N_1 (b) et son nombre (c). Derrière un verbe transitif, tout complément direct est présumé essentiel et non suppressible. Les tests (M. GROSS 1982) pour déterminer les degrés de figement correspondent aux opérations que l'on peut effectuer de façon naturelle en syntaxe libre sur celui-ci. Ce sont la pronominalisation (d), l'expansion de N_1 par un complément de N ou une relative (e), sa modification par un adjectif (f), son extraction par une question partielle (g) ou par la focalisation *è … che/ c'est… que* (h). S. VIETRI (2004 : 153-56) développe le test de passivation selon deux formes : Passivo-1 qui correspond à la forme du passif avec permutation des sujet et objet (i) et Passivo-2 qui désigne la forme passive sans agent et sans permutation du sujet et de l'objet *Maria ha mangiato la mela* ↔ *è stata mangiata la mela* (j). Dans les tableaux, le signe (-) indique l'impossibilité de variation ou de transformation. Illustrons ces tests avec les deux expressions n'ayant pas posé de problème de repérage.

> (6) E la categoria che dimostra l'egoismo maggiore anche sul piano della risposta fiscale è sempre quella dei commercianti che *piangono miseria* e ipotetici danni ogni qualvolta il Comune tenti di mettere ordine al traffico e alle varie anarchie, come le occupazioni abusive di aree esterne con i ristoratori (*La Repubblica* 26/04/2012).
>
> (7) « E quando il sindaco è cambiato e mi ha mandato a casa, *non ho detto beo* per cinque anni. Secondo me i civici si devono comportare così » (*Il Resto del Carlino* 29/06/2010).

La variation du déterminant (b) restitue le sens référentiel (*piangere la miseria*) ou est simplement impossible (**non ho detto il beo*) ; la variation en nombre de N_1 (c) est bloquée comme elle le sera avec l'ensemble des expressions du groupe par la nature de mention de celui-ci[8]. La pronominalisation (d) équivaudrait à une reprise anaphorique de *risposta fiscale* en (6) et non de N_1 et aboutit à une phrase absurde en (7). Si (e) n'est réalisable qu'en co-occurrence avec l'Art. Def., en revanche, Modif par adjectif (f) est possible (*i commercianti piangono*

---

[2] Sauf dans une occurrence où N_1 est mention : « Ma io mi chiamo Viulante, e *dico : mea culpa*! che sono stata io, a dare la figlia mia a quell'assassino! » (Elsa MORANTE 1957 *L'isola di Arturo* , p. 209).

[3] Sur plus de vingt ans (depuis 1998), seulement 21 occurrences de *dire mea culpa*, y compris dans ses formes conjuguées (*dic*, dett*..). Pourtant GRADIT (entrée *mea culpa* CO) énumère sur le même plan *fare, recitare, dire il mea culpa*.

[4] La variante n'a toutefois pas la vitalité de l'expression originellement recensée : le rapport est de 1 à 16.

[5] L'originel *gridare pietà* ne présente que très peu d'occurrences : une seule recensée dans DE MAURO 2007 et dans *Factiva*, on n'en relève que deux en deux ans tandis que *chiedere pietà*, toutes formes morphologiques confondues, présente 121 occurrences sur la même période.
Avec N_1 *vendetta*, l'apport du corpus journalistique pour *chiedere* par rapport à *gridare* est de 1 à 8 (119 occurrences sur deux ans et 947 *gridare* -). L'apport journalistique est donc inverse selon les N_1.

[6] Ce ne sont pas moins de 346 occurrences sur deux ans.

[7] Six occurrences journalistiques en deux ans. D'origine régionale (*beo* >*bello* ?), cette expression n'est ni citée dans des dictionnaires d'usage ni dans le GDLI.

[8] N_1 *pietà* est en outre invariable.

miseria grande/ miseria ciclica e profonda) ainsi que la question partielle (g) et la focalisation (h) pour (6) et non pour (7). Les passivations sont impossibles.

	a	b	c		d	e	f	g	h		i	j
N_o piangere miseria	+ hum	-	-		-	-	+	+	+		-	-
N_o non dire beo	+ hum	-	-		-	-	-	-	-		-	-

On le voit, les traits de figement sont plus ou moins denses selon l'expression.

2. Les variations suggérées relèvent-elles véritablement du figement ?

2.1. La variation *gridare/cantare* (2a et 2b)

Comparons les comportements d'occurrences présentant les deux verbes.

> (8) ma i secoli lo vinceranno con la pazienza, millimetro per millimetro, fino a quando le tranquille acque napoletane *canteranno vittoria* in una bella giornata come questa, come fanno già sui tre o quattro scogli superstiti della villa di Pollione sotto Capo Posillipo (R. LA CAPRIA *Ferito a morte* 1961: 30).
>
> (9) Ora la Provincia *canta vittoria*: « la vicenda serva da utile monito per il futuro », dice il presidente Dario Allevi (*Il Giorno* 1/05/2012).
>
> (10) « Avevamo *gridato vittoria* troppo presto, appena saputo del via libera del provvedimento che riduceva i consiglieri », dice il presidente degli industriali sardi, Massimo Putzu, « un'occasione sprecata, l'ennesima, della politica sarda (...) (*Unione sarda* 1/10/11).

Quelle que soit la variante lexicale de V, la variation en nombre de N_1 (c) est bloquée et bien des transformations ne sont applicables que si N_1 reprend son statut de mention : développement par une relative (e) (*la Provincia canta: « vittoria », che ha meritato*), modification par un adjectif (f) *la provincia canta: « vittoria meritata*! »), focalisation (h) (*È « vittoria » che avevamo gridato troppo presto*), passivations (*Era stato gridato « vittoria » troppo presto*). La relative (e) n'est réalisable qu'en co-occurrence avec l'Art. Def. (*gridavano la vittoria che speravano*). Beaucoup d'indices convergent vers un enrichissement du domaine figé (*cantare + gridare*) *vittoria*.

Mais il nous faut ajouter quelques nuances. Outre la plus faible apparition de la variante en *gridare* (cf. note 4), certaines occurrences comme 11 (et non toutes, cf. 10) sont paraphrasables par « dire en V gérondif/ *dicevano vittoria gridando* », paraphrase confortée par des éléments du co-texte (*un canto rude e solenne*) qui situent l'expression proche de son sens référentiel.

> (11) I figli dei loro pronipoti rimasti lassù avanzavano ora verso il mare *gridando vittoria* con un canto rude e solenne, rivolto alla loro guida che aveva fatto leva su una sete di riscatto definitivo *Druze Tito, ljubicica bijela* (F. TOMIZZA *La miglior vita* 1977:181).

Plusieurs occurrences des relevés nous invitent ainsi à compléter les indications de CZER1 originelle pour *cantare vittoria* : N_0 n'est pas seulement [+hum] mais [± hum], comme l'illustrent (8) et (9) et c'est peut-être ce qui distingue *cantare* de *gridare vittoria* (uniquement +hum dans nos relevés).

	a	b	c		d	e	f	g	h		i	j
N_o cantare vittoria	± hum	-	-		-	-	-	-	-		-	-
N_o gridare vittoria	+hum	-	-		-	-	-	-	-		-	-
N_o gridare (*pietà + vendetta*)	± hum	-	-		-	-	-	-	-		-	-

Nous suggérons la même correction de détail pour l'expression 3a (*gridare vendetta*). N_0 originellement annoté [+hum] nous semble devoir être dotée de [± hum] cf. 12 et 13 où N_0 correspond à des inanimés (*storia* et *finanziamento*) versus *gridare pietà*, uniquement [+hum] (14).

> (12) « Le sembra possibile? Non *grida vendetta* a Dio, tutta questa storia? » Io non sapevo cosa dire. Avevo freddo (Dino BUZZATI *Sessanta racconti* 1958:289).
>
> (13) Anche il finanziamento ai partiti, mascherato da rimborso elettorale *grida vendetta* (*La Nazione* 17/04/12).
>
> (14) Tra coloro che *gridano pietà* ci sono anche imprenditori capaci e onesti, impegnati nel settore produttivo da trenta anni (*Unione Sarda* 25/03/11).

2.2. Le statut des variantes avec le verbe *chiedere* (3b)

Soient les occurrences qui combinent N_1 *vendetta* (15-16) et *pietà* (17-18) :

(15) Invocava un aiuto il drago, e *chiedeva vendetta* per i suoi figli. Ma a chi? alle montagne forse, aride e disabitate? (Dino BUZZATI *Sessanta racconti* 1958: 97).

(16) Antonella, la vedova dell'operaio, non *chiede vendetta*. Esige giustizia (*La Repubblica* 31/03/11).

(17) Lui si butta in ginocchio, pronto a *chiedere pietà*. E sapete cos'era ? Roba da non crederci (Gaia MAZZUCCO *Vita* 2003: 292).

(18) [una delle zone più devastate dallo tsunami] Morti neri, gonfiati dall'acqua, qualcuno con le mani protese verso il cielo come avesse voluto *chiedere pietà* nell'ultimo istante di coscienza (*Il Giorno* 12/04/12).

	a	b	c	d	e	f	g	h		i	j
(3a) N_o *gridare* (*pietà + vendetta*)	± hum	-	-	-	-	-	-	-		-	-
(3b) N_o *chiedere* (*pietà + vendetta*)	+hum	-	-	-	-	+	+	+		+	+

Si les variations de détermination et de nombre (b et c) demeurent impossibles, si la relative (e) demeure liée à l'Art.Def, d'autres transformations sont parfaitement réalisables : Adj. (f) (*chiedeva vendetta eterna per i suoi figli*), focalisation (h) (*è* vendetta *che* chiedeva il drago*/ non è vendetta che* chiede la vedova *ma giustizia; è* pietà *che* è pronto a chiedere; *è* pietà *che* avesse voluto chiedere). La variante avec *chiedere* ne semble donc pas relever du domaine figé. En outre, un fait a éveillé notre curiosité : le corpus présente également l'objet « figé » *vendetta* coordonné à d'autres N_1 (19).

(19) [le mugissement des bêtes] con ogni loro muggito che ci giungeva un po' sordo per la lontananza ma come se fosse rivolto proprio a noi, accusando e implorando, e *chiedendo* ad un tempo *vendetta e misericordia* (Elio VITTORINI *Le donne di Messina* 1949 : 423).

D'où la question : pourquoi ne pas envisager également d'autres $N_{1\ «\ propos\ »}$ derrière le verbe *chiedere* ? En d'autres termes, puisque l'on peut s'exclamer *aiuto !* ou demander *Permesso?* avant d'entrer dans une pièce, *scusa* ou *perdono*, n'est-il pas légitime de reconstruire un paradigme à N_1 « formulaire » ? [9]

(20a) Le famiglie che *chiedono aiuto* alle associazioni caritatevoli sono sempre di più (*Unione sarda* 17/05/12)

(20b) Li scontò tutti « nel carcere di Aversa senza mai *chiedere grazia* ». Una volta libero, per volontà regia « fu sottoposto a speciale sorveglianza di polizia » (*La Repubblica* 30/10/12).

(20c) Lo stesso McCullin che si racconta, non vuol essere definito né artista, né reporter di guerra, *chiede perdono* per le sue foto cruente che hanno ritratto le atrocità cruciali del ventesimo secolo (*Il Resto del Carlino* 12/05/12).

(20d) Le <dipendenti> intrusive non *chiedono permesso* e partecipano alla seduta e dicono la loro (*La Repubblica* 24/04/12).

(20e) L'amministratore delegato ora *chiede scusa* e fa mea culpa per l'accaduto ma rassicura: la banca è solida e i clienti non hanno subito perdite (*Il Giorno* 12/05/12).

Seule la passivation s'effectue dans des conditions analogues à celles des figées à savoir avec co-occurrence de Det Def. ou Ind. (*Un aiuto è chiesto dalle famiglie /L'aiuto è chiesto dalle famiglie*, etc.) ; on ne l'exclue pas totalement avec Det ∅ quelquefois (*grazia è chiesta, perdono è chiesto da Mc Cullin/ scusa è chiesta dall'amministratore delegato*). Le reste des tests (extraction, modification, etc.) ne permet pas de leur assigner le statut de figées.

Alors pourquoi Det.∅ dans ces expressions qui ne relèvent pourtant pas du figement ? L'analyse de RENZI (1991) n'a pu nous aider. Pour KORZEN (1996 : 147-148), l'absence de Det provient du fait que N_1 est un « constituant incorporé/ *costituente incorporato* » au verbe et qu'il fonctionne, non pas comme un substantif objet « obiectum affectum » (DE ROBERTO 2011 : 984) mais comme un modifieur verbal. M. HERSLUND (1999) explicite avec clarté les systèmes romans où l'italien - comme l'ancien français - distingue une transitivité où l'objet maintient des traits d'individualité et d'autonomie par rapport au verbe (et N est précédé d'un Art Def ou Ind.) d'une autre à Nom nu, à « référentialité réduite » et avec « coalescence entre verbe et objet, c'est-à-dire incorporation ». Dans ce cas, N_1 « se trouve réduit à une sorte de qualification adverbiale du procès verbal »

[9] Un peu comme la consultation de l'index permet d'établir pour le français N_o *demander* (grâce + pouce + réparation) (C1R) et crier (famine + gare + grâce + misère + miséricorde + vengeance + victoire) (C1R).

(HERSLUND 1999:46). Le paradigme *chiedere* (*aiuto + grazia + misericordia + perdono + permesso + scusa + vendetta*) relève donc de l'incorporation et non du figement.

Tout serait clairement défini si le corpus ne nous réservait pas des emplois comme (21), extrait d'un texte où est exprimé le mécontentement pour des aménagements urbains intempestifs.

> (21) Per non parlare della piccola oasi di verde che c'era nella piazzetta una « statua delle quattro stagioni » che è stata sostituita con la ghiaia. Tutto ciò *chiede vendetta*. Vorrei sottolineare anche una cosa invasiva che è la cartellonistica pubblicitaria stradale, perché toglie la vista a quel poco di bello ancora rimasto (*Il Giorno* 12/04/12).

	a	b	c		d	e	f	g		h	i	j
N_o *gridare* (*pietà + vendetta*)	± hum	-	-		-	-	-	-		-	-	-
N_o *chiedere vendetta*	- hum	-	-		-	-	-	-		-	-	-

Par les tests (cf. tableau), l'expression (21) revient à une simple variante de celle originellement enregistrée, *gridare vendetta*[10]. Encore une fois, ne serait-ce pas la nature [± hum] de N_0 qui fait basculer de l'emploi référentiel ou libre à l'acceptation métaphorique figée ? Il nous faut admettre qu'il y a deux expressions *chiedere vendetta* (l'une variante figée, peu usuelle mais existante et l'autre relevant de la syntaxe libre) ; d'où la possibilité d'intégrer la variante *chiedere vendetta* à la table CZER1 originelle, à condition toutefois de mentionner la possibilité d'un N_0 [-hum].

En nous inspirant du tableau de HERSLUND (1999: 45) et en traduisant les faits selon des conventions Lexique Grammaire, ceux-ci peuvent être ordonnés ainsi :

N_0 [-hum] V Det ∅ N_1	N_0 [+hum] V Det ∅ N_1	N_0 V Det (Def+Ind) N_1
Expr. Figée	Nom nu incorporé	N référentiel
Chiedere vendetta	*Chiedere permesso*	*Chiedere un permesso di soggiorno*
← V-N = procès verbaux →		*Chiedere il permesso di entrare*

Pour J-C. ANSCOMBRE (plus particulièrement 1991 :103), les GN à Det zéro en français « renvoient systématiquement à des procès » ; M. HERSLUND (1999 :38) parle d'*activité*. Quels qu'en soient les termes, l'observation vaut pour nos données. Ainsi ce n'est ni Det ∅, ni le caractère formulaire de N_1 qui déterminent le figement ; dans les limites de notre corpus, c'est la rupture de l'isotopie par le caractère [-hum] de N_0.

Les relevés proposent une autre variante (3b) avec *invocare*.

> (22) I militari cominciarono a gridare, qualche prigioniero a piangere, a pregare, a *invocare pietà*. Sole non aveva capito dove fosse (Ugo RICCARELLI *il dolore perfetto* 2004:289).

Pietà étant un mot tronqué, le nombre (c) perd de sa pertinence ; à part la corrélation obligatoire avec Det Def. (e) *qualche prigioniero cominciò a invocare la pietà dei militari/ la pietà che i militari non conoscevano*), les autres traits - *qualche prigioniero cominciò a invocarla* (d); *a invocare pietà cristiana/ profonda pietà* (f) , *è invocata dal prigioniero* (i) - nous inviteraient à ranger cette variante au côté de *chiedere*, dans ce secteur très particulier de l'incorporation qui confine avec le figement.

	a	b	c		d	e	f	g	h		i	j
N_o *chiedere pietà*	+hum	-	-		-	-	+	+	+		+	+
N_o *invocare pietà*	+hum	-	-		+	-	+	+	+		+	+

2.3. Que *faire* du *mea culpa* ?

L'expression concurrente de (1a), *fare* (un+il+∅) *mea culpa* apparaît avec différentes déterminations dans les proportions suivantes : 61 % *fare* ∅ mea culpa ; 25 % *fare il* - ; 11% *fare un* - [11].

> (23) Sarebbe opportuno che molte persone cominciassero *a fare mea culpa* e si facessero da parte (Il *Giorno* 17/04/12).

> (24) Molti incivili, sapendo che l'intervento dei vigili è lento, spesso inesistente o tollerante; e chiamare il carroattrezzi per far rimuovere la macchina di solito provoca ritorsioni pesanti da parte dei maleducati che non esitano a prendersela con il padrone di casa invece di *fare un mea culpa* (*Il Resto del Carlino* 19/03/12).

[10] *gridare vendetta* CO fam, scherzoso 'essere di pessimo gusto' (GRADIT).

[11] Les 3% restants se répartissent entre le verbe réfléchi *farsi* (un+il) *mea culpa* et d'autres déterminants (*fare qualche* -).

(25) Umberto Bossi *fa il mea culpa* solo perché è stato scoperto (*Il Giorno* 12/04/12).

Non seulement la détermination varie dans l'expression mais elle est interchangeable entre les réalisations : (23b) *cominciassero a fare* il *mea culpa; cominciassero a fare* un *mea culpa*, etc. Nous ne sommes donc pas dans les cas relevés par I. MIRTO (2010) où toute modification de déterminant entraîne des changements sémantiques. Devons-nous pour cela écarter tout processus de figement ?

Que nous enseignent les transformations/tests ? N_0 (a) est toujours [+hum] et la variation en nombre de N1 (c) est bloquée par la forme latine du mot. Les tests (e) et (f), volontiers corrélés avec Det (Def+Ind) s'appliquent régulièrement sauf pour (23) à Det ∅ (**molte persone cominciassero a fare mea culpa garbato/ *molte persone cominciassero a fare mea culpa che uno aspetta*). Il en est de même avec les passivations (**Sarebbe opportuno che mea culpa sia fatto da molte persone/ ? che sia fatto mea culpa*).

	a	b	c		d	e	f	g	h		i	j
N_o *fare* ∅ *mea culpa*	-	+	-		+	-	-	+	+		-	?
N_o *fare* il *mea culpa*	-	+	-		+	+	+	+	+		+	+
N_o *fare* un *mea culpa*	-	+	-		+	+	+	+	+		+	+

Or, modification du déterminant, variation possible de la négation, introducteur d'un modifieur de N_1 source de relative, passivation sont les propriétés spéculaires des verbes supports par rapport au figement (S. LANGER 2004 :172 ; E. MARINI 2003). La substitution de l'adjectif par un adverbe est possible (A.CICALESE 1995 : 131 ; *fare un mea culpa garbato* ↔ *fare garbatamente un mea culpa*). Quant au remplacement de l'expression entière par un verbe, on peut avancer *confessare*/ avouer. Ce qui ferait classer *fare* (un+il) *mea culpa* parmi les expressions à verbes supports si l'on admet *il/un mea culpa* comme nom prédicatif à la morphologie un peu particulière. Nous rangerions volontiers dans ce groupe la variante *recitare il mea culpa* dont les propriétés sont analogues.

> (26) Bossi arriva al mattino da Gemonio nella sede del suo movimento. E *recita il mea culpa*: « Io adesso devo stare lontano, non posso fare altro, devo stare un passo indietro (...) (*Il Giorno* 08/04/12).

Quel statut accorder à *fare* ∅ *mea culpa*[12] et à l'expression de départ (1a) *dire* ∅ *mea culpa* ? Il ne peut s'agir de verbes opérateurs (GIRY-SCHNEIDER 1978) puisque l'on ne peut paraphraser la construction verbale par un verbe associé morphologiquement à N_1. Sur la base d'occurrences comme (27) et (28), il nous semble que (*dire+fare*) peuvent être considérés comme des pro-verbes, quelque peu interchangeables, destinés à actualiser N_1 comme le montrent les guillemets qui l'entourent en (28).

> (27) « Cofferati *dica mea culpa* e rimedi ai suoi errori » (*Il Resto del Carlino* 1/06/2007).

> (28) I padroni di casa devono *fare "mea culpa"* per le importanti occasioni sbagliate durante la prima frazione di gioco (*Unione sarda* 19/03/12).

Des relevés en français (dont nous n'excluons pas l'existence en italien même si nous n'en avons pas encore repéré) montrent le fonctionnement autonome de ce N_1 particulier, fragment du *Confiteor*, qui vient redire un verbe (29) ou bien articuler l'argumentation comme un connecteur (30).

> (29) « Faites bien la différence entre une consigne directive qu'on a peut-être donnée au début du premier mandat, je l'avoue *mea culpa*, et 10 ans après (*Mediapart* 7/04/2012)

> (30) Alors oui, il faut l'admettre, être séparée du téléphone portable est une épreuve. *Mea culpa.* Mais il y a quand même certaines situations où, naturellement, chacun décroche (*La voix du Nord*, 6 /04/2012).

Dans le statut spécial de *mea culpa*, emprunt au latin et fruit d'une métonymie à partir de propos spécifiques, nœud prédicatif, résident des potentialités qui le mettent au cœur de constructions très variées.

3. Autres apports du corpus

3.1. Une meilleure appréhension de la structure argumentale

Le classement français de *pleurer - misère auprès de* dans la table C1RPN questionne le classement de l'expression italienne analogue (4) en CZER1. Or les 200 occurrences journalistiques qui ont servi de champ d'observation confirment pleinement la place de l'expression en CZER1. Les seuls syntagmes prépositionnels que l'on trouve dans le co-texte explicitent le motif (*per i tagli, un taglio, i pochi soldi* = 5 %) et seuls quelques locatifs (*in Cina, su Radio 24* = 2%) pourraient être assimilés à des *auprès de* mais en forçant les faits.

[12] Nous n'avons pas pu vérifier si l'expression figure dans la table FC (N_0 fare C1, D'AGOSTINO et alii 2004:131).

3.2. Des expressions pas si équivalentes d'une langue à l'autre

Le contexte dans lesquels apparaissent les expressions renseigne sur leur acception et cela est précieux lorsque l'on doit traduire ou comparer (C. CAMUGLI GALLARDO 2010 §.1.2 ; §.2). *Cantare vittoria* semble avoir comme équivalent structurel, lexical et sémantique *chanter victoire*. Or en français, les co-textes associent souvent l'excès exprimé par l'expression à une inadéquation temporelle, connotation très souvent absente du corpus italien (rare exception, notre exemple 10, supra). Ainsi *crier victoire* ne sera jamais *cantare vittoria* [13].

(31a) On ne va pas *chanter victoire trop vite*, nous savons que le plus dur nous attend (*Sud Ouest* 4/03/12)

(31b) A *crier victoire trop tôt*, il y en a qui y ont laissé des plumes (*Midi Libre* 17/04/12).

(31c) Pas de triomphalisme, a pourtant prévenu le PS. Pas question de se laisser griser et de *crier victoire avant l'heure* (*Ouest France* 16/04/12).

En travaillant hors corpus et en raisonnant encore une fois sur l'analogie structurelle, on penserait volontiers *à ne dire mot* (C1R) pour traduire (5) *non dire beo*, N_1 jouant le rôle de renforceur de négation (CORBLIN; TOVENA 2003). Or, dans bien des cas, l'équivalent français qui vient à l'esprit, dans le contexte précis de l'occurrence, est plutôt l'informel *sans moufter/ sans l'ouvrir* (cf. supra ex 7), avec un changement diamésique bien connu en traduction (CAMUGLI GALLARDO 2005).

Conclusion[14]

Le secteur V. *dire* Det ∅ N_1 *propos*, intuitivement délimité à l'intérieur d'un ensemble plus vaste qu'est la table CZER1, n'a été qu'une commodité méthodologique, en réduisant le champ d'observation. On n'a pu dégager de corrélation entre ce qui aurait pu être une classe lexico-sémantique (à partir de N_1 potentiellement employés de façon isolée en exclamations) et un comportement syntaxique homogène.

En revanche, la manipulation de nombreuses réalisations en contexte permet d'en préciser les caractéristiques, coupant court aux doutes d'une analyse reposant sur la simple intuition du chercheur et dans laquelle celui-ci en vient quelquefois à douter fortement. Ce n'est pas tant l'absence superficielle d'article qui a été un critère discriminant quant au degré de figement mais la nature de N_0 ou des réactions diverses aux tests.

Sans prétendre à un quelconque caractère péremptoire du corpus, ce travail partiel a permis
- de proposer un classement autre pour (1) *dire mea culpa*,
- d'élargir la variation sur V et de préciser le trait ± hum de N_o pour (2) (*cantare + gridare*) *vittoria*,
- de confirmer la validité de (3) avec l'ajout d'un trait N_o ± hum *gridare* (*pietà + vendetta*),
- d'y insérer une variation partielle de V avec restriction sur N_o - hum *chiedere vendetta*,
- de confirmer la légitimité de (4) *piangere miseria* dans cette table.

Bibliographie

ANSCOMBRE Jean Claude (1991), La détermination zéro : quelques propriétés, *Langages 102*, Paris, Larousse, p.103-124.

BATTAGLIA Salvatore (1966-2002), *Grande dizionario della lingua italiana*, Turin, UTET (dit GDLI).

CAMUGLI GALLARDO Catherine (2005), Niveaux de langue et variations linguistiques dans la comparaison interlangue de tables du *Lexique-Grammaire*, in : *Lingvisticae Investigationes*, n° 28 : 2. Amsterdam / Philadelphia, Benjamins, p. 169 -188

CAMUGLI GALLARDO Catherine (2010), Jusqu'où la syntaxe construit-elle le sens ? Réflexions autour d'une comparaison italien-français des locutions verbales figées, *Langages* 179-180, p. 243-258.

CICALESE Anna (1995), L'analisi dei nomi operatori con il verbo fare, in : D'AGOSTINO Emilio (ed), *Tra sintassi e semantica. Descrizione e metodi di elaborazione automatica della lingua d'uso*, Naples, Loffredo, p. 267-286.

CORBLIN Francis ; TOVENA Lucia M. (2003), L'expression de la négation dans les langues romanes, in : GODARD Danièle (ed), *Les langues romanes. Problèmes de la phrase simple*, Paris, éditions du CNRS, p. 281- 343.

D'AGOSTINO Emilio ; ELIA Annibale ; VIETRI Simonetta (2004), Lexicon-grammar, electronic dictionaries and local grammars of italian, *Lingvisticae Investigationes*, Supplementa 24, p. 125-136.

DE MAURO Tullio; MANCINI Federica, VEDOVELLI Massimo, VOGHERA Miriam. (1993), *Lessico di frequenza dell'italiano parlato* (LIP), Milan, Etaslibri.

DE MAURO, Tullio (éd.) (2007), Primo tesoro della lingua letteraria italiana del Novecento, Turin, UTET.

DE MAURO Tullio (éd.) (1999), *Grande dizionario dell'uso*, Turin, UTET (dit GRADIT)

[13] Le Robert les cite dans cet ordre *crier, chanter victoire.*

[14] Tous les exemples en italien auraient mérités d'être traduits mais l'espace imparti ne le permet pas. Nous le regrettons.

DE ROBERTO Elisa (2011), Oggetto, in : SIMONE Raffaele (ed.), *Enciclopedia dell'italiano*, Roma, Istituto dell'Enciclopedia Italiana G. Treccani, p. 983-987.
GIRY-SCHNEIDER Jacqueline (1978) Les Nominalisations en français. L'opérateur « faire » dans le lexique, Genève-Paris, Droz.
GIRY-SCHNEIDER Jacqueline (1981), Les compléments nominaux du verbe *dire*, *Langages* 63, Paris, Larousse, p.75-97.
GROSS Maurice (1982), Une classification des phrases « figées » du français, *Revue québécoise de linguistique*, vol. 11, n° 2, p. 151-185.
HERSLUND Michael (1999), Incorporation et transitivité dans les langues romanes, *Verbum* XXI, 1 Nancy, Presses Universitaires de Nancy, p. 37-47.
KORZEN Jørn (1996), L'articolo italiano fra concetto ed entità : uno studio semantico-sintattico sugli articoli e sui sintagmi nominali italiani con e senza determinante, Copenhague, Museum Tusculanum Press.
LANGER Stefan (2004), A linguistic test battery for support verb constructions, in : GROSS Gaston ; DE PONTONX Sophie (eds), *Verbes supports : nouvel état des lieux*, Lingvisticae Investigationes, Fasc. Spécial 27/2, Benjamins Publishing Co., Amsterdam/Philadelphia, p. 171-184.
MARINI Emanuela (2003), Tipologia delle costruzioni a verbo supporto a det. Ø in italiano antico e moderno in: MARASCHIO Nicoletta ; POGGIO SALANI Teresa (eds.) *Italia linguistica anno mille. Italia linguistica anno duemila*, Roma, Bulzoni, p. 259-272.
MIRTO Ignazio (2010), Nomi post-verbali e articolo zero in italiano, in: N. PRANTERA; A. MENDICINO; C. CITRARO (éd.) *Parole. Il lessico come strumento per organizzare e trasmettere gli etnosaperi*, Centro Editoriale e Librario, Università della Calabria, Rende, p. 589-607.
RENZI Lorenzo (1991), L'articolo in : RENZI Lorenzo (ed) *Grande grammatica di consultazione*, vol.1, Bologna, Il Mulino, p. 357-423.
VIETRI Simonetta (1985), Lessico e sintassi delle espressioni idiomatiche. Una tipologia tassonomica dell'italiano, Napoli, Liguori.
VIETRI Simonetta (2004), Lessico-grammatica dell'italiano. Metodi, descrizioni e applicazioni. Turin : UTET.

Étude du traitement de certains compléments de phrase dans le cadre d'une meta-grammaire

Éric de La Clergerie

Résumé: Dans le cadre de FRMG, une méta-grammaire (MG) à large couverture du français, nous examinons comment les propriétés d'héritage et de modularité des MG peuvent être utilisées pour représenter les compléments de phrase, à la fois en terme de diversité et en terme de position dans la phrase.

Introduction

Introduites sous différentes formes (CANDITO 1999; *GAIFFE & al 2002; BARRIER & al 2004; PARMENTIER & al 2005;* THOMASSET & al 2005*)*, les méta-grammaires (MG) permettent des descriptions grammaticales modulaires exploitant des propriétés d'héritage. Toutes s'appuient sur des hiérarchies de classes, qui permettent de se focaliser sur la description d'un phénomène syntaxique, comme la notion de sujet, en la rafinant progressivement au travers de l'héritage pour en capturer ses différentes formes. Un processus de compilation permet ensuite de combiner les classes pour produire les structures d'une grammaire dans un formalisme syntaxique cible, comme, par exemple, des arbres élémentaires dans le cas de grammaires d'arbres adjoints [TAG – *Tree Adjoining Grammars*] (JOSHI & al 1975).

Historiquement, les travaux autours des MG ont surtout été illustrés, avec succès, par le traitement des aspects valenciels des verbes (*CANDITO 1999; CRABBE 2005)* et éventuellement d'autres catégories comme les noms et les adjectifs (*BARRIER 2002)*. Pour résumer, il s'agit de prendre en compte les diverses dimensions descriptives que forment les fonctions syntaxiques (sujet, objet, etc), les positions (canoniques, extraites, etc), et les réalisations (nominales, prépositionnelles, phrastiques, etc).

Nous nous proposons de montrer que les MG facilitent également la description des compléments de phrase, vus comme généralisant la notion d'adverbiaux. Du point de vue de la description dans le cadre de grammaires symboliques opérationnelles, ces compléments regroupent un ensemble de situations bien moins étudiées et finalement très protéiformes (et très floues). Au final, ce travail, effectué pour le français dans le cadre de la méta-grammaire à large couverture FRMG (DE LA CLERGERIE 2005), permet de capturer un large éventail de possibilités avec une forte économie en terme descripif. Il est également l'occasion de présenter, en les illustrant, les caractéristiques de notre variante de méta-grammaires.

Nous commençons par lister quelques types de compléments que nous avons identifiés et que souhaitons décrire, en mettant en avant les problèmes potentiels. Ensuite, nous fournissons quelques éléments d'information sur les méta-grammaires utilisées dans cet article. Enfin, nous étudions comment on peut capturer de manière économique une partie de la richesse des compléments de phrase avec une méta-grammaire. Pour finir, nous indiquons quelques difficultés résiduelles.

Comment caractériser les compléments de phrase

Les compléments de phrase regroupent un ensemble très hétérogène de compléments. Dans la plupart des cas, ils sont assimilables à des adverbiaux, dans le sens où ils peuvent être remplacés, de manière plus ou moins équivalente, par des adverbes. En tant que tels, ces compléments peuvent porter sur divers aspects du temps (moment, durée, périodicité, etc), de lieu, de manière, de moyen, de concession, de cause, de conséquence, etc. Par nature, ils sont optionnels dans la phrase et coïncident bien avec la notion d'adjoints dans le cadre des TAG. Par contre, d'un point de vue descriptif, ils souffrent des problèmes suivants:

- Les compléments de phrase peuvent prendre de très nombreuses réalisations. Certaines, comme les adverbes, les groupes prépositionels, les participiales, ou les subordonnées, dénotent des modifieurs et ne prêtent pas trop à confusion. Mais nous avons aussi les cas de groupes nominaux, de groupes adjectivaux pouvant prêter à confusion avec d'autres fonctions dans la phrase (objet et attribut par exemple). Enfin, nous avons le cas de locutions adverbiales et de constructions idiomatiques plus ou moins figées, souvent difficiles à recenser et demandant des efforts spécifiques en terme de description.
- Les compléments présentent une forte mobilité dans la phrase. Ils peuvent apparaître en tête ou fin de phrase, mais aussi autours des arguments verbaux (par exemple entre un sujet non cliticisé et un verbe) ainsi qu'entre un auxiliaire et un verbe.
- En fait, les points d'accroches sont même plus variés: à la liste précédente de points directement liés à la phrase, on peut aussi citer le cas de compléments apparaissant à droite de mots coordonnants (*il mange une pomme et* ***parfois****, une poire*), de prépositions (*il part, avec,* ***toutefois****, une pointe de regret*) et des noms déverbaux (*L'annonce,* ***ce matin****, d'un remaniment a surpris tous les commentateurs*).
- Enfin, les compléments de phrases, comme d'autres types de compléments, peuvent être isolés du reste de la phrase par l'utilisation de diverses marques d'incises, en général la virgule, mais aussi les

parenthèses '(X)' et les tirets '- X –'. Ces marques sont obligatoires dans certains cas, mais en général optionnelles et, de fait, souvent manquantes dans la pratique.

Ces divers points sont illustrés ci-dessous par quelques types de compléments que nous traitons dans le cadre de notre méta-grammaire et qui montrent bien cette grande diversité de situations.

Le cas le plus typique est fourni par les adverbes (*il a **parfois** envie de partir; **Désormais**, il veut partir).* Nous avons aussi les groupes prépositionnels (***Avec son ami**, il a décidé de partir **sans tarder***), ainsi que les subordonnées (*il est arrivé **pendant que tu parlais***). Dans les cas relativement spécifiques, nous pouvons ajouter les participiales (***déçu par ses résultats**, il se retire; **Sa société n'allant pas bien**, il doit la vendre*) et adjectifs flottants (*il prend le train, **soucieux des deniers publics***).

À coté de ces cas *faciles*, nous avons des réalisations par des groupes nominaux ou des pronoms, avec

- des expressions temporelles dans « ***ce matin**, il est parti, sans rien dire »*.
- des indications de rang dans « *il a, **le premier**, fini l'exercice »*.
- des expressions nominales liés à des parties du corps humain, comme dans « ***mains sur la tête**, il recule contre le mur »*
- des expressions dénotant des positions spatiales (au sens large) ou des addresses
    - o ***couloir de droite**, vous avez la classe de Mr Louis.*
    - o *vous trouverez**, chapitre 22**, les explications nécessaires à ce devoir.*
    - o *il attend, **rue des Bourdonnais**, que ses amis arrivent.*
- des pronoms liés au sujet de la phrase, comme dans *« il a, **lui aussi**, décidé de partir »*

Enfin, nous avons les cas de réalisations par des locutions, pouvant prendre des formes très variées. Ainsi, *il y a* peut être vu comme une préposition mais son origine verbale n'a cependant pas disparu comme illustré dans « *il est parti, **il n'y a pas deux jours**, avec des amis »*. De même, nous pouvons continuer à considérer *voici* et *voilà* comme des formes verbales (impératives) utilisables pour dénoter des expressions temporelles dans « *Il est parti, **voici deux semaines**, avec des amis »*. Dans cette série, nous pouvons ajouter des expressions plus exotiques comme celles construites sur le verbe *obliger*, dans « ***service oblige**, je dois vous quitter »*.

Nous pouvons aussi mentionner le cas des constructions concessives, comme dans « *il a, **quoi que tu en penses**, toutes ses chances »* et également des cas de comparatives « *Paul a, **plus que son frère**, le sens de la famille »*. Pour clore cette liste, qui n'a rien d'exhaustive, nous avons les cas de compléments de phrase qui ne sont pas des adverbiaux comme les apostrophes (de l'audience) dans « *je vous annonce, **chers collègues**, ma démission ferme et définitive »*.

Cette liste suggère qu'il existe un cadre général commun à ses divers exemples, en terme de positions, de points d'accroche et d'incise. Par contre, un contrôle relativement fin est nécessaire pour autoriser et bloquer certaines possibilités (par exemple incise obligatoire dans certaines positions pour certains types de compléments). Et nous avons à faire face à la diversité des réalisations. Ainsi, pour les réalisations nominales, on peut constater que certaines requièrent des groupes saturés avec un déterminant (*ce soir),* d'autres non (*couloir de droite*); certaines doivent être nécessairement modifiées (*mains sur la tête*), et souvent avec des contraintes de nature sémantique. Comme nous allons le voir, les méta-grammaires fournissent un cadre possible pour cette tâche.

Rappels sur les Méta-Grammaires

La section suivante sera l'occasion de présenter concrètement certaines des caractéristiques de notre variante de Méta-Grammaire (THOMASSET & al 2005). Néanmoins, nous la présentons ici de manière plus synthétique et abstraite.

Une MG est constituée d'une hiérarchie de classes reliées par héritage (**<: SuperClass**). Une classe est un ensemble de contraintes pouvant portant sur des noeuds d'un arbre d'analyse (**node(Subject)** ou **Subject** selon les contextes) ou sur la structure (arbre) en cours de construction dans son ensemble (via le mot clé **desc**). Les contraintes sur les noeuds permettent d'exprimer la dominance (directe avec S >> **Subject** ou indirecte avec S >>+ **V**) ainsi que la précédence (**Subject < V :** le sujet précède le verbe). Il est possible d'avoir des contraintes sur le contenu d'un noeud, en terme de valeur de traits dans des structures de traits (**node(V): [cat: v, bot:[mode: indicative|subjonctive]]**). On peut aussi indiquer que deux valeurs doivent être identiques (ou plus précisemment unifiables) via une équation entre chemins dans des structures de traits (**node(Nc).top.number=node(Det).top.number :** accord en nombre entre un déterminant et un nom). Sans entrer dans les détails, il est également possible d'indiquer qu'un noeud est présent ou absent selon des conditions exprimées par des formules logiques sur des équations entre chemins **(~ Subject => node(V).top.mode = value(imperative) :** pas de sujet pour les verbes à l'impératif).

Enfin, une classe peut fournir une fonctionalité, vue comme une ressource **(+ agrement**) tandis qu'une autre classe peut réquérir cette fonctionalité **(- agrement**), éventuellement dans le cadre d'un espace de nom (**- det:agrement)**. Ce méchanisme est une alternative plus flexible à l'héritage, dans le sens où la même fonctionalité peut être requérie plusieurs fois dans des espaces de noms différents, alors qu'on ne peut hériter qu'une seule fois des contraintes d'une classe parente.

L'héritage, les ressources et les contraintes sont utilisées pour combiner les classes de manière à obtenir des classes *neutres* (toute fonctionalité/ressource demandée est fournie et toute ressource fournie est consommée) et *satisfiables*, du point de vue des contraintes (ainsi, deux contraintes de précédence **A < B** et **B < A** sont incompatibles dans une même classe). Les classes survivantes accumulent les contraintes de toutes les classes combinées, ces contraintes étant ensuite utilisées pour produire des arbres élementaires TAG minimaux.

Rappelons qu'une grammaire TAG se compose d'un ensemble d'arbres élémentaires initiaux et auxiliaires. Les arbres initiaux sont *substitués* au niveau de noeuds feuilles étiquetés par des non-terminaux. Les arbres auxiliaires sont *adjoints* au niveau de noeuds (internes) et permettent d'adjoindre du materiel à droite et/ou à gauche du noeud d'adjonction. De manière naturelle, les arbres auxiliaires et l'adjonction sont généralement utilisés pour traiter les modifieurs (optionels). Enfin, les arbres élémentaires possèdent généralement un noeud (feuille) ancre lexicalisé, correspondant à la tête de l'arbre, par exemple un noeud ancre de catégorie **v** (verbe) pour un arbre de racine **S** (phrase).

Décrire le comportement des compléments de phrase

En premier lieu, les compléments de phrase sont des modifieurs et sont représentés comme des arbres auxiliaires dans le cadre des TAG. Ils héritent donc des propriétes des arbres auxiliaires dans les grammaires TAG.[1]

La classe **auxiliaire** est très générique et indique seulement qu'un arbre auxiliaire comprend un noeud racine (de type **std** et nommé **Root**) dominant (>>+) un noeud pied (de type **foot** et nommé **Foot**), et que les deux noeuds portent la même catégorie syntaxique.

```
class auxiliary {                                  % Class for TAG auxiliary trees
   Root >>+ Foot;                                  % root dominates foot
    node(Root).cat = node(Foot).cat;               % same syntactic category on foot and root
    node Root : [id:Root];
    node Foot : [id:Foot];
    node(Foot).type = value(foot);
    node(Root).type = value(std);
    node(Foot).top = node(Foot).bot;
}
```

En fait, en tant qu'adjoints, les compléments se comportent comme des modifeurs assez simples, pour lesquels le noeud racine est le père du noeud fils (aucun noeud ne venant s'insérer entre la racine et le pied). Ceci est capturé par la classe **shallow_auxiliary**, qui hérite de la classe **auxiliary** et qui est utilisée par de très nombreux modifieurs. Elle se contente de préciser que **Root** est le père (>>) de **Foot**, et indique qu'elle fournit la fonctionalité **shallow_auxiliary**.

```
class shallow_auxiliary {                          %% Generic class for shallow auxiliary trees
    <: auxiliary;
    + shallow_auxiliary;                           %% provide functionality 'shallow_auxiliary'
    Root >> Foot;                                  %% root is parent of foot
}
```

En continuant à raffiner, on constate que les compléments partagent souvent la propriété de pouvoir être encadrés par des marques d'incises, donnant lieu à la classe **modifier_on_x** qui requiert la fonctionnalité **shallow_auxiliary** et qui introduit un nœud **Incise** de catégorie **incise** dominant le modifieur. Ce nœud **Incise** sert d'accroche potentielle pour une adjonction des marques d'incises. La propriété **incise_kind** permet de contrôler les marques d'incises autorisées et un ensemble de quelques d'arbres adjoints permettent d'adjoindre des virgules (**coma** et **comastrict)**, des parenthèses (**par**) et des tirets (**dash**). Les arbres pour l'ajout des virgules sont en peu particuliers, pour tenir compte des cas de début et fin de phrase, ainsi que des cas où une virgule physiquement présente sert en fait de marque partagée pour plusieurs incises.

```
class modifier_on_x {
    + x_modifier;
    - shallow_auxiliary;                                  % require functionality shallow_auxilary
    Root >> Incise;                                       % root parent of node incise
    node Incise : [cat:incise, id:incise, type: std];
   Incise >>+ Modifier;                                   % incise dominâtes the modifier
   node(Incise).bot.incise_kind = value(coma|comastrict|par|dash);  % incise_kind controls the marks
}
```

[1] À ne pas confondre avec la notion de verbe auxiliaire !

Il est à noter que la classe **modifier_on_x** est en fait utilisée par de nombreux types de modifieurs comme les adjectifs sur les noms, les adverbes sur les adjectifs, les adverbes sur les adverbes, etc.

À ce stade, il devient pertinent de différencier les cas où le modifieur se situe avant ou après l'élément modifié (représenté par le nœud pied), en raffinant (par héritage) la classe **modifier_on_x** et en utilisant les valeurs **ante** et **post** de la propriété **position** pour permettre un contrôle.

```
class modifier_before_x {
   <: modifier_on_x;
   Incise < Foot;                                   % Modifier on the left side of the modifiee
   node(Incise).adj = node(Incise).ante.adj;
  desc.position = value(ante);
}
class modifier_after_x{
   <: modifier_on_x;
   Foot < Incise;                                   % Modifier on the right side of the modifiee
   node(Incise).adj = node(Incise).post.adj;
   desc.position = value(post);
}
```

Nous pouvons maintenant nous focaliser sur les compléments de phrase, qui, en première approximation, ne sont que des **modifier_on_x** avec x=S comme catégorie syntaxique de l'élément modifié. En pratique, nous considérons aussi la catégorie syntaxique artificielle **VMod** qui chapeaute tous les arguments verbaux (sujets non clitiques inclus) et qui permet d'accrocher des modifieurs de phrase à leur droite (*Paul, **ce matin**, est parti à la chasse*).

```
class modifier_at_S_level {
   + s_modifier;
   - x_modifier;
   - s_adj_pos;
   node(Root).bot = node(Foot).top;                 %% S-modifiers do not alter S properties
   node(Root).cat = value(S|VMod);                  %% may adjoin on S or VMod
}
class modifier_on_S {                               %% specialization for S
   + s_adj_pos;
   node(Root).cat = value(S);
}
```

Il est maintenant facile de partager ces propriétés communes pour divers types de compléments, en commençant par les adverbes, qui héritent d'autres propriétés générales des adverbes (par exemple un peu de sous-catégorisation), avec de plus la contrainte que les adverbes modifieurs de phrase ne peuvent être l'adverbe *très*, des adverbes d'intensification (comme *plus*) ou des adverbes de comparaison (comme *autant* ou *davantage).*

```
class adv_s {                                       %%  Adverbs on sentences (may adjoin on both sides)
   <: adv;                                          %% inherit properties for adverbs
   - s_modifier; Adv = Modifier; node(Foot).dummy.cat = value(adv);
   node(Adv).bot.adv_kind = value(~très|intensive|equalizer);          % restrictions on adverbs
   ...
}
```

Mais nous avons aussi la classe **cnoun_as_adv** qui permet d'utiliser les groupes nominaux (cat=N2) portant le trait **time** (avec une valeur non négative) comme modifieur de temps.

```
class cnoun_as_adv {
  node N2: [cat: N2, id:time_mod, type: subst];
   node(N2).top.time = value(~ -);                  % N2 should carry a time property (other than -)
   - s_modifier; N2 = Modifier; node(Foot).dummy.cat = value(adv);          % N2 acts as an adv
}
```

La classe **bodypart_cnoun** (qui hérite de la classe **cnoun** des groupes nominaux) permet d'utiliser des parties du corps humain comme modifieurs phrastiques, à la condition qu'elles soient elles-mêmes modifiées (à droite), par exemple par un adjectif (***tête courbée**, il avance dans le vent*).

```
class bodypart_cnoun_as_modifier {
  <: cnoun;                                                    %% inherit from nominal phrase
   desc.@kind0 = value(-);
   - s_modifier; Modifier = N2Root; node(Foot).dummy.cat = value(adv);    % s_modifier acting as adv
   node(Root).cat = value(~ N2);                               %% modifier can be attached on event nouns
   node(Root).bot = node(Foot).top;
   node(Anchor).bot.semtype = value(bodypart);                 %% semantic property bodypart
   node(N2).adj = value(strict);                               %% right adjoining is mandatory on N2
   node(N2).adjleft = value(no);
   node(N2).adjwrap = value(no); }
```

De nombreuses classes requérant - **s_modifier**[2] existent, par exemple pour les groupes prépositionnels (**_prep_s_modifier**), les subordonnées (**_csu_s_modifier**), les participiales (**participiale**), les concessives (**consessive_relatives**), les comparatives (**comparative_as_vmod_mod**), les interjections (**_pres_s**), le rang (**cnoun_as_position_mod**), les apostrophes (**audience_on_s**), les positions spatiales (**position_on_s**, **address_on_s**), les pronoms modifieurs (**pronoun_as_mod**), certaines locutions (**verb_ilya_as_time_mod**, **verb_voici_as_time_mod**, **verb_oblige_as_mod**), etc.

Nous illustrons (partiellement) le cas de *il y a*, où nous combinons la notion de s_modifier avec les contraintes d'une structure verbale canonique (**<: *verb*_canonical**), qui elle-même, en cascade, va requérir les divers composants formant une phrase complète autour d'un verbe.

```
class verb_ilya_as_time_mod {
   <: _verb_canonical;                                         % inherits canonical verb construction
  - s_modifier; S = Modifier;
   node v:[cat:v, lex: avoir];
   desc.ht.imp = value(+);                                     % impersonal subject expected
   desc.ht.extraction = value(-);                              % argument extraction not allowed
   ... }
```

Par ailleurs, il est progressivement apparu que les modifieurs de phrases se retrouvent autour d'autres catégories syntaxiques que '**S**' et '**VMod'**, comme par exemple après un mot coordonnant (catégorie '**coo'**), après une préposition ('**prep'**) ou un nom déverbal dénotant un événement. On voit que l'on force le modifieur à suivre le coordonnant (post position).

```
class modifier_on_coo {
   + s_adj_pos;
   node(Root).cat = value(coo);
   desc.position = value(post);                                % modifier on the right side of coo
   node(Foot).top.modifier = value(-);          % restriction: no more than one modifieur per coo node
   node(Root).bot.modifier = value(+); }
```

Dans les cas des compléments venant se positionner à droite d'une préposition, nous contraignons (de manière très arbitraire) la préposition a être *avec*, *pour*, ou *contre*, ce qui semble recouvrir les cas les plus fréquents. Bien sûr, cette liste est facilement extensible, ou peut être remplacée à terme par une propriété. Il est aussi à noter que les marques d'incises sont considérées comme obligatoires, ce qui n'est pas toujours vérifié dans la réalité.

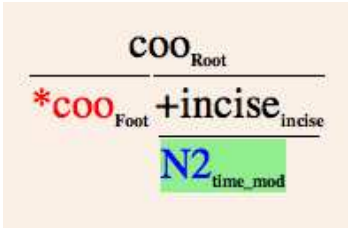
```
class modifier_on_prep {
   + s_adj_pos;
   node(Root).cat = value(vmodprep);
   desc.position = value(post);                                % modifier on the right side of
prep
   node(Root).bot = node(Foot).top;
   node(Incise).adj = value(strict);                           % mandatory parenthetical marks
   node(Incise).bot.incise_kind = value(comastrict|par|dash);
   node(Foot).top.pcas = value(avec|pour|contre);              %% only some prep may be modified
```

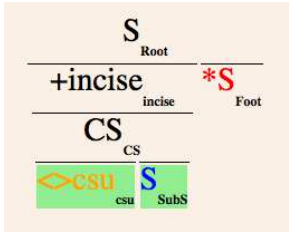
Enfin, l'application de certains adverbiaux sur des groupes nominaux est possible mais controlée par la présence de la valeur **event** pour le trait **semtype**, portée par les groupes nonimaux dont la tête est un nom évenementiel (souvent un déverbal) comme *départ* ou *annonce*. Il est à noter que ces modifieurs se distinguent des autres modifieurs du nom (adjectifs, groupes prépositionnels, relatives, etc) ainsi que des adverbes modifieurs de noms (*Deux cents francs environ ont été volés*).

[2] ou parfois des sous-composants de **s_modifier** (plus hauts dans la hiérarchie) dans le cas de contraintes plus spécifiques.

```
class mod_on_N2 {                    %% some event nouns may be modified like verbs
   + s_modifier;                     %% example: la tenue, ce matin, d'une assemblée a évité le désastre
   - x_modifier;
   - s_adj_pos;
   node(Root).cat = value(N2);
   node(Root).bot.semtype = value(event);                    %% semantic property event
}
```

Au final, l'utilisation de ces classes permet d'engendrer un ensemble conséquent d'arbres, à savoir 79 arbres, ce qui est beaucoup si on compare avec les 38 arbres de FRMG actuellement ancrés par des verbes pour couvrir les constructions canoniques et celles avec extractions (interrogatives, relatives, clivées, topicalisation), ce pour les voix actives et passives. Cependant, à la différence des arbres verbaux qui sont très complexes, les arbres pour les compléments sont en général très simples, comme illustré par les deux arbres ci-dessous, le premier pour les groupes nominaux temporels accrochés sur une coordination et le second pour les subordonnées en tête de phrase.





Un décompte effectué sur les résultats d'analyse syntaxique de 10096 phrases du French TreeBank (sur du style journalistique) montre que ces 79 arbres représentent presque 8% de tous les arbres utilisés. Le plus fréquent (4,3%) correspond aux groupes prépositionnels accrochés sur les **Vmod**, le deuxième (0,8%) étant pour les groupes prépositionels en tête de phrase. Les deux suivants sont pour les adverbes. Seulement ensuite arrivent les nominaux temporels (0,35%) et subordonnées (0,13%). Il semble donc nécessaire de multiplier les arbres pour capturer ces divers types de compléments, en notant cependant que beaucoup de ces compléments ont un taux d'occurrence très faible.

Limitations

Malgré l'aide apportée par les classes présentées ci-dessus (particulièrement celles fournissant **+s_modifier**), il reste difficile de capturer la diversité des compléments de phrase, en particulier ceux liés à des locutions. Il faut effectuer un travail de recensement long et laborieux, pouvant conduire à un ensemble important d'arbres très rarement utilisés. Pour éviter de trop multiplier le nombre d'abres, il faut continuer à rechercher des généralisations derrière ces locutions.

Par ailleurs, certains compléments sont faciles à décrire mais sont en fait contrôlés par des propriétés sémantiques sur le modifieur (temps, spatial, partie du corps humain, évenements, etc). La non-prise en compte de ces contraintes peut avoir un impact important en terme de sur-génération et en terme d'efficacité pour un analyseur syntaxique. Or, la mention de ces propriétés sémantiques au sein des entrées lexicales est une tâche difficile et de longue haleine. Notons cependant que l'utilisation de versions non contraintes sur un treebank ou sur un large corpus peut permettre d'apprendre les classes sémantiques en question, pour les injecter dans le lexique ou, de manière moins stricte, pour les utiliser en tant que préférences pendant la phase de désambiguisation.

De manière plus générale, il y a nécessité d'un dialogue entre les développeurs de FRMG et du lexique Lefff (SAGOT 2010) pour recenser, implanter et exploiter les informartions lexicales utiles. Ainsi, Lefff fournit des informations assez précises sur divers types d'adverbes (SAGOT B & al 2009), informations non encore complètement prises en compte pour le contrôle des adverbes dans FRMG. Encore plus précisement, on peut espérer que certaines de ces informations puissent indiquer si les marques d'incise sont obligatoires ou non pour un adverbe donné dans certains contextes (par exemple entre l'auxiliaire et le participe).

Régulièrement, de nouvelles positions d'accroche des compléments de phrase émergent, comme par exemple sur les noms évenementiels. Comme nous l'avons vu, l'ajout dans la méta-grammaire d'un nouveau lieu n'est pas très complexe, en s'appuyant sur les classes existantes. Néanmoins, l'impact sur les temps d'analyse peut être important.

Enfin, il est à noter que l'accroche des compléments sur certains lieux comme les coordinations et les prépositions relève plutôt d'une approche de syntaxe superficielle, résultant des limites des TAG. Il serait plus judicieux d'avoir un traitement plus sémantique avec l'indication d'une portée (*scope*) pour ces compléments. Ceci sera peut-être rendu possible par une transformation des sorties syntaxiques (sous forme de dépendances) vers des structures plus profondes. Une autre alternative envisagée est l'utilisation de TAG multi-composants (MC-TAG) comme formalisme, en place des TAG actuelles.

Conclusion

Nous avons montré comment la description modulaire de diverses propriétés syntaxiques, comme la notion de modifieur et de marques d'incises, permet une gestion plus facile et finalement assez économique des compléments de phrase, qui recouvrent un ensemble large et hétéroclite de constructions. La définition de fonctionnalités de base (**shallow_auxiliary**, **x_modifier**, **s_modifier**, **incise**, etc) associée à des mécanismes de contrôle (via des valeurs dans des structures de traits comme **desc.position = value(post)**) fournit rapidement une bibliothèque dans laquelle on peut piocher pour décrire de nouvelles constructions syntaxiques. En jouant le jeu, parfois un peu contraignant, de la recherche du bon niveau de généralité, on facilite également à terme l'application à des cas connexes non initialement prévus. Ainsi, l'extension de l'accroche des compléments de phrase sur les coordonnants, les prépositions et noms évènementiels s'est faite très naturellement. Ce travail nous semble au final une confirmation de l'utilité des méta-grammaires pour des descriptions syntaxiques relativement fines, même si nous restons conscient des difficultés des langues et du besoin d'informations complémentaires (sémantiques en particulier).

Références

BARRIER S.; BARRIER N. (2004), Metagrammars: a new implementation for FTAG. Actes du colloque TAG+7, Vancouver.

BARRIER N. (2002), Une Métagrammaire pour les adjectifs du français. Actes du colloque TALN 2002, Nancy.

CRABBE, B. (2005), Représentation informatique de grammaires fortement lexicalisées - Application à la grammaire d'arbres adjoints. Ph. D. thesis, Université Nancy 2.

CANDITO M.-H. (1999), Organisation modulaire et paramétrable de grammaires électroniques lexicalisées. Thèse de doctorat de l'Université Paris 7.

DE LA CLERGERIE E. (2005). From metagrammars to factorized TAG/TIG parsers. In Proceedings of IWPT'05 (poster), p. 190–191, Vancouver, Canada.

GAIFFE B.; CRABBÉ B; ROUSSANALY A. (2002), A New Metagrammar Compiler. In Proceedings of the 6th International Workshop on Tree Adjoining Grammars and Related Frameworks (TAG+6), Venice.

JOSHI A. K.; LEVY L.; TAKAHASHI M. (1975), Tree Adjunct Grammars. Journal of Computer and System Science 10, 10(1), p. 136–163.

PARMENTIER Y; LE ROUX J. (2005), XMG: a Multi-formalism Metagrammatical Framework. - ESSLLI'2005.

SAGOT B. (2010), The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In Proceedings of the 7th Language Resources and Evaluation Conference (LREC'10), La Valette, Malte.

SAGOT B.; FORT K.; VENANT F. (2009), Extension et couplage de ressources syntaxiques et sémantiques sur les adverbes. In Linguisticæ Investigationes 32(2), p. 305-315.

THOMASSET F ; DE LA CLERGERIE E. (2005), Comment obtenir plus des méta-grammaires. In Proceedings of TALN'05, Dourdan, France.

Connecteurs de discours adverbiaux : Problèmes à l'interface syntaxe-sémantique

Laurence Danlos

1 Introduction

Parmi les adverbiaux (simples ou composés) a été identifiée la classe des adverbiaux connecteurs de discours (les conjonctions de subordination et de coordination étant aussi des connecteurs de discours). Divers critères ont été mis en avant dans la littérature pour identifier la classe des adverbiaux connecteurs, voir entre autres (Moliner, 1990) et (Bonami et al., 2004). Les principaux sont :

– non intégration au contenu propositionnel de la phrase dans laquelle ils apparaissent, ce qui se manifeste par l'interdiction de focalisation dans une clivée,

– nécessité que la phrase dans laquelle ils apparaissent ait un contexte discursif gauche non vide,

– ce ne sont pas des expressions référentielles ni anaphoriques et le sens des adverbiaux composés n'est pas compositionnel.

A ces critères syntaxiques ou sémantiques s'ajoutent des critères proprement discursifs (Roze, 2009). Cet ensemble de critères a servi à construire une base lexicale des connecteurs de discours du français, LEXCONN (Roze et al., 2012) qui est librement disponible sur le site http: //www.linguist.univ-paris-diderot.fr/~croze/D/Lexconn.xml. Cette base lexicale répertorie pour chaque item (simple ou composé) sa catégorie syntaxique et une indication sur son sens via la « relation de discours » qu'il exprime. Nous n'entrerons pas dans une présentation des relations de discours, en nous contentant d'indiquer qu'un même connecteur peut avoir différents sens comme illustré en (1) : en (1a), le connecteur adverbial ensuite indique la succession temporelle des voyages de Fred, tandis qu'en (1b) il participe à une énumération.

(1) a. Fred ira à Dax pour Noël. Ensuite, il ira à Pau.

b. Fred présente trois qualités majeures. Il est gentil. Ensuite, il est intelligent. Enfin, il a de beaux yeux.

Il existe donc deux entrées de *ensuite* dans LEXCONN. Cette base lexicale compte 206 entrées de connecteurs adverbiaux et 430 entrées de connecteurs au total.

En admettant que la classe des connecteurs adverbiaux ait été clairement identifiée, nous nous concentrons dans cet article sur les problèmes posés à l'interface syntaxe-sémantique pour les éléments de cette classe. Le premier de ces problèmes, bien connu, est le fait qu'un connecteur adverbial n'a qu'un seul argument syntaxique — comme tout adverbial — et deux arguments sémantiques — comme tout connecteur. La prolongation d'un analyseur syntaxique (phrastique) par un analyseur sémantique (discursif) demande donc de mettre au point un mécanisme qui permette d'attribuer un argument sémantique supplémentaire (considéré comme le premier argument) aux connecteurs adverbiaux. Diverses solutions ont été proposées, entre autres dans les formalismes D-LTAG (Forbes-Riley et al., 2006) et D-STAG (Danlos, 2009), mais ce n'est pas l'objectif de cet article de discuter de cette question. Nous voulons nous concentrer sur la question suivante : dans quelle mesure le second argument sémantique d'un connecteur adverbial correspond-il à son argument syntaxique ?

Nous allons procéder en deux temps : à la Section 2, nous examinons les connecteurs adverbiaux apparaissant dans des phrases simples, c'est-à-dire sans phrase enchâssée. Cette section nous permettra de présenter notre terminologie et nos conventions. A la Section 3, nous examinons des connecteurs adverbiaux apparaissant dans des phrases complexes (avec au moins une phrase enchâssée) et mettrons en évidence des cas où le second argument sémantique du connecteur ne correspond pas à son argument syntaxique. La Section 4 examinera l'extension de ces cas qui posent problème à l'interface syntaxe-sémantique.

2 Portées syntaxique et sémantique dans les phrases simples

Nos exemples sont toujours composés de deux phrases typographiques (séparées par un "."), le connecteur adverbial apparaissant dans la seconde. Nous adoptons les conventions typographiques suivantes (inspirées du PDTB (Penn Discourse Tree Bank, (PDTB Group, 2008)) : si l'exemple est acceptable, le connecteur est souligné, l'empan de son premier argument sémantique est mis en italiques, celui de son second argument sémantique en gras, comme en (2).

(2) *Fred ira à Dax pour Noël.* Ensuite, **il ira à Pau.**

Ces conventions vont nous permettre d'éviter toute discussion sur la nature exacte des arguments sémantiques des connecteurs qui varie selon les auteurs et les exemples (événement, proposition, acte de parole, etc.). Cette discussion est hors sujet et nous nous situons hors d'un cadre discursif particulier. Comme nous nous concentrons sur le second argument sémantique du connecteur, nous considérons que déterminer l'empan de ce second argument revient à déterminer la « portée sémantique » du connecteur.

Sur le plan syntaxique, un adverbial est considéré comme un « modifieur », « ajout » ou « adjoint » selon les auteurs. Nous utiliserons le terme ajout sans nous situer dans un cadre syntaxique particulier. Lorsqu'un connecteur adverbial figure dans une phrase simple (sans phrase enchâssée), il peut généralement occuper diverses positions[1] : à l'initiale, (3a), à l'intérieur, (3b), ou à la finale de cette phrase, (3c). Le connecteur dans ces exemples est un ajout initial ou final sur la phrase entière ou un ajout sur un élément du VP de cette phrase, quoiqu'il en soit un ajout sur un élément de (la structure syntaxique) de cette phrase. Nous appelons « phrase hôte » du connecteur adverbial la phrase dont un élément est le site d'adjonction de l'adverbial.

(3) a. *Fred ira à Dax pour Noël.* Ensuite, **il ira à Pau**.

b. *Fred ira à Dax pour Noël.* **Il ira** ensuite **à Pau.**

c. *Fred ira à Dax pour Noël.* **Il ira à Pau**, ensuite.

Sur le plan sémantique, quelle que soit la position du connecteur en (3), sa portée sémantique est sa phrase hôte. Ces exemples permettent donc d'avancer le principe suivant :

Principe 1 La portée sémantique d'un connecteur adverbial est sa phrase hôte.

Ce principe conduit à une interface syntaxe-sémantique triviale pour les connecteurs adverbiaux, avec des portées syntaxique et sémantique identiques, mais nous allons montrer que ce principe ne tient pas dans des cas plus complexes.

Auparavant une remarque. Comme remarqué dans (Bonami et al., 2004), la portée sémantique d'un connecteur adverbial apparaissant dans une phrase simple est sa phrase hôte, y compris les éventuels adverbiaux à prosodie « détachée (incidente) ». En (4a), la portée de ensuite, qui a une prosodie « intégrée », englobe *probablement* qui est en position détachée à prosodie incidente. Les adverbiaux connecteurs se distinguent ainsi des autres adverbiaux dans la mesure où (Bonami et al., 2004) avancent une contrainte postulant qu'un adverbial à prosodie incidente a portée large sur tout adverbe à prosodie intégrée, contrainte qui permet d'expliquer le contraste d'acceptabilité entre (4b) et (4c).

(4) a. *Fred ira à Dax pour Noël.* **Probablement, il ira** ensuite **à Pau**.

b. Probablement, Fred ira souvent à Pau.

c. * Souvent, Fred ira probablement à Pau.

Quoi qu'il en soit, le Principe 1, lorsque le connecteur adverbial figure dans une phrase simple, est à interpréter de la façon suivante : la portée sémantique d'un connecteur adverbial est TOUTE sa phrase hôte.

3 Portées syntaxique et sémantique dans les phrases complexes

Considérons maintenant les connecteurs adverbiaux apparaissant dans des phrases complexes comportant un verbe d'attitude propositionnelle (e.g. *croire* considéré comme un « verbe pont » en syntaxe) introduisant une complétive objet, soit des phrases de forme *Conn, N0 V que P* dans lesquelles nous ferons varier la position du connecteur *Conn*. Nous distinguons deux cas selon que le contexte gauche comporte ou non un verbe d'attitude propositionnelle.

Le premier cas est illustré en (5a) enchaînant deux phrases complexes qui comportent chacune un verbe d'attitude propositionnelle. Dans cet exemple, *ensuite* peut être déplacé à l'intérieur du VP de la phrase matrice sans changement de sens, (5b), mais il ne peut pas être déplacé dans la complétive sans induire un changement de sens radical, (5c). Syntaxiquement, en (5a-b), le connecteur est un ajout initial sur la phrase matrice ou un ajout sur un élément du VP de la phrase matrice : celle-ci est donc la phrase hôte de l'adverbial. Sémantiquement, ces discours décrivent la succession temporelle de croyances de Jane concernant les voyages de Fred : la portée sémantique de *ensuite* est donc la phrase matrice (en gras).

(5) a. *Jane a cru que Fred irait à Dax pour Noël.* Ensuite, **elle a cru qu'il irait à Pau**.

b. = *Jane a cru que Fred irait à Dax pour Noël.* **Elle a cru** ensuite **qu'il irait à Pau**.

c. 6= Jane a cru que Fred irait à Dax pour Noël. Elle a cru qu'ensuite il irait à Pau.

En conclusion, le Principe 1 est respecté pour (5a-b). Il en est autrement lorsque le contexte gauche d'une phrase de forme *Conn, N0 V que P* ne comporte pas de verbe d'attitude propositionnelle — ni de marqueur évidentiel de forme prépositionnelle tel que *selon/d'après Nhum* — comme en (6a). Dans cet exemple, le déplacement de *ensuite* à l'intérieur du VP matrice débouche sur une incohérence, (6b), mais son déplacement à l'intérieur de la complétive n'induit pas de différence de sens perceptible, (6c).

(6) a. *Fred ira à Dax pour Noël.* Ensuite, Jane croit qu'**il ira à Pau.**

[1] Ce n'est pas toujours le cas : ainsi certains connecteurs adverbiaux comme *à ce moment-là* ne peuvent figurer qu'en position initiale (Roze, 2009). Par ailleurs, d'autres connecteurs changent de sens selon leur position comme montré pour *alors* par (Bras, 2008).

b. # Fred ira à Dax pour Noël. Jane croit ensuite qu'il ira à Pau.

c. = *Fred ira à Dax pour Noël.* Jane croit qu'ensuite **il ira à Pau**.

Comme *ensuite* ne peut dénoter que la succession temporelle des voyages de Fred, sa portée sémantique en (6a) est la phrase enchâssée, comme c'est le cas en (6c). De ce fait, on peut envisager pour (6a) une analyse par « extraction » : *ensuite* serait extrait de la complétive.[2] Néanmoins, cette analyse ne tient pas lorsqu'on substitue *puis* à *ensuite* : le paradigme en (7) montre que *puis* ne peut se trouver qu'à l'initiale de la phrase matrice, toute autre position étant syntaxiquement interdite.[3] De ce fait, on ne peut pas envisager pour (7a) une analyse où *puis* serait extrait de la complétive.[4]

(7) a. *Fred ira à Dax pour Noël*. Puis, Jane croit qu'**il ira à Pau**.[5]

b. * Fred ira à Dax pour Noël. Jane croit puis qu'il ira à Pau.

c. * Fred ira à Dax pour Noël. Jane croit que puis il ira à Pau.

Une analyse possible pour (7a), qui est aussi valable pour (6a), consiste à considérer que le verbe pont *croit* est un ajout sur la phrase enchâssée, le connecteur étant lui aussi un ajout sur la phrase enchâssée. Cette analyse des verbes ponts, qui est défendue en TAG (Grammaire d'Arbres Adjoints, (Joshi, 1985)), s'appuie sur le fait que (7a) et (8a) décrivent la même situation, comme le font (6a) et (8b), les séquences *d'après Jane* et *croit Jane* étant à l'évidence des ajouts syntaxiques.

(8) a. *Fred ira à Dax pour Noël.* Puis, d'après Jane, **il ira à Pau**.

b. *Fred ira à Dax pour Noël.* Ensuite, croit Jane, **il ira à Pau**.

L'analyse à la TAG des verbes ponts permet donc d'expliquer la portée sémantique des connecteurs adverbiaux en (6a) et (7a) en suivant le Principe 1.[6] Néanmoins, considérons le connecteur *par contre*. Le paradigme en (9) montre que cet adverbial peut apparaître à l'initiale de la phrase matrice, (9a), à l'intérieur de la phrase matrice, (9b), ou à l'initiale de la phrase enchâssée, (9c), sans qu'un changement de position induise un changement de sens clairement perceptible.

(9) a. *Fred ira à Dax pour Noël.* Par contre, Jane croit que **Luc n'ira pas**.

b. = *Fred ira à Dax pour Noël*. Jane croit, par contre, que **Luc n'ira pas**.

c. = *Fred ira à Dax pour Noël*. Jane croit que, par contre, **Luc n'ira pas**.

L'exemple (9b), qui est acceptable contrairement à (6b) avec *ensuite* ou (7b) avec *puis*, pose problème pour le Principe 1 : étant donné la position de *par contre* à l'intérieur du VP matrice, la phrase hôte du connecteur est la phase matrice, or sa portée sémantique se limite à la phrase enchâssée. Certes, on pourrait arguer que sa portée sémantique n'est pas la phrase enchâssée mais la phrase matrice, mais montrons que ceci n'est pas tenable. En effet, une portée large de *par contre* (positionné à l'intérieur du VP matrice) sur la phrase matrice est possible, par exemple en (10a) où il y a conflit d'opinions entre l'auteur et Jane sur le même événement, i.e. le voyage de Fred à Dax pour Noël.[7] Mais nous pensons que *par contre* ne peut pas avoir cette portée large dès que les opinions entre

---

[2] Comme décrit dans (Bonami & Godard, 2007), certains adverbiaux comme les adverbiaux de localisation temporelle (*demain*) peuvent être topicalisés en tête de leur phrase hôte, (i). Il est alors possible d'introduire un verbe pont entre l'adverbial et sa phrase hôte, (ii), sans changer la portée sémantique de l'adverbial qui porte sur la phrase enchâssée, comme en témoigne le fait que (ii) et (iii) décrivent la même situation.

(i) Demain, Fred ira à Pau.

(ii) Demain, Jane croit que Fred ira à Pau.

(iii) = Jane croit que, demain, Fred ira à Pau.

L'adverbial en (ii) est considéré comme extrait de la phase enchâssée. Cette analyse peut s'appliquer mutadis mutandis à (6a). Soulignons toutefois que (Bonami & Godard, 2007) proposent cette analyse par extraction mais soutiennent qu'un connecteur adverbial ne peut pas se trouver à l'initiale d'une phrase sans avoir portée sémantique sur toute cette phrase, ce qui est manifestement faux : voir (6a) où *ensuite* ne porte sémantiquement que sur la phrase enchâssée.

[3] La position de *puis* est contrainte comme celle d'une conjonction de coordination, c'est-à-dire toujours à l'initiale d'une phrase non enchâssée, voir **Jane a puis cru qu'il irait à Dax* et les inacceptabilités de (7b-c). Il est donc envisageable que puis soit une conjonction de coordination et non un adverbial.

[4] Dans (Bonami & Godard, 2007), un complément « extrait » n'est pas obligatoirement acceptable en position « canonique ». Ces auteurs peuvent donc considérer que *puis* est extrait en (7a). Quoiqu'il en soit, leur position sur l'extraction ne change pas le fait qu'un exemple comme (9b) ci-dessous pose problème à l'interface syntaxe-sémantique.

[5] Cet exemple peut paraître plus naturel avec une virgule avant *puis* au lieu du point, changement qui affecte la prosodie.

[6] Cette analyse peut aussi s'appliquer pour les connecteurs non adverbiaux, les conjonctions de subordination, (i), les conjonctions de coordination, (ii).

(i) *Fred ira à Dax pour Noël* alors que Jane croit que **Luc n'ira pas.**

(ii) *Fred ira à Dax pour Noël*. Mais Jane croit que **Luc n'ira pas**.

[7] Le discours en (10a) est amélioré par la présence d'une séquence comme *Je sais que* qui introduit l'opinion (factive) de l'auteur sur le voyage de Fred à Dax pour Noël. Le connecteur ne peut pas être déplacé à l'intérieur de la phrase enchâssée, (i), ce qui va de pair avec une portée large de ce connecteur et le fait que (ii) est mal formé.

l'auteur et Jane concernent des événements (éventualités) différentes. Ainsi, en (10b), il n'y a pas conflit d'opinions sur la gentillesse de Fred — Jane peut penser comme l'auteur que Fred est gentil — mais un contraste entre les qualités de Fred, gentillesse versus radinerie, l'une étant assertée par l'auteur, l'autre soutenue par Jane : *par contre* a portée sémantique sur la phrase enchâssée. Ajoutons qu'un discours construit sur le modèle de (9b) ou (10b), où la première phrase et la phrase enchâssée décrivent des éventualités différentes, n'est cohérent qu'en présence d'un contraste entre ces éventualités : les discours en (10c) et (10d) sont incohérents. Si *par contre* avait une porté large dans ce type de discours, rien n'expliquerait la différence d'acceptabilité entre (9-10b) d'une part et (10c-d) d'autre part. A rebours, en posant que ce connecteur a portée sur la phase enchâssée, on explique cette différence d'acceptabilité par l'absence de contraste en (10c-d) : il n'y a pas de contraste entre les voyages de Fred et Luc en (10c) ou entre les voyages de Fred et Fred avec son chien en (10d). En revanche, il y a contraste entre les voyages de Fred et Luc en (9b) ou entre la gentillesse et la radinerie de Fred en (10b).

(10) a. *(Je sais que) Fred ira à Dax pour Noël.* **Jane croit**, par contre, **qu'il n'ira pas**.

b. *Fred est gentil*. Jane croit, par contre, qu'**il est radin**.

c. # Fred ira à Dax pour Noël. Jane croit, par contre, que Luc ira aussi.

d. # Fred ira à Dax pour Noël. Jane croit, par contre, qu'il ira avec son chien.

En résumé, nous pouvons affirmer que *par contre* en (9b) ou (10b) a portée syntaxique sur la phrase matrice et portée sémantique uniquement sur la phase enchâssée : c'est donc un contre-exemple pour le Principe 1 (contre-exemple pour lequel l'analyse à la TAG des verbes ponts ou l'analyse par extraction n'apporte pas de solution). Nous sommes donc amenée à réviser le Principe 1 selon le principe suivant :

Principe 2 La portée sémantique d'un connecteur adverbial est sa phrase hôte ou une phrase enchâssée dans sa phrase hôte.

Le Principe 2 demande une interface syntaxe-sémantique pas du tout triviale.

4 Principe 1 versus Principe 2 : extension

Deux questions se posent d'emblée concernant le choix du verbe d'attitude propositionnelle, d'une part, et le choix du connecteur de discours, d'autre part :

1. déterminer si, pour un connecteur donné, le choix du verbe d'attitude propositionnelle et de sa polarité a une influence : les exemples présentés à la section précédente avec le connecteur *par contre* ayant portée sémantique sur la phrase enchâssée sont tous construits autour du verbe *croire* — qui est non factif — construit sans marque de polarité négative. A notre connaissance, les données ne changent pas si on utilise d'autres verbes non factifs (*dire, penser*), des verbes factifs (*savoir, ignorer, reconnaître*) ou contre-factifs (*s'imaginer, prétendre*), si tant est qu'un contraste puisse être établi entre les propositions décrivant deux événements distincts (e.g. *Fred ira à Dax* versus *Luc n'ira pas*). Il faut donc faire attention, entre autres, à distinguer un verbe « neg-raising » comme croire d'un verbe factif comme *savoir* : la cohérence d'un exemple comme (11a) repose sur le fait que le second argument sémantique du connecteur reçoit une polarité négative induite par celle syntaxiquement portée par le verbe *croire*, les discours (11a) et (11b) recevant la même interprétation. Avec un verbe factif comme *savoir*, la cohérence de (11c-d) s'oppose à l'incohérence de (11e) due à un manque de contraste entre *Fred ira à Dax pour Noël* et *Luc ira.*

(11) a. *Fred ira à Dax pour Noël*. Jane ne croit pas, par contre, que **Luc ira**.

b. *Fred ira à Dax pour Noël.* Jane croit, par contre, que **Luc n'ira pas**.

c. *Fred ira à Dax pour Noël*. Jane sait, par contre, que **Luc n'ira pas.**

d. *Fred ira à Dax pour Noël*. Jane ne sait pas, par contre, que **Luc n'ira pas**.

e. # Fred ira à Dax pour Noël. Jane sait, par contre, que Luc ira.

Indiquons que le Principe 2 s'applique pour un niveau quelconque d'enchâssement. Ainsi en (12), la portée sémantique de *par contre* est une phrase enchâssée dans une phrase enchâssée dans la phrase hôte du connecteur.

(12) *Fred est gentil.* Jane sait, par contre, qu'on raconte qu'**il est radin**.

2. distinguer les connecteurs adverbiaux qui se comportent comme *par contre* — pour lesquels le Principe 2 doit obligatoirement être adopté — de ceux qui se comportent comme *ensuite* — pour lesquels on peut soit adopter le Principe 2 soit conserver le Principe 1 avec une analyse par extraction ou à la TAG lorsque le connecteur est à l'initiale de sa phrase hôte comme en (6a). Le critère déterminant pour cette distinction est l'acceptabilité versus l'inacceptabilité des discours où le connecteur se trouve à l'intérieur du VP matrice avec portée sémantique sur la phrase enchâssée, voir l'acceptabilité de (9b) versus l'inacceptabilité de (6b). Ainsi on constate que *par exemple* se

(i) * (Je sais que) Fred ira à Dax pour Noël. Jane croit que, par contre, il n'ira pas.
(ii) * Fred ira à Dax pour Noël. Par contre, il n'ira pas.

comporte comme *par contre*, (13a). Il en est de même pour *en réalité*, (13b). En revanche, *en effet* ne peut pas apparaître dans la phrase matrice avec portée sémantique sur la phrase enchâssée : (13c) est inacceptable contrairement à (13d).

(13) a. *Cette banque a fait des investissements imprudents*. Son directeur reconnaît, par exemple, qu'**elle a fait un investissement de 440 millions d'euros en Cratupie**.

b. *Léa soupçonne que son mari la trompe*. Jane pense en réalité qu'**il ne la trompe pas.**

c. # Léa soupçonne que son mari la trompe. Jane pense en effet qu'il la trompe bel et bien.

d. *Léa soupçonne que son mari la trompe*. Jane pense qu'en effet **il la trompe bel et bien.**

Après avoir distingué les connecteurs adverbiaux qui se comportent comme *par contre* de ceux qui se comportent comme *ensuite* ou *en effet*, il faudra examiner s'il y a une quelconque régularité concernant la sémantique de ces classes de connecteurs. Les résultats obtenus devraient permettre de trancher entre les deux solutions pour la classe *ensuite* : garder le Principe 1 ou adopter le Principe 2.

L'annotation entreprise dans le projet FDTB (Danlos et al., 2012) devrait permettre d'apporter des embryons de réponse aux questions précédentes. Elle devrait aussi permettre de confirmer notre intuition qu'il n'existe aucun connecteur adverbial respectant le Principe 3 :

Principe 3 La portée sémantique d'un connecteur adverbial peut être une phrase enchâssant sa phrase hôte.

Si cette intuition est confirmée, on aurait une règle postulant que la portée sémantique d'un connecteur adverbial est égale à sa portée syntaxique ou plus étroite qu'elle. Indiquons que cette règle ne s'applique pas pour un connecteur qui est une conjonction de subordination. Dans (14), la portée sémantique de *parce que* est plus large que sa portée syntaxique.

(14) *Luc est de très mauvaise humeur* parce qu'**il a perdu ses clefs. De plus, il a aussi perdu son permis de conduire.**

5 Conclusion

Les connecteurs adverbiaux, classe bien délimitée d'adverbiaux, posent des problèmes à l'interface syntaxe-sémantique. Outre celui bien connu de la différence entre leur nombre d'arguments sur les plans sémantique et syntaxique, nous avons mis en évidence des exemples où le second argument sémantique du connecteur adverbial ne correspond pas à son argument syntaxique (en étant strictement inclus dans celui-ci), exemples pour lesquels toute solution proposée dans la littérature sur la syntaxe est inopérante. L'extension de ce phénomène demande à être approfondie.

Le problème que nous avons mis en avant pour l'interface syntaxe-sémantique concernant certains connecteurs adverbiaux demande de mettre au point une solution pour pouvoir, par exemple, prolonger un analyseur syntaxique (phrastique) par un analyseur sémantique (discursif).

Faut-il créer la notion de connecteur adverbial flottant par analogie avec celle de quantifieur flottant ?

Références

Bonami O. & Godard D. 2007. Adverbes initiaux et types de phrases en français. In A. Cunita, C. Lupu & L. Tasmowski, Eds., Studii di Lingvistica i Filologie Romanica, p. 50–57. Editura Universitatii din Bucuresti.

Bonami O., Godard D. & Kampers-Manhe B. 2004. Adverb classification. In F. Corblin & H. de Swart, Eds., Handbook of French Semantics, p. 143–184. Stanford : CSLI Publications.

Bras M. 2008. Entre relations temporelles et relations de discours. Université de Toulouse le Mirail : Dossier d'HDR.

Danlos L. 2009. D-STAG : un formalisme d'analyse automatique de discours basé sur les TAG synchrones. Revue TAL, 50(1), p. 111–143.

Danlos L., Antolinos-Bassos D., Braud C. & Roze C. 2012. Vers le FDTB : French Discourse Tree Bank. In Actes de TALN 2012, Grenoble, France.

Forbes-Riley K., Webber B. & Joshi A. 2006. Computing discourse semantics : The predicate-argument semantics of discourse connectives in D-LTAG. Journal of Semantics, 23(1).

Joshi A. 1985. Tree-adjoining grammars. In D. Dowty, L. Karttunen & A. Zwicky, Eds., Natural language parsing, p. 206–250. Cambridge University Press.

Moliner C. 1990. Une classification des adverbes en -ment. Langue française, 88.

PDTB Group 2008. The Penn Discourse Treebank 2.0 Annotation Manual. Rapport interne, Institute for Research in Cognitive Science, University of Philadelphia.

Roze C. 2009. LEXCONN : Base lexicale des connecteurs discursifs du français. Mémoire de Master, Université Paris Diderot.

Roze C., Danlos L. & Muller P. 2012. LEXCONN : a French lexicon of discourse connectives. Discours, 10.

Intransitivité scindée, passif et sujet impersonnel en vietnamien

Huy Linh Dao

Résumé: Le but de cet article est d'examiner de plus près, à travers des données du vietnamien, la distinction inergatif/inaccusatif initialement proposée par Perlmutter (1978) sous le nom de l'Hypothèse d'Inaccusativité. Nous avançons quelques arguments empiriques en faveur de cette bipartition au sein de la classe des verbes intransitifs, en mettant l'accent sur les affinités structurales entre les verbes inaccusatifs et les constructions passives adversatives dans cette langue. Nous suggérons que le marqueur du passif adversatif *bị* peut être traité comme un indice formel de l'inaccusativité en vietnamien contemporain. Ce dernier point est suivi d'une brève discussion sur le rapport reliant les propriétés syntaxiques et sémantiques des inaccusatifs à la légitimation d'un explétif sujet dont la distribution est restreinte en vietnamien parlé aux contextes irrealis-résultatifs. De ces observations se dégage l'idée que les notions de changement d'état et de résultativité s'avèrent particulièrement centrales dans l'opposition entre les inergatifs et les inaccusatifs. Ceci semble se vérifier dans la mesure où certains prédicats inergatifs et statifs, n'encodant pas lexicalement de telles spécifications aspectuelles, peuvent se comporter comme des verbes intrinsèquement inaccusatifs, à condition d'être enrichis de matériels lexicaux prévus à cet effet.

1. Introduction

L'objectif principal de cette présente étude est d'enrichir les recherches portant sur les manifestations syntaxiques et sémantiques de l'Hypothèse d'Inaccusativité (cf. Permmuter, 1978) en la confrontant aux données du vietnamien. Selon cette hypothèse, la classe des verbes intransitifs est hétérogène et doit être divisée en deux sous-groupes, à savoir les inaccusatifs d'un côté et les inergatifs de l'autre. Cette scission repose essentiellement sur le comportement syntaxique et sémantique de l'unique argument d'un verbe intransitif donné, lequel fonctionne comme l'objet, pour les premiers, et comme le sujet, pour les derniers, d'un verbe transitif ordinaire. Si cette hypothèse semble avoir trouvé sa confirmation dans d'importants travaux effectués sur des langues particulières ou dans une perspective translinguistique (cf. Burzio, 1986 ; Zribi-Hertz, 1987 ; Legendre, 1988 ; Levin, Rappaport Hovav, 1995 ; Alexiadou *et al*., 2004, *inter alia*), aucune recherche systématique n'a été, à notre connaissance, menée sur le vietnamien, exception faite de la contribution de Duffield (2011) qui, malgré la discussion centrée sur l'opposition inergatif *versus* inaccusatif dans des constructions causatives de cette langue, prend pour acquise une telle distinction sans pour autant en apporter une démonstration explicite. On pourrait alors se demander jusqu'à quelle mesure l'Hypothèse d'Inaccusativité se vérifie en vietnamien et quelles constructions spécifiques permettent de mettre en évidence cette intransitivité scindée. Dans le cadre de cet article, nous tenterons de montrer que cette bipartition au sein de la classe des intransitifs s'observe effectivement dans cette langue et que l'appartenance d'un verbe intransitif donné à une des deux sous-classes n'est pas entièrement encodée dans le lexique mais se définit pour une grande partie au niveau des structures syntaxiques dans lesquelles ils peuvent être insérés.

Pour ce faire, après un bref rappel, dans la section 2, des principales caractéristiques des verbes inergatifs et inaccusatifs, nous proposons une série de tests afin d'étayer cette division en vietnamien, l'accent étant mis sur deux nouvelles observations empiriques, à savoir que (i) le marqueur du passif adversatif *bị* semble n'être compatible qu'avec les inaccusatifs ; (ii) ceux-ci peuvent apparaître en présence du sujet impersonnel *nó* alors que les inergatifs sont exclus de cette configuration. Dans la section 3, nous tenterons de défendre l'hypothèse que les notions de changement d'état et de résultativité jouent un rôle important dans la distinction inergatif/inaccusatif en vietnamien et que celle-ci constitue plutôt un continuum qu'une opposition binaire. Pour cela, nous examinerons a) quelques verbes intransitifs dénotant des manières de mouvement et b) certains prédicats statifs (verbes de qualité correspondant aux adjectifs d'une langue comme le français). Ces deux types de prédicats, quand ils sont enrichis de matériels lexicaux spécifiant l'état résultant final dans lequel se trouve leur unique argument, peuvent se comporter syntaxiquement comme les inaccusatifs de base, en vertu des tests avancés dans la première partie.

2. Quelques arguments en faveur de l'Hypothèse d'Inaccusativité en vietnamien

2.1. Principales caractéristiques des verbes inergatifs et inaccusatifs

L'extrême abondance de la littérature dévolue à cette dichotomie rendant toute tentative de récapitulation ici délicate et nécessairement déficiente, nous nous bornons à résumer dans le tableau ci-dessous les principales propriétés qui séparent les inergatifs des inaccusatifs. Il convient de noter que les recherches conduites dans le cadre des travaux adoptant l'Hypothèse d'Inaccusativité sont, en effet, allées au-delà du débat initial, centré sur les différents comportements syntaxiques et sémantiques de l'unique argument[1] d'un verbe intransitif donné, en établissant un lien entre la sous-classe à laquelle appartient ce dernier et les propriétés aspectuelles (lexicales) du groupe verbal qu'il forme avec son argument (cf. Tenny, 1994 ; Alexiadou *et al.*, 2004 ; Borer, 2005).

[1] La distinction entre inergatif et inaccusatif dépasse le cas des verbes mono-argumentaux et la discussion a été étendue aux prédicats dyadiques (cf. Belletti, Rizzi, 1988 ; Landau, 2010) et au domaine adjectival (Cinque, 1990 ; Bennis, 2004 ; Anscombre, 2005 ; Tayalati, 2008). Pour des raisons d'espace, nous ne les abordons pas ici.

		Inergatif	Inaccusatif
Argument unique	Rôle sémantique	Proto-agent	Proto-patient
	Comportement syntaxique	Sujet d'un verbe transitif (= Argument externe)	Objet d'un verbe transitif (=Argument interne)
Aspect lexical (Aktionsart) (Classes aspectuelles)		Atélique (Activités)	(A)télique[2] (Etats, achèvements, accomplissements)

Tableau 1 : Principales caractéristiques des verbes inergatifs et inaccusatifs

2.2. Inaccusativité versus Inergativité en vietnamien : quelques arguments empiriques

2.2.1. Positionnement de l'argument unique

Le vietnamien étant une langue isolante à ordre SVO, l'objet canonique suit donc normalement le verbe. L'unique argument nominal d'un verbe inaccusatif peut se trouver dans une position préverbale ou postverbale (ex. 1 et 2), alors que celui d'un verbe inergatif ne peut apparaître que préverbalement (ex. 3, 4) :

(1) a. **Chuyện** **gì** đã xảy ra thế ?
Histoire quoi PERF se passer PART-Inter
b. Đã xảy ra **chuyện** **gì** thế ?
PERF se passer histoire quoi PART-Inter
'Qu'est-ce qui s'est passé ?'

(2) a. **Nhà** cháy rồi kìa !
Maison brûler PERF DEICT
b. Cháy **nhà** rồi kìa !
Brûler maison PERF DEICT
'(Regarde), la maison a brûlé !'

(3) a. **Thằng kẻ trộm** chạy rồi kìa !
Voleur courir PERF DEICT
b. *Chạy **thằng kẻ trộm** rồi kìa !
Courir voleur PERF DEICT
'(Attention), le voleur s'est sauvé/a couru !'

(4) a. **Paul** đã nhảy/hát/hò hét suốt đêm.
P. PERF danser/chanter/hurler durant nuit
b. *Đã nhảy/hát/hò hét **Paul** suốt đêm.
PERF danser/chanter/hurler P. durant nuit
'Paul a dansé/chanté/hurlé toute la nuit'

La possibilité d'occuper la position postverbale de l'unique argument des verbes comme *xảy ra* 'se passer' ou *cháy* 'brûler' (intr.) (ex. 1-2), qui sont clairement non agentifs, permet de le rapprocher de l'objet d'un verbe transitif ordinaire. A l'inverse, les verbes tels que *chạy* 'courir/se sauver', *nhảy* 'danser', *hát* 'chanter', *hò hét* 'hurler' dans les exemples (3-4) sélectionnent tous un sujet volitionnel, réalisé uniquement dans la position préverbale, comportement typiquement associé au sujet d'un verbe transitif des langues SVO.

2.2.2. Sous-extraction du noyau nominal d'un SN quantifié

En vietnamien, il est possible d'extraire le noyau nominal d'un syntagme nominal (SN) complexe en position postverbale et de l'antéposer au verbe, laissant derrière le matériel lexical restant. Cette situation s'observe avec les verbes inaccusatifs (ex. 5). Une telle opération syntaxique donne cependant lieu à l'agrammaticalité dans le cas des inergatifs (ex. 6).

(5) a. Trước đây, ở trang trại này từng chết **hàng chục con bò**.
Avant DEICT LOC ferme DEICT EXP mourir PL dix CL bœuf

[2] Les inaccusatifs statifs qui dénotent des états comme *exister*, *rester*, etc. sont atéliques.

b. Trước đây, ở trang trại này **bò** từng chết **hàng chục con.**
Avant DEICT LOC ferme DEICT bœuf EXP mourir PL dix CL
'Avant, dans cette ferme, il est arrivé que des dizaines de bœufs soient morts'

(6) a. **Năm người bạn** đang sôi nổi thảo luận trước cửa thư viện.
Cinq CL ami PROG vif discuter devant porte bibliothèque
b. ***Bạn năm người** đang sôi nổi thảo luận trước cửa thư viện.
Ami cinq CL PROG vif discuter devant porte bibliothèque
'Cinq amis sont en train de discuter vivement devant la bibliothèque'

Dans (5a), l'unique argument nominal du verbe *chết* 'mourir', à savoir le SN complexe *hàng chục con bò* 'des dizaines de bœufs', apparaît en position postverbale. Le noyau de celui-ci, en l'occurrence le nom *bò* 'bœuf', est extrait de l'ensemble et précède le verbe dans (5b). Le caractère mal formé de (6b) suggère que l'extraction, illicite avec un verbe agentif comme *thảo luận* 'discuter', n'est possible qu'à partir de la position d'objet. Il est à noter en effet que le SN complexe argument de *chết* 'mourir' peut aussi être réalisé dans son intégralité dans la position préverbale mais que l'extraction depuis ce site est difficile, voire impossible.

2.2.3. Construction à possession externe

Chapelle (1999), dans son étude portant sur plusieurs langues sinitiques, fait observer que la construction qu'elle appelle *Double Unaccusative*, schématisée dans (7), ne peut prendre comme noyau prédicatif qu'un verbe inaccusatif (ex. 8). Cette tournure, qui s'observe aussi en vietnamien, exclut systématiquement les inergatifs (ex. 9). Elle comporte deux SN qui entretiennent entre eux une relation de type partie/tout, le premier étant le plus souvent interprété comme le possesseur du second :

(7) SN1 V_{INACC} SN2
POSSESSEUR/TOUT POSSESSUM/PARTIE

(8) a. **Cái cây này** rụng nhiều **lá** thật !
CL arbre DEICT tomber beaucoup feuille vraiment
'Cet arbre a vraiment perdu beaucoup de feuilles'
'Litt. (De) cet arbre sont tombées beaucoup de feuilles'

b. **Anh chàng này** gẫy cả **hai tay**.
Jeune homme DEICT se casser tout deux bras
'Ce jeune homme s'est cassé les deux bras'

(9) * **Anh chàng này** hét **cổ họng.**
Jeune homme DEICT hurler gorge
'#La gorge de ce jeune homme a hurlé/ce jeune homme a hurlé sa gorge'

Cette construction nous rappelle le phénomène d'extraction précédemment abordé, dans le sens où le SN1 semble former avec le SN2 un SN complexe fonctionnant comme l'unique argument du verbe inaccusatif[3], dans lequel le premier SN est un sous-constituant syntaxique et un argument sémantique du second SN. La situation est pourtant inversée ici, car le vrai argument du verbe reste dans la position postverbale alors que dans le cas de l'extraction, il est réalisé préverbalement.

2.3. Inaccusativité, passif et sujet impersonnel

2.3.1. Inaccusativité et le marqueur du passif adversatif bị en vietnamien

Nous tenterons de montrer dans cette section que le mot *bị*, qui sert à former le passif adversatif en vietnamien, peut être qualifié de marqueur potentiel de l'inaccusativité dans cette langue. Comme nous le montrent les exemples (10b) et (11b), *bị* permet de construire, à partir des phrases actives correspondantes (a), deux types de passif, à savoir le passif *direct* et le passif *possessif*, où le sujet passif est interprété respectivement comme l'objet direct et le possesseur de l'objet direct du verbe transitif dans les contreparties actives :

(10) a. Marie đánh Paul
M. frapper P.
'Marie a frappé Paul'
(**Actif**)

b. Paul **bị** Marie đánh
P. BI M. frapper
'Paul a été frappé par Marie'
(**Passif direct**)

[3] Une telle situation est attestée, étant donné la grammaticalité des variantes « non discontinues » de (8a-b) :
(i) **Lá cái cây này** rụng nhiều thật (= 8a)
(ii) **Hai tay anh chàng này** gẫy cả. (=8b)

(11) a. Paul trộm tiền của Marie b. Marie **bị** Paul trộm tiền
P. voler argent POSS M. M. BI P. voler argent
'Paul a volé de l'argent à Marie' 'Marie s'est fait voler de l'argent par Paul'
(**Actif**) (**Passif possessif**)

Le traitement de l'inaccusativité dans le cadre génératif reconnaît, depuis les travaux de Burzio (1986), les formes verbales passives comme un sous-type de verbes inaccusatifs. Cette assimilation repose sur l'analogie structurale que partagent les deux constructions : le sujet formel d'un verbe inaccusatif et celui d'une phrase passive ont été l'objet sous-jacent à une étape de la dérivation syntaxique. Il ne serait donc pas étonnant de constater ici les affinités entre le passif avec *bị* et les différents patrons inaccusatifs décrits *supra*. Ainsi la Construction Inaccusative Double (cf. section 2.2.3) semble-t-elle se rapprocher du passif possessif. Les similitudes vont plus loin encore : *bị* est incompatible avec les inergatifs, comme en témoigne l'agrammaticalité de (12) :

(12) * Pierre **bị** chạy /nhảy / hát /ngủ rồi.
P. BI courir/danser/chanter/dormir PERF
'Pierre a couru/dansé/chanté/dormi'

Ceci contraste fort avec le fait que les inaccusatifs se combinent sans difficulté avec *bị*. Observons :

(13) a. Nhà cháy rồi kìa ! a'. Nhà **bị** cháy rồi kìa !
Maison brûler PERF DEICT Maison BI brûler PERF DEICT
'(Regarde), la maison a brûlé'
b. Cháy nhà rồi kìa. b'. **Bị** cháy nhà rồi kìa.
Brûler maison PERF DEICT BI brûler maison PERF DEICT
'(Regarde), la maison a brûlé'

Bị peut également apparaître dans la Construction Inaccusative Double (ex. 14b/b') :

(14) a. Cái bình của bố tôi vỡ rồi.
CL vase POSS père 1SG se-casser PERF
a'. Cái bình của bố tôi **bị** vỡ rồi.
CL vase POSS père 1SG BI se-casser PERF
'Le vase de mon père s'est cassé/s'est brisé'
b. Bố tôi vỡ cái bình rồi.
Père 1SG se-casser CL vase PERF
b'. Bố tôi **bị** vỡ cái bình rồi.
Père 1SG BI se-casser CL vase PERF
'Le vase de mon père s'est cassé/s'est brisé'
'Litt. Mon père subit le fait que son vase s'est cassé'[4]

(15) a. Cái cây này **bị** rụng nhiều lá thật ! (=8a)
CL arbre DEICT BI tomber beaucoup feuille vraiment
'Cet arbre a vraiment perdu beaucoup de feuilles'
'Litt. (De) cet arbre sont tombées beaucoup de feuilles'
b. Anh-chàng này **bị** gẫy cả hai tay. (=8b)
Jeune homme DEICT BI se casser tout deux bras
'Ce jeune homme s'est cassé les deux bras'

Nous avons vu dans la section (2.2.2) que la sous-extraction du noyau nominal à partir d'un SN complexe est possible seulement avec les inaccusatifs. Le même phénomène s'observe avec une phrase passive construite à l'aide de *bị* :

(16) a. Người-ta giết **năm con bò**. (actif)
On abattre cinq CL bœuf
'On a abattu cinq bœufs'
b. **Năm con bò** bị người-ta giết. (passif direct)
Cinq CL bœuf BI on abattre
'Cinq bœufs ont été abattus'

[4] Nous n'entrons pas dans les détails des différences liées à la structure informationnelle entre (14a-a') et (14b-b'), qui méritent une discussion plus approfondie.

c. **Bò** bị người-ta giết **năm con** (passif direct avec extraction)
Bœuf BI on abattre cinq CL
'Quant aux bœufs, il y en a cinq d'abattus'/ cinq en ont été abattus'

(17) a. Bị rách **ba cái áo**. b. **Áo** bị rách **ba cái**.
BI être-déchiré trois CL chemise Chemise BI être-déchiré trois CL
'Il y a trois chemises de déchirées'
'Quant aux chemises, il y en a trois de déchirées'

Les ressemblances structurales entre les structures passives avec *bị* et les verbes inaccusatifs présentées ci-dessus, si elles sont correctes, permettent de motiver davantage l'hypothèse selon laquelle *bị* serait le marqueur potentiel de l'inaccusativité en vietnamien et corroborent la distinction inergatif/inaccusatif dans cette langue.

2.3.2. Inaccusativité et sujet impersonnel

En vietnamien, la forme pronominale *nó* '3SG' peut avoir une interprétation non référentielle dans les contextes irrealis-résultatifs. Cet emploi explétif de *nó* ne s'obtient qu'en position de sujet. Seuls les verbes inaccusatifs sont permis dans ces constructions. Comparons :

(18) a. Đừng làm thế, **nó** chết **mấy con cá của tao** bây giờ.
NEG faire ainsi Expl mourir PL CL poisson POSS 1SG maintenant
b. Đừng làm thế, **mấy con cá của tao** chết bây giờ.
NEG faire ainsi PL CL poisson POSS 1SG mourir maintenant
'Ne fais pas ça ! Sinon, mes poissons vont tous mourir'

(19) a. Mày làm thế, **nó** vỡ **cái bình cổ của bố** bây giờ.
2SG faire ainsi Expl se-casser CL vase ancien POSS père maintenant
b. Mày làm thế, **cái bình cổ của bố** vỡ bây giờ.
2SG faire ainsi CL vase ancien POSS père se-casser maintenant
'Si tu fais ça, le vase du père va se casser'

(20) a. Anh làm thế, **thằng bé** chạy/khóc/hét/ngã bây giờ.
2SG faire ainsi garçon courir/pleurer/crier/tomber maintenant
b. * Anh làm thế, **nó** chạy/khóc/hét **thằng bé** bây giờ.
2SG faire ainsi Expl courir/pleurer/crier garçon maintenant
'Si tu fais ça, le garçon va courir/pleurer/crier'
c. Anh làm thế, **nó** ngã **thằng bé** bây giờ.
2SG faire ainsi Expl tomber garçon maintenant
'Si tu fais ça, le garçon va tomber'

Les phrases (18-20) dénotent toutes des situations complexes que l'on pourrait appeler « irrealis-résultatives ». Elles se composent de deux membres, séparés formellement par la virgule (ou une pause à l'oral). La relation qui s'établit entre les deux peut être interprétée comme celle reliant la protase à l'apodose d'un système conditionnel. Les exemples (18) et (19) comportent des verbes inaccusatifs de type « achèvement » (cf. Vendler, 1967), à savoir *chết* 'mourir', *vỡ* 'se casser'. Ceux-ci sont des prédicats téliques décrivant des changements d'état et spécifiant lexicalement l'état résultant final dans lequel se trouve leur argument unique. La présence de l'explétif sujet *nó* est licite quand le vrai argument nominal de ces verbes se réalise postverbalement. Tel n'est pas le cas dans (20b) où les verbes inergatifs comme *chạy* 'courir', *khóc* 'pleurer', *hét* 'crier', dénotant des activités non délimitées temporellement et donc ne spécifiant pas lexicalement un état résultant final, semblent requérir que leur position sujet soit occupée par leur argument nominal unique (cf. ex. 20a). Cette position n'étant plus libre, l'insertion de l'explétif *nó* entraîne l'agrammaticalité de la phrase entière. Si la position préverbale doit nécessairement être remplie pour les inergatifs, elle ne l'est qu'optionnellement dans le cas des inaccusatifs, comme nous l'indiquent (20a) et (20c) : le SN *thằng bé* 'le garçon' peut précéder le verbe *ngã* 'tomber' dans (20a) ou le suivre dans (20c). L'introduction du sujet explétif est rendue possible et donne un résultat parfaitement bien formé dans ce dernier cas.

La légitimation de l'explétif sujet *nó* en présence de verbes inaccusatifs tels que *chết* 'mourir', *vỡ* 'se casser'et *ngã* 'tomber' laisse supposer que les notions de changement d'état et de résultativité peuvent être caractéristiques de l'inaccusativité. Nous avons observé qu'un état résultant final est directement inférable du sens lexical de ces verbes. Il paraît donc naturel de se demander si certains prédicats, qui n'encodent pas dans leur sens une telle spécification aspectuelle mais qui l'acquièrent via l'ajout explicite d'éléments lexicaux, peuvent être qualifiés d'inaccusatifs. Si cela est effectivement le cas, il sera légitime de penser que l'appartenance à l'une ou à l'autre

sous-classe des verbes intransitifs se décide en syntaxe et que l'opposition inergatif/inaccusatif cesse d'être binaire et constitue plutôt un continuum.

3. Changement d'état, résultativité et la distinction inergatif/inaccusatif

A la question de la nature compositionnelle de l'inaccusativité en vietnamien, la réponse que nous tentons d'apporter semble être positive, du moins pour ce qui nous concerne ici. Pour des raisons d'espace, nous nous limitons dans cette section à deux types de situations qui permettent de vérifier cette hypothèse.

3.1. Verbes de manières de mouvement

Tous les arguments antérieurement avancés militent en faveur du statut inergatif des verbes de manières de mouvement comme *chạy* 'courir', *nhảy* 'danser', etc. Ainsi *chạy* 'courir' ne peut-il se combiner avec le marqueur du passif adversatif *bị* (ex. 21a) ou entrer dans une structure à possesseur externe (cf. Double Unaccusative Construction) (ex. 21b). L'ajout de la particule résultative *mất* 'perdre, disparaître, ne plus être présent' à ce verbe permet de former un prédicat complexe doté d'une structure événementielle biphasée dans laquelle la particule sert à préciser l'état final résultant de l'action dénotée par le verbe *chạy* 'courir'. L'ensemble *chạy mất* peut donc être paraphrasé par « courir jusqu'à disparaître, au point de ne plus être vu ». Cet enrichissement aspectuel semble aller de pair avec l'acquisition des propriétés syntaxiques d'un verbe inaccusatif de base, comme nous le montrent (21a') et (21b') :

(21) a. *Ba con gà bị chạy. b. *Paul (bị) chạy ba con gà.
Trois CL poule BI courir P. BI courir trois CL poule
'Trois poules ont couru' 'Trois poules de Paul ont couru'

a' Ba con gà bị chay mất. b. Paul (bị) chay mất ba con gà.
Trois CL poule BI courir-perdre P. BI courir-perdre trois CL poule
'Trois poules se sont égarées' 'Parmi les poules de Paul, trois s'en sont égarées'

En outre, le SN *ba con gà* 'trois poules' peut se réaliser dans une position postverbale, tout comme l'argument unique d'un verbe inaccusatif simple (ex. 22). Son noyau nominal *gà* 'poule' peut être extrait et antéposé au prédicat complexe *chạy mất* (ex. 23). Finalement, ce dernier peut légitimer l'explétif sujet *nó* dans des contextes irrealis-résultatifs (ex. 24).

(22) Chạy mất **ba con gà** rồi !
Courir-perdre trois CL poule PERF
'Trois poules se sont (déjà) égarées'

(23) **Gà** chay mất **ba con** rồi !
Poule courir-perdre trois CL PERF
'Trois poules se sont (déjà) égarées'

(24) Anh quên đóng cửa là **nó** chay mất **mấy con gà của em** đấy !
2SG oublier fermer porte COP Expl courir-perdre PL CL poule POSS 1SG DEICT
'Si tu oublies de fermer la porte, mes poules vont s'égarer !'

3.2. Prédicats statifs (verbes de qualité correspondant aux adjectifs du français)

Un verbe de qualité comme *đen* 'être noir' peut être « transformé » en un prédicat inchoatif complexe dénotant un changement d'état non-volitionnel et cela au moyen de particules aspectuelles comme *đi* 'aller', *hết-cả* 'finir-tout/complètement'. A la différence du cas précédent, les prédicats complexes construits ont une structure événementielle monophasée car ils décrivent le passage d'un état initial (absence d'une propriété X) à un autre (acquisition de cette nouvelle propriété), sans événement causateur explicité. Les combinaisons *đen + đi* 'devenir noir, (se) noircir' et *đen + hết cả* 'devenir complètement noir' se comportent comme des inaccusatifs simples, ce que nous pouvons observer dans (25-26) :

(25) a. Da cô ấy **bị** đen đi. b. Cô ấy **bị** đen da đi.
Peau 3SGFem BI noir-aller 3SGFem BI noir peau aller
'Sa peau s'est noircie'

(26) a. Thằng bé đi nắng nhiều nên da **bị** đen hết cả !
Garçon aller soleil beaucoup alors peau BI noir-finir-tout
'Le garçon s'est tellement exposé au soleil que sa peau s'est noircie complètement'

b. Thằng bé **bị** đen hết cả da.
Garçon BI noir-finir-tout peau
'La peau du garçon s'est complètement noircie'

c. Trời nắng thế này, ra đường để **nó** đen hết cả da à ?
Ciel soleil ainsi sortir rue pour Expl noir-finir-tout peau PART-Int
'(Tu ne vois pas ça ?) Le soleil tape fort ! Tu veux qu'on sorte pour que nos peaux deviennent complètement noires ?'

4. Conclusion

Dans le présent travail, nous avons cherché à montrer que l'Hypothèse d'Inaccusativité, qui consiste à scinder la classe des verbes intransitifs en deux sous-groupes : les inaccusatifs et les inergatifs, se vérifie aussi en vietnamien. Nous avons avancé quelques arguments empiriques en faveur de cette subdivision en faisant observer de plus près le lien entre l'inaccusativité, la légitimation de l'explétif sujet *nó* dans des contextes irrealis-résultatifs et le marqueur du passif adversatif *bị*. De ces observations se dégage l'idée que ce dernier peut être traité comme un marqueur de l'inaccusativité en vietnamien. Un examen approfondi des verbes de manières de mouvement et de certains prédicats statifs (verbes de qualité) révèle que ceux-ci, une fois enrichis de matériels lexicaux spécifiant un changement d'état et un état résultant final, s'approprient des propriétés syntaxiques normalement associées aux inaccusatifs. Ces observations corroborent l'idée que l'appartenance à l'une ou à l'autre sous-classe n'est donc pas exclusivement dépendante du lexique mais se décide en partie en syntaxe. Ceci nous amène également à réévaluer l'hypothèse d'une opposition binaire entre inergatif et inaccusatif et à favoriser une analyse en termes de continuum.

Abréviations

1SG : 1ère personne, singulier ; 2SG : 2ème personne, singulier ; 3SGFem : 3ème personne, singulier, féminin ; CL : classificateur ; DEICT : déictique ; EXP : aspect expérientiel ; Expl : explétif ; LOC : locatif ; NEG : négation ; PART-Inter : particule interrogative ; PERF : aspect perfectif ; PL : pluriel ; POSS : possession ; PROG : aspect progressif.

Références

Alexiadou Artemis.; Anagnostopoulou Elena.; Everaert Martin (éds.) (2004), The Unaccusativity Puzzle. Explorations of the Syntax–Lexicon Interface, Oxford, Oxford University Press.

Anscombre Jean-Claude (2005), Temps, aspect et agentivité, dans le domaine des adjectifs psychologiques, LIDIL 32.

Belletti Adriana; Rizzi Luigi (1988), Psych-Verbs and the theta-Theory, Natural Language and Linguistic Theory 6, p. 291-352.

Bennis Hans. (2004), Unergative adjectives and psych verbs. In : Alexiadou Artemis.; Anagnostopoulou Elena.; Everaert Martin. (éds.), The Unaccusativity Puzzle. Explorations of the Syntax–Lexicon Interface, Oxford, Oxford University Press, p. 84-114.

Borer Hagit (2005), Structuring sense. Volume II: The normal course of Events, Oxford, Oxford University Press.

Burzio Luigi (1986), Italian Syntax, Dordrecht, Reidel.

Chappell Hilary (1999), The Double Unaccusative Construction in Sinitic Languages, In : PAYNE Doris.-L., Barshi Immanuel (Eds.), External Possession, Amsterdam, John Benjamins, p. 195-228.

Cinque Guglielmo (1990), Ergative Adjectives and the Lexicalist Hypothesis, Natural Language and Linguistic Theory 1, p. 1-39.

Duffield Nigel (2011), Unaccusativity in Vietnamese and the Structural Consequences of Inadvertent Causes, In : Folli Raffaella & Ulbrich Christiane (éds.), Researching interfaces in linguistics, Oxford, Oxford University Press.

Landau Idan (2010), The locative syntax of experiencers, Cambridge MA, MIT Press.

Legendre Géraldine (1988), Two Classes of Unergatives in French?, CLS 24, p. 259-274.

Levin Beth ; Rappaport Hovav Malka (1995), Unaccusativity: At the syntax-lexical semantics interface, Cambridge (Mass.), MIT press.

Perlmutter Davis (1978), Impersonal passives and the Unaccusative Hypothesis, Proceedings of the Berkeley Linguistic Society 4, p. 57-189.

Tayalati Fayssal (2008), La distinction ergatif /inergatif et son incidence sur le placement des clitiques datifs dans les constructions causatives avec Faire et Rendre, Probus 20/2, p. 301-321.

Tenny Carol (1994), Aspectual roles and the syntax-semantics interface, Dordrecht: Kluwer.

Vendler Zeno (1967), Linguistics in Philosophy, Ithaca & London, Cornell University Press.

Zribi-Hertz Anne (1987), La réflexivité ergative en français moderne, *Le français moderne* LV, 1-2, p. 23-54.

In search of knowledge: text mining dedicated to technical translation

Annibale Elia, Alberto Postiglione, Mario Monteleone, Johanna Monti, Federica Marano

Abstract: Although a vast amount of contents and knowledge has been made available in electronic format and on the web in recent years, translators still do not have friendly and targeted tools at their disposal for the various aspects of a translation process, i.e., the analysis phase, automatic creation and management of the linguistic resources needed and automatic updating with the relevant information generated by the computer translation tools used in the process (Machine Translation, Translation Memories, and so on).

Text mining and information retrieval are not typically connected with the translation process and no existing online translation workspace integrates text mining or information retrieval facilities that are specifically aimed at improving the documentary competence of translators in order to process unstructured (textual) information, and make the information on the web or in texts accessible to translators.

This paper explores a new approach to helping translators look for different types of information (glossaries, corpora, Wikipedia, and so on) related to the specific translation work they have to perform which can then be used to update the lexical base needed for the translation workflow (both human or machine-aided). This new approach is based on CATALOGA, a text mining tool, which can be combined with an IR application and/or an MT/TM system and used for different purposes.

1. INTRODUCTION

Information acquisition is a very crucial aspect in the translation workflow that has been underlined by many scholars in the field of translation theory and practice. However, up to now, very little help has come from the IT community to make this task faster and easier.

Information extracted from the web or from texts has a valuable role to play in the analysis phase of a translation process (whether human or computer-assisted) since it helps detect typical translation reference material (texts, glossaries, dictionaries, comparable or parallel corpora and so on), identify the subject domain of texts and relevant concepts and terminology and detect similarities between documents or how they are related to other variables of interest.

Translators spend a lot of time on the preliminary phases of translation and this is particularly true when working on technical translations.

Surprisingly, computer translation tool producers focus mainly on the post-editing of raw machine translations and completely ignore the initial phases of the translation workflow, i.e., the search for information needed for the translation task and, for technical translations, mainly terminological compound words; if done properly, this can lead to a significant improvement in the translation engines and can speed up the whole translation process from the very beginning, i.e., analysis of the source text.

This paper explores a new approach to helping translators look for different types of information (glossaries, corpora, Wikipedia, and so on) related to the specific translation work they have to perform which can then be used to update the lexical base needed for the translation workflow (both human or computer-aided). This new approach is based on the combination of CATALOGA, a text mining tool, an IR application and/or an MT/TM system.

2. CATALOGA

2.1 General features of CATALOGA

CATALOGA is a text mining software based on matching digitised texts and electronic dictionaries[1] of terminological compound words developed according to the Lexicon-Grammar (LG) lexical formalization method[2]. The monolingual version was designed and developed by Elia, Postiglione and Monteleone and the bilingual version is under development.

At present, it is configured as a stand-alone software which can be integrated in web sites and portals. Its main goal is to extract terminological compound words from a given scientific or technical text and to automatically determine – without human intervention – the main knowledge domains it deals with and detect the terminological compound words in the analysed texts.

The tasks performed by the monolingual version of CATALOGA can be summarised as follows:

1. automatic reading of a text;
2. computation of all the occurring terminological multi-word units, i.e., location and computation of all the occurrences of any of a finite number of compound words;

[1] The electronic dictionaries used by CATALOGA are part of the DELA system, developed according to the Lexicon Grammar (LG) approach. This system is formed by Simple-Word Electronic Dictionaries (DELAS-DELAF), and Compound-Word Electronic Dictionaries (DELAC-DELACF). CATALOGA uses the DELAC-DELACF dictionary which includes mainly terminological compound nouns. Each entry of the dictionaries is given a consistent ontological description, being coherently tagged with reference to the knowledge domain(s) in which it is commonly used (i.e., in which it has a terminological unambiguous meaning). For more information see ELIA Annibale; POSTIGLIONE Alberto; MONTELEONE Mario, (2011)

[2] For more information about the Lexicon-Grammar approach refer to http://infolingu.univ-mlv.fr/english/Bibliographie/biblieng.html.

3. statistical computation of the ratio between terminological and non-terminological occurrences;
4. statistics-based listing of all the terminological occurrences in decreasing order, classed on the basis of the relevant knowledge domains.

2.2 The bilingual version of CATALOGA

The initial phases of a scientific or technical translation process imply several tasks that have to be performed by translators i.e., reading of the source text, identification of the main concepts and relevant terminology, documentary search using traditional documentary tools (paper dictionaries, thesauri, etc.) or web pages on the Internet, use of general, and specialized, monolingual, bilingual and multilingual electronic dictionaries on the Internet or on CD-ROM, consulting reference material provided by the customer or text corpora on the Internet or on CD-ROM, looking up information in a personal text corpus by means of text analysis or concordance software programs and updating and tailoring the linguistic resources or the translation tools according to the specific translation task that has to be performed. No tools are available on the market that speed up these complex and time-consuming activities.

The approach we would like to propose here is to introduce a higher degree of automation and integration for this crucial phase of the translation cycle which could also be beneficial to the subsequent translation phase.

An ideal documentary tool, in this respect, should contain a text mining and information extraction facility from corpora which enables:

- document classification (identification of domain and extraction of relevant concepts) and automatic indexing based on linguistic information;
- retrieval of useful reference material by users such as appropriate terminology resources, parallel corpora, etc. which are automatically assigned to a specific translation project;
- pre-translation of the source text and /or updating of the translation tools (both MT and TM) with the relevant information found during the query phase.

This tool would allow users to semi-automate the translation analysis phase with regard to the retrieval of reference material (documents, terminology, corpora) for a particular translation project. Unlike state-of-the art collaborative translation workspaces, this would provide an advanced and indispensable feature based on linguistic knowledge within a typical translation workflow.

With this idea in mind, we are developing the bilingual version of CATALOGA. In addition to the features described for the monolingual version, the main features of the bilingual version of Cataloga are:

1. listing of all the terminological occurrences in decreasing order classed on the basis of the relevant knowledge domains with their translation;
2. tagging of all the terminological compound words with their translation in the source text in XML format;
3. automatic replacement of the translations in the target text.

These features are very useful for different purposes:

1. the list of words obtained at the end of the text analysis process can be used in a specific crawling tool, such as BootCat[3] for instance, to automatically retrieve useful reference material such as parallel or comparable corpora;
2. the tagged text can be used for training purposes in conjunction with MT, in specific SMT and TM applications, in order to identify and pre-translate linguistically significant phrases, with the aim of improving the computer-assisted translation results;
3. the pre-translated target text can be used as a basis during a traditional human-based translation cycle.

In the following section, we provide the results of some experiments performed with CATALOGA both on Italian texts and their translations and some examples of the possible use of CATALOGA in the initial stages of a scientific or technical translation process.

[3] Bootcat (http://bootcat.sslmit.unibo.it/?section=home) is an open source crawling tool that creates random tuples from a seed term list and runs a query for each tuple (on the Bing search engine). It constructs a URL list on the basis of the first 10 results obtained from the query and downloads the corresponding web pages. Bootcat is also available on the Sketchengine webpage (http://www.sketchengine.co.uk/?page=Website/SketchEngine) and as BootCat front-end, a web service front-end and a graphical user interface to the core tool, respectively.

3. USE OF CATALOGA IN A SCIENTIFIC/TECHNICAL TRANSLATION PROCESS

3.1 Text analysis performed by CATALOGA: results

In order to provide a concrete example of how CATALOGA processes texts and automatically extracts meanings, we will consider the following short passage which a human reader with an average cultural level could straightforwardly define as dealing with the field of medicine:

> *La vitamina A (Retinolo) svolge un'azione protettiva delle mucose e degli epiteti. Inoltre ha un ruolo nella crescita, favorendo lo sviluppo scheletrico. La carenza di vitamina A è una delle più comuni carenze vitaminiche. È comune soprattutto nei Paesi in via di sviluppo, rappresentando una della principali cause di cecità. La carenza di vitamina A è spesso dovuta a malassorbimento lipidico, ad alcolismo, e si osserva più comunemente negli anziani. Un sintomo precoce di carenza di vitamina A è la cecità notturna, seguita da secchezza della congiuntiva, macchie di Bitot (macchie biancastre della sclera). Questa risposta fatta da me su altro sito le fa capire a che cosa è dovuta la macchia di Bitot e di che colore è ovvero biancastro. La sua sembra più o un piccolo nevo nevocellulare piano oppure una zona di assottigliamento sclerale, completamente innocua e sine materia dal punto di vista patologico, che lascia intravedere la componente bluastra sottostante. (Available on http://www.medicitalia.it/consulti/Oculistica/65819/Macchianella- sclera)*[4]

After reading and analysing it, CATALOGA automatically produces a table with the results of the text processing:

[4] Vitamin A (Retinol) exerts a protective action on the mucous membranes and epithets. It also has a role in growth, supporting skeletal development. Lack of vitamin A is one of the most common vitamin deficiencies. It is especially common in developing countries, representing one of the main causes of blindness. Vitamin A deficiency is usually due to fat malabsorption and alcoholism and is most commonly seen in elderly people. An early sign of vitamin A deficiency is night blindness, followed by dryness of the conjunctiva and Bitot's spots (white spots on the sclera). This answer I gave on another site helps you understand the origins of Bitot's spots, and their colour, i.e. whitish. Your spot looks more like a small melanocytic nevus or a scleral thinning area that is completely harmless and sine materia from a pathological point of view, with a bluish part underneath. (Available on http://www.medicitalia.it/consulti/Oculistica/65819/Macchianella-sclera; English translation by Mario Monteleone).

```
CATALOGA - Rel. 4.8 del 9 mar 2010 - 11:00
Global number of knowledge domains in the database: 180
************************************************************
Total Number of lines in the input text:        1
Total Number of words in the input text:        154
Total Number of chars in the input text:        972
Longest line in the input text:             972
Average sentence length (in words):         9.6
Average word length (in syllables):         2.2
Flesh index for this paper:             62.0
************************************************************
Generic Dictionary Occurrences:             1
Thematic dictionaries occurrences:          14
Therefore, the input text is thematic.

************************* ANALYSIS *************************
The input Text deals with (in frequency order): MED (Medicine), ANAT (Anatomy).
******************ORDERED FREQUENCIES ********************
MED (MEDICINE)              12  92.9%
ANAT (ANATOMY)              1   7.1%
DIGE (GENERIC DICTIONARY)   1   7.1%
************************************************************
File name: Medicina.txt
Number of different compound words: 12
************************************************************************************************************************
COMPOUNDS                   OCC.    MORPH   INFL    DOM     ENG                 MORPH       INFL
************************************************************************************************************************
assottigliamento sclerale   1   N+NA    ms-+MED     scleral thinning    N+AN        s+
carenze vitaminiche         1   N+NA    fs-+ MED    vitamin deficiencies    N+NN        p+
cecità notturna     1   N+NA    fs--    MED     night blindness     N+NN        s+
macchia di Bitot            1   N+NPN   fs-+ MED    Bitot's spot        N+NPN       s+
macchie biancastre della sclera 1   N+NAPA  fp-+ MED   white spots of the sclera   N+ANPDETN   p+
macchie di Bitot            1   N+NPN   fp-+ MED    Bitot's spots       N+NPN       p+
malassorbimento lipidico    1   N+NA    ms-+    MED     fat malabsorption       N+NN        s+
nevo nevocellulare piano    1   N+NAA   ms-+    MED     small melanocytic nevus N+AAN       s+
punto di vista      1   N+NPN   ms-+    DIGE    point of view       N+NPN       s+
secchezza della congiuntiva 1   N+NPN   fs-+ MED    dryness of the conjunctiva  N+NPDETN   s+
sviluppo scheletrico        1   N+NA    ms-+ANAT    sketelal development    N+AN        s+
vitamina A          4   N+NN    fs--    MED     Vitamin A           N+NN        s-
```

Table 2 - CATALOGA Analysis Results

This table shows that CATALOGA has inferred that the input text deals with medicine as a result of the analysis and computation of the terminological compound words in it, i.e., it has made the same conclusions as the human reader.

Similar precise results have been obtained with the analysis of a corpus consisting of 1,070 text files (approx. 10 MB) extracted from Italian online newspapers. The results of these analyses are shown in the following table:

Analysis Results (1,070 files tested)		
Correct	Partially faulty	Faulty
71%	29%	0%

Table 3 - CATALOGA Analysis Results on a corpus of 1, 070 files (approx. 10MB)

It is worth noting that partially faulty results depend on terminological entries of the dictionaries that have not yet been updated and that CATALOGA also achieves detailed and successful analyses with very short text files.

We have shown that CATALOGA analytic routine can extract terminology, and more generally, semantic information from given texts in a precise way. In addition, these results confirm that:

- the information given in any terminological text is mainly conveyed by terminological compound words;
- the automatic retrieval of terminological compound words from texts allows automatic retrieval of its general meaning in the form of lexical ontologies made up of one compound word matched with one non-ambiguous knowledge domain tag.

Therefore, some more specific considerations can be made based on the results obtained so far, i.e. CATALOGA:

- achieves more efficient and less "noisy" automatic terminological information retrieval compared with statistical approaches to IR;
- may allow automatic information-based data storage and bypass human reading;
- can support the creation of information data bases in which items (i.e. digitised texts) may be linked/grouped according to the terminological compound keywords and/or the non-ambiguous knowledge domain tags they share;

- can offer more rapid management and updating procedures as far as textual terminological information is concerned;
- allows the automatic creation of bilingual lists of terminological compound words which can be used for web crawling purposes or as training set for MT and TM applications.

3.2 Use of CATALOGA with a crawling tool

The bilingual list of terminological compound words produced by CATALOGA can be used to automatically produce a precise and specific list of 'seed terms', both in the source and the target language, tailored on the source text, to be used in queries on the web with a crawling tool. For our experiment we used the BootCat toolkit (BARONI *et al*., 2004), a well-known suite of Perl scripts for bootstrapping specialized language corpora from the web.

Taking as input the key terms extracted by means of the automatic text analysis procedure performed by CATALOGA, BootCat draws upon web data to automatically build a specialised corpus for the domain of interest and tailored on the specific text to be translated. In this way, the most relevant web pages which specifically refer to the subject matter of the text to be translated can be collected.

For instance, if we take the list of English terminological compound words (refer to Table 2) produced during the text analysis phase illustrated in the previous section and we use it as 'seed terms' in Bootcat, we obtain the following list of web sites:

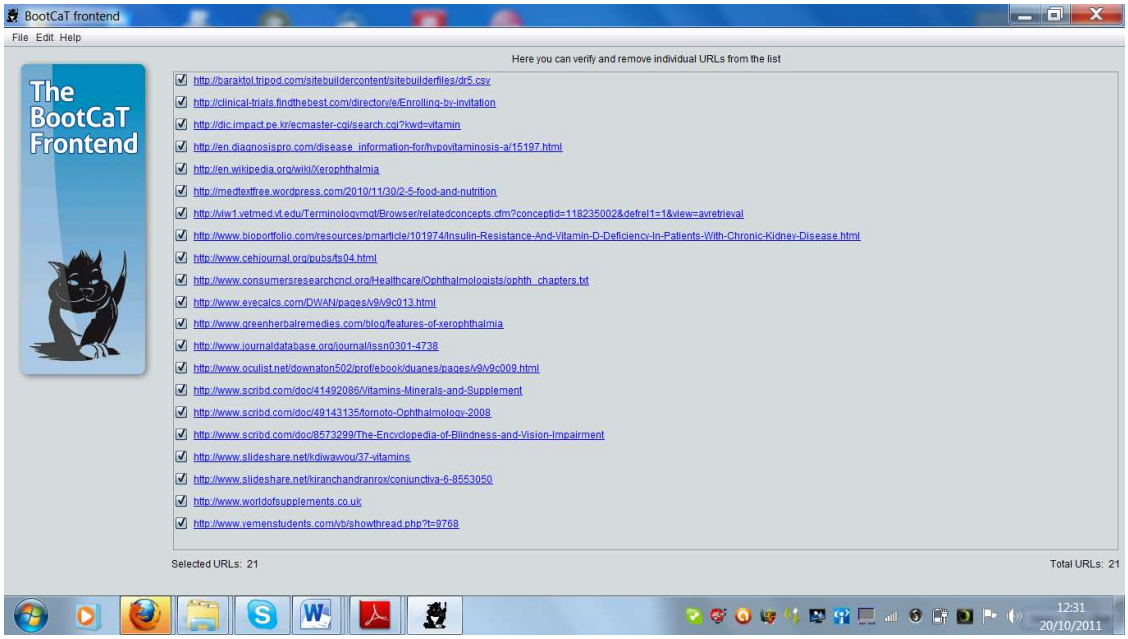


Figure 1 - URL list generated on the basis of the CATALOGA list of compound words

The list of web sites contains relevant information sources such as medical texts, glossaries, thesauri and text corpora related to the subject matter of the analysed text.

3.3 Use of CATALOGA with MT

The handling of multi-word units in MT is a well-known problem. The importance of a correct processing of multi-word units in Machine Translation (MT) and Computer Aided Translation (CAT) has been highlighted by several who have pointed out that the issue of MWU identification and accurate translation has remained an unsolved problem for current MT systems.

One of the proposed solutions for overcoming translation problems in MT and in SMT in particular is based on the idea that multi-word units should be identified and bilingual multi-word units should be grouped prior to statistical alignment. More recently, ZHIXIANG REN *et al*. (2009) have underlined that experiments show that the integration of bilingual domain MWUs in SMT could significantly improve translation performance.

Since terminological compound words are multi-word units with limited or no variability of co-occurrence among words, they have to be considered as a single lexical unit with specific semantic and syntactic features and therefore treated as one single token rather than a sum of different tokens. In this way, translation problems can be substantially solved since terminological compound words are mono-referential and unambiguous translations can be assigned prior to an MT process.

In the wake of WU *et al*. (2008), who proposed a method for constructing a phrase table using a manually-made translation dictionary in order to improve SMT performance, especially when translating domain texts, we intend to experiment on a large scale how the results of text analysis performed by CATALOGA can be used in a

translation process based on SMT and phrase-based MT in particular. The bilingual list of compound words provided by this text mining tool represents the specific dictionary of the analysed text and can therefore be used for domain adaptation purposes, adding it as a bilingual phrase table to an SMT system, which seems to outperform the use of dictionaries as a training corpus (WU *et al.*, 2008), to improve the quality of a baseline SMT system.

3.4 Use of CATALOGA to pre-translate source text

In addition to the above mentioned options for integrating CATALOGA in a translation environment, a further possibility, already available, is to use the list of compound words generated during the analysis of the source text performed by CATALOGA to pre-translate the source text, thereby ensuring coherent use of terminology throughout the whole target text.

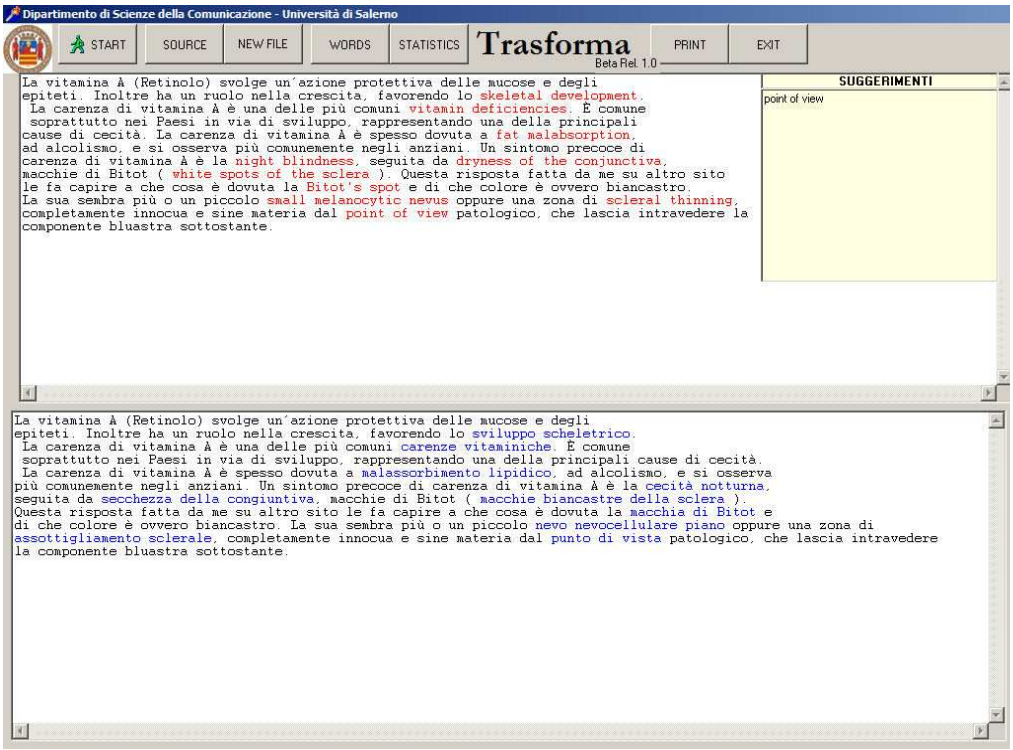


Figure 2 - Pre-translation of the source text using the CATALOGA bilingual compound word list

4. FUTURE RESEARCH PERSPECTIVES

CATALOGA and the linguistic resources used by it are based on the adoption of a well-founded, empirically coherent and solid linguistic formalisation method, i.e. the Lexicon-Grammar approach. Future research perspectives concern both the text mining tool and the DELAF/DELACF dictionaries and will include:

1. the creation of terminological electronic dictionaries for newly created knowledge domains such as e-government, bioethics, biomedicine, and so on;
2. the implementation of multilingual semantic-based terminological analysis;
3. the construction of an ontology-based query interface for information retrieval;
4. the testing and validation of the lingware on large corpora;
5. the integration of CATALOGA in web sites and portals in order to let internet surfers test the whole system.
6. the creation of bilingual and multilingual electronic dictionaries in order to use CATALOGA in automatic and semi-automatic machine translation routines as far as terminology is concerned;
7. the creation of an automatic smart text storing system, by means of which files can be automatically read and categorized on the basis of the main knowledge domain(s) they belong to and the terminological compound words they include. Information retrieval from textual relational data bases of this type will be achieved by means of queries structured both on knowledge domain tags and on terminological compound words.

These steps will undoubtedly be a launching pad towards new forms of experimentation of the system in translation environments.

REFERENCES

BARONI Marco; BERNARDINI Silvia, (2004), Boot-CaT: Bootstrapping corpora and terms from the web, *Proceedings of LREC 2004*, Lisbon, ELDA, p. 1313-1316.

D'AGOSTINO Emilio; ELIA Annibale, (1998), Il significato delle frasi: un continuum dalle frasi semplici alle forme polirematiche, in: ALBANO LEONI Federico; GAMBARARA Daniele; GENSINI Stefano; LO PIPARO Franco; SIMONE Raffaele, *Ai limiti del linguaggio*, Bari, Roma, p. 287-310.

ELIA Annibale; POSTIGLIONE Alberto; MONTELEONE Mario; MONTI Johanna; GUGLIELMO Daniela, (2011), CATALOGA: a Software for Semantic and Terminological Information Retrieval, *WIMS '11 Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, Sogndal, Norway, 25-27 May 2011.

ELIA Annibale; DE BUERIIS Giustino (eds.), (2008), *Lessici elettronici e descrizioni lessicali, sintattiche morfologiche ed ortografiche, Risultati del Progetto PRIN 2005 Atlanti Tematici Informatici – ALTI*, Collana "Lessici & Combinatorie", n. 2, Dipartimento di Scienze della Comunicazione dell'Università degli Studi di Salerno, Salerno, Plectica.

GROSS Maurice; SENELLART Jean, (1998), Nouvelles bases statistiques pour les mots du français, in: *4`emes Journées internationales d'Analyse statistique des Donnees Textuelles (JADT'98)*, Nice, p. 335–349.

MONTI Johanna, (2010), Alla ricerca della conoscenza. quali strumenti per la traduzione saggistica?, in: MONTELLA Clara (ed.), *Tradurre Saggistica*, Milano, Franco Angeli, p. 143-161.

MONTI Johanna, (2010), La E-translation da Google a Second Life: le più recenti applicazioni di Traduzione Automatica online, *Atti del XLIII Congresso della Società Linguistica Italiana*, 24-26 settembre 2009, Roma, Bulzoni Editore.

OLVERA LOBO Maria Dolores; ROBINSON Bryan; CASTRO PRIETO Rosa Maria; QUERO GERVILLA Enrique; MUÑOZ MARTÍN Ricardo; MUÑOZ RAYA Eva; MURILLO MELERO Miguel; SENSO RUIZ Jose Antonio; VARGAS QUESADA Benjamín; DÍEZ LERMA Jose Luís, (2007), A professional approach to Translator training (PATT), *Meta*, 52 (3): 518.

PYM Anthony, (2003), Redefining Translation Competence in an Electronic Age. In Defense of a Minimalist Approach, *Meta*, 48 (4): 481-497.

REN Zhixiang; YAJUAN Lü; CAO Jie; LIU Qun; HUANG Yun, (2009), Improving statistical machine translation using domain bilingual multiword expressions, *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, Singapore, August, p. 47-54.

SAG Ivan A; BALDWIN Timothy; BOND Francis; COPESTAKE Ann; FLICKINGER Dan, (2002), Multiword Expressions: A Pain in the Neck for NLP?, *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CI-CLING 2002)*, Mexico City, Mexico, p. 1-15 available on http://lingo.stanford.edu/pubs/WP-2001-03.pdf

WILLS Wolfram, (1977), Übersetzungswissenschaft. Probleme und Methoden, Stuttgart, E. Klett, (trans. The Science of Translation. Problems and methods, Tübingen, Günther Narr).

WU Hua; WANG Haifeng; ZONG Chengqing, (2008), Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora, *Proceedings of Conference on Computational Linguistics (COLING)*, p. 993-1000.

Extension du dictionnaire électronique grec de termes boursiers à partir d'un corpus spécialisé

Evangelia Fista, Tita Kyriacopoulou, Eleni Tziafa

Résumé : Un des problèmes essentiels en traitement automatique des langues (TAL) est celui des mots non reconnus par les systèmes d'analyse automatique, quelle que soit l'approche adoptée, linguistique, statistique ou hybride. Dans ce travail, nous définissons comme mots inconnus les mots non reconnus dans un corpus donné, précisément dans le corpus boursier grec car ils ne sont pas répertoriés dans les dictionnaires électroniques généraux et terminologiques du grec auxquels ont recours les systèmes de TAL. Dans un domaine de spécialité, ce problème s'avère l'un des plus délicats, suite à l'évolution rapide des langues techniques ou scientifiques. Pour l'enrichissement de ces ressources et afin d'exploiter de nouveaux domaines, il est nécessaire d'acquérir rapidement la nouvelle terminologie et de mettre à jour les ressources existantes.
Parmi les mots inconnus, figurent des néologismes, mais aussi des mots étrangers, transcrits en grec ou en alphabet latin, des mots en écriture hybride (caractères grecs et latins), des noms propres, des sigles, des mots mal orthographiés et en principe des mots non accentués. Ces mots non reconnus freinent l'analyse automatique des textes boursiers. L'objet du présent travail est l'étude de mots inconnus du corpus boursier (CoBourse), ce qui nous permettra l'ajout de termes néologiques dans le dictionnaire électronique des termes du domaine boursier. Nous nous limitons aux mots simples, les unités polylexicales demandant une approche de traitement différente. À partir de données extraites, nous proposons des heuristiques pour l'annotation semi-automatique des mots inconnus détectés à l'aide du système Unitex (Paumier, 2003), afin de les intégrer dans le dictionnaire de termes boursiers.

Abstract: The problem of unknown words (words not recognized by automated language analysis systems) is one of great importance for Natural Language Processing (NLP). In this paper, we consider as unknown words the words not recognized in a given corpus, the corpus of Greek Stock Exchange texts, since they are not included in the general dictionaries and terminologies for the Greek language, as used by the NLP systems. In this special domain, it is a critical issue, due to the rapid development of technical and scientific languages. In order to expand our resources, especially as regards new domains, it is necessary to acquire as soon as possible new terms and include them in the existing resources. Many of the unknown words are actually neologisms, and also loan words, written in Latin or Greek alphabet, words in hybrid form (both Latin and Greek alphabet), proper names, abbreviations, wrong spelled words, words without accents etc. The aim of this work is to study the unknown words found in the Stock Exchange corpus (CoBourse) and to make them part of the dictionary of the Stock Exchange terms. In this paper, we are studying simple words, as multiword expressions require a different approach.

Introduction

Dans le domaine des langues de spécialité, les progrès scientifique, technique et culturel ont pour effet la création incessante d'un nombre important de termes nouveaux qui reflètent toutes les composantes essentielles de la spécialité (DINCĂ, 2009). Selon Lerat (1993 : 132) le rôle des néologismes est d'enrichir et de moderniser le vocabulaire pour les besoins de dénomination, d'expression et de communication. Un des problèmes essentiels en traitement automatique des langues de spécialité est celui des mots inconnus, c'est-à-dire des mots qui ne sont pas répertoriés dans les dictionnaires électroniques de termes auxquels ont recours des systèmes d'analyse comme Unitex (PAUMIER, 2003).

Des recherches ont été menées sur le problème de ces mots étiquetés comme inconnus et sur la typologie des néologismes identifiés à partir des corpus statiques ou dynamiques (cf. DISTER & FAIRON, 2004, WALTHER & SAGOT, 2011, BLANCAFORT *et al.*, 2010). Parmi les mots inconnus, figurent des néologismes, mais aussi des mots étrangers, transcrits en grec ou en alphabet latin, des mots en écriture hybride (caractères grecs et latins), des noms propres, des sigles, des mots mal orthographiés et en principe des mots non accentués.

Notre approche s'inscrit dans les travaux de la linguistique de corpus, proposant une étude de mots inconnus à partir de données (corpus et dictionnaires) particulières (CARTONI, 2006, DISTER & FAIRON, 2004, MAUREL, 2004). Il est évident que, même si les néologismes *ψευτοάνοδος* (fausse hausse), *φουσκοεταιρεία* (société bulle), *ξενόχαρτο* (action étrangère), *μακροπρόβλεψη* (prévision macroéconomique)[1], sont absents de nos dictionnaires et par conséquent ne sont pas identifiés par Unitex, ils expriment des réalités bien nouvelles et communes. De plus, leur introduction dans un dictionnaire électronique nécessite pour chaque entrée une validation par un spécialiste, et cette validation n'est pas entièrement automatisable[2]. Dans la section 1 nous présentons brièvement le corpus et les ressources lexicales terminologiques existantes pour la langue grecque dans le domaine de la bourse, c'est-à-dire le corpus sur lequel notre travail repose et le dictionnaire électronique des termes boursiers.

En section 2 nous présentons les termes néologiques du domaine boursier dans le but de les intégrer dans le dictionnaire des termes de la bourse. Nous décrivons ensuite, en section 3, les mots repérés comme inconnus par Unitex ainsi que les solutions semi-automatiques de filtrage de ces mots inconnus. Puis, nous concluons en section 4 présentant quelques applications vers lesquelles nous nous engagerons.

[1] La traduction est mot-à-mot et il faut noter qu'en grec le sens de ces mots composés est transparent.

[2] Le rôle de l'informatique est certes d'automatiser certaines tâches, mais en l'occurrence, et pour des dictionnaires électroniques d'une couverture étendue, comme il en existe pour le français, des décisions d'inclusion de mots nouveaux prises de façon entièrement automatique auraient manifestement un taux d'erreur excessif (LAPORTE, 2009).

1. Ressources textuelles et lexicales pour le grec

1.1. Le corpus du domaine boursier (CoBourse)

Certes, le grec est l'une des langues de faible présence sur internet et par conséquent les textes disponibles numérisés ainsi que les ressources textuelles et lexicales existantes sont d'une couverture relativement limitée, étant donné que l'anglais s'impose de plus en plus comme la lingua franca des marchés internationaux. Quand il s'agit d'une langue de spécialité, ces ressources sont beaucoup moins étendues. Notre corpus est constitué[3] de textes spécialisés du domaine boursier tirés de sources et de registres très divers dont la publication s'échelonne sur 11 ans : de 1999 à 2010, une période marquée par deux crises majeures en Grèce, la crise boursière et la crise de la dette. Ces événements ont permis l'émergence d'un grand nombre de néologismes. Comme corpus de référence nous avons utilisé le Corpus de Textes Grecs (*Σώμα Ελληνικών Κειμένων - ΣΕΚ*, environ 30 millions de mots) (GOUTSOS, 2003) et celui formé du journal « Ta Nea » (environ 120 millions de mots)[4]. Le premier est un corpus équilibré, tandis que l'autre est un corpus de source unique. Notre corpus est relativement de grande taille puisqu'il comporte environ 19 millions de mots grecs et il se compose de quatre sous-corpus explicités ci-après. Les textes sont complets et authentiques. Ce corpus se compose de quatre sous-corpus de textes grecs.

1.2. La structure du corpus boursier

Le sous-corpus A est constitué de messages publiés dans les débats publics dans deux forums sur internet, tous deux, consacrés à la bourse. Ce genre de forum est apparu en Grèce les trois dernières années.

Le sous-corpus B provient de textes journalistiques, numérisés et couvre la période 1999-2000. Il a été complété par des articles sous format électronique de 2000 à 2010, écrits dans le même registre de langue.

Le sous-corpus C provient du site de la Bourse d'Athènes et contient des avis, des rapports annuels et des articles parus en 2000. Le sous-corpus C pourrait constituer une base pour une étude plus approfondie des textes parallèles, puisque les textes inclus sont accompagnés de leurs traductions en anglais.

Le sous-corpus D contient des textes académiques essentiellement axés sur les marchés monétaires et les marchés boursiers dérivés, fournis à partir de modules universitaires. De plus, ont été utilisés des thèses de troisième cycle et de doctorat, disponibles en ligne.

1.3. Dictionnaire boursier des termes grecs

À ce jour, le dictionnaire boursier des termes grecs simples et composés comprend 71.717 formes différentes classées sous 9.526 lemmes. Les analyses portent sur un corpus de 18.800.000 occurrences, tous textes confondus. Les termes du domaine boursier étudiés sont soit, des mots simples soit, des unités polylexicales (KYRIACOPOULOU & TZIAFA, 2011)[5].

Tous ces termes spécialisés sont codés et formalisés dans le dictionnaire électronique du grec et par conséquent des informations comme le genre (m pour masculin, f pour féminin et n pour neutre) le nombre (*s* pour singulier et *p* pour pluriel) ou le cas (*N* pour Nominatif, *G* pour Génitif, *A* pour Accusatif, *V* pour Vocatif dans les exemples ci-après) sont traitées. Voici un extrait du dictionnaire électronique grec des termes simples et composés :

splits,split.N+[Eco]:Nnp:Gnp:Anp:Vnp
spread,spread.N+[Eco]:Nns:Gns:Ans:Vns:Nnp:Gnp:Anp:Vnp
spreads,spread.N+[Eco]:Nnp:Gnp:Anp:Vnp
stock option,.N+[Eco]:Nns:Gns:Ans:Vns:Nnp:Gnp:Anp:Vnp
stock option,.N+[Eco]:Nfs:Gfs:Afs:Vfs:Nfp:Gfp:Afp:Vfp
time sharing,.N+[Eco]:Nns:Gns:Ans:Vns:Nnp:Gnp:Anp:Vnp
time-sharing,.N+[Eco]:Nns:Gns:Ans:Vns:Nnp:Gnp:Anp:Vnp
venture capital,.N+[Eco]:Nns:Gns:Ans:Vns:Nnp:Gnp:Anp:Vnp
AB μετοχή,.N+[Eco]:Nfs:Afs:Vfs
AB μετοχής,AB μετοχή.N+[Eco]:Gfs
AB μετοχές,AB μετοχή.N+[Eco]:Nfp:Afp:Vfp
AB μετοχών,AB μετοχή.N+[Eco]:Gfp
A B μετοχή,.N+[Eco]:Nfs:Afs:Vfs
A B μετοχής,A B μετοχή.N+[Eco]:Gfs
A B μετοχές,A B μετοχή.N+[Eco]:Nfp:Afp:Vfp
A B μετοχών,A B μετοχή.N+[Eco]:Gfp
A ομόλογο,.N+[Eco]:Nns:Ans:Vns
A ομολόγου,A ομόλογο.N+[Eco]:Gns[6]

[3] A l'aide de *WordSmith* (SCOTT 2011).
[4] Ce corpus nous a été fourni par Cédric Fairon, Université Catholique de Louvain.
[5] À noter que dans le dictionnaire des termes boursiers que nous avons constitué, les unités polylexicales représentent 80% des entrées.
[6] Où N indique le nom, A indique l'adjectif et [Eco] désigne le trait sémantique *économie*.

2. Formes des termes néologiques

L'analyse automatique du corpus et malgré l'utilisation du dictionnaire DELAF[7] et du dictionnaire boursier (71.717 formes fléchies) s'est heurtée à plusieurs problèmes relevant pour la plupart des spécificités de la langue boursière, comme des emprunts à l'anglais, des symboles, des abréviations, des noms propres, des néologismes morphologiques de création récente[8].

2.1. Emprunts à l'anglais

Selon Anastasiadi-Symeonidi, (1986: 61) « un vocabulaire de spécialité d'une autre langue peut être ''source de néologismes dans un vocabulaire de spécialité''. La création néologique, soit sous forme d'emprunts soit sous forme de calques est étroitement liée à l'internationalisation de la science et de la technologie ».

Dans le marché boursier grec, dominé par le marché boursier anglo-américain, on rencontre des néonymes[9] comme *split*, mais également leurs dérivés suffixaux ou flexionnels du grec comme *split* < *σπλιτάρω ή splitάρω* (faire un split), *short* < *shortarω*, *shortάρω*, *sortάρω* et/ou *σορτάρω* (être vendeur), *σορτάρισμα*, *sortαρισμα*, *σορτάκιας*[10], *hedge* < *hedgarω* et/ou *χετζάρω* (effectuer une opération de couverture), *long* < *longarω*, *λογκάρω* et/ou *λονγκάρω* (être acheteur). Les exemples cités ci-dessus montrent que de nombreux néologismes apparaissent sous plusieurs formes graphiques.

2.2. Termes néologiques de création récente

La création lexicale a pour moteur le besoin de désigner des objets nouveaux, des phénomènes nouveaux, des idées nouvelles, des concepts nouveaux dans un monde en évolution constante. De nouveaux termes, recensés dans le corpus boursier (CoBourse), illustrent la dynamique néologique toujours en relation étroite avec les changements socio-économiques, culturels qui ont marqué ce domaine jusqu'en 1999. Citons par exemple : *ελδάρχης, ελδεάρχης, ελδετζής*[11] (propriétaire d'une société de type SICAV), *δεικτοβαρής* (relatif à une obligation indexée). Il s'agit de formes qui ont une existence éphémère puisqu'elles apparaissent en fonction des évènements occasionnels liés à des métiers particuliers et par conséquent leur longévité n'est pas assurée.

2.3. Termes dérivés par suffixation

Nous citons comme particulièrement productifs les suffixes *-ότητα, -ποίηση, -λογία, -σμα, -ισμός, -τράπεζα, -χαρτο: διακυμανσιμότητα* (volatilité ou instabilité), *αποπαγιοποίηση* (déconsolidation), *βιοχαλκολογία* (discussion autour de la société Viohalco), *λογκάρισμα* (action d'achat), *μαρφινοτράπεζα* (Marfin Banque).

Le suffixe nominal *-ισμα* entre dans la formation de nombreux substantifs neutres qui expriment une action ou le résultat de cette action: *λογκάρισμα* (action d'être acheteur), *σορτάρισμα* (action d'être vendeur), *μανατζάρισμα* (manœuvre de bourse). Les suffixes *-ικός, -ιμος* sont productifs dans la formation des adjectifs à base de noms ou de verbes: *αξιογραφικός* (relatif aux titres), *απορρυθμιστικός* (qui peut provoquer une déréglementation), *αποτιμήσιμος* (quantifiable).

Nous avons remarqué que, par exemple, l'acronyme *ΕΛΔΕ* (SICAV) a servi de base pour la formation du mot *ελδάρχης* (ANASTASIADI-SYMEONIDI, 1986 : 54 et 241). On note plusieurs termes, dérivés ou composés, issus de noms propres tels que *βγενό**χαρτο**, βαγγελό**χαρτο**, πανουσό**χαρτο**, παπαελληνό**χαρτο*** (< Βγενόπουλος/Vgenopoulos, Βαγγέλης/Vangelis, Πανούσης/Panousis, Παπαέλληνας/Papaellinas). Le formant ***-χαρτο*** (*χαρτί* = papier, *l'action* au jargon boursier) est très productif et il a servi pour l'identification des néologismes (cf. Figure 2).

2.4. Termes dérivés par préfixation

Après avoir examiné la liste de mots candidats (3.836 mots), nous avons eu à résoudre un autre problème : un grand nombre de mots n'a pas été reconnu parce que ces mots étaient préfixés. Les préfixes formant les termes nouveaux sont, pour la plupart, d'origine grecque (préfixes savants). On a recensé un grand nombre de préfixes dont les plus productifs[12] sont :

> **υπερ-** (hyper) : *υπερτράπεζα* (banque énorme), *υπερσυγκέντρωση* (énorme concentration), *υπερχρέη* (dettes énormes)

---

[7] Dictionnaire électronique des formes fléchies du vocabulaire général du grec moderne, élaboré suivant la méthodologie du LADL (COURTOIS & SILBERZTEIN, 1990).

[8] Nous n'étudions pas les néologismes sémantiques, comme par exemple le nom *φούσκα* (bulle).

[9] Selon Kocurek (1991 :174) le terme de « néonyme » est utilisé par Cellard et Sommaelt (1979) mais Rondeau (1984) réserve à la néologie terminologique la dénomination de« néonymie ».

[10] Le terme *σορτάκιας* est formé sur le terme anglais *short selling* et désigne la personne qui vend des actions qui ne possède pas mais qui espère "qu'elles baissent pour tirer un bénéfice important (MATHIOPOULOS, 1999: 164).

[11] Les suffixes *-τζής, -άρχης* forment des noms de métier dans un registre de langue populaire.

[12] A titre indicatif, nous avons relevé 104 formes préfixées en *υπερ-* (hyper), 60 formes en *επανα-*, 32 formes en *υπο-* (hypo), 24 formes en *ψιλο-* (psilo), 22 formes en *μικρο-* (micro), 15 formes en *νεο-* (neo) et 13 formes en *αυτο-* (auto).

επανα-[επι-+ανα] (epana) : *επαναπώληση* (revente), *επαναδημοσιοποίηση* (nouvelle communication)
υπο- (hypo) : *υποαγορά* (la marchée peu efficace), *υποαντίδραση* (réaction hypotonique), *υποαπόδοση* (baisse de revenu)
ενδο- (endo) : *ενδοσυνεδριακά* (pendant la séance boursière), *ενδοδίκτυο (intranet)*, *ενδοημερήσια* (pendant la journée)
εξω- (exo/extra) : *εξωεταιρικός* (en dehors d'une société), *εξωχρηματιστηριακός* (transaction hors la Bourse)
ιδιο- (idio) : *ιδιοχρηματοδοτούμενος* (autofinancé), *ιδιοχρηματοδότηση* (autofinancement), *ιδιοχρησιμοποιώ* (auto-utiliser)
αυτο- (auto) : *αυτοπαλινδρόμηση* (auto-régression), *αυτοπεραίωση* (auto-achèvement)
βραχυ- (brachy) : *βραχυδιακύμανση*, (fluctuation à court terme), *βραχυμεσοπρόθεσμος* (court et moyen terme)
ημι- (hemi) *: ημιδιακύμανση* (semi fluctuation)
μικρο- (micro) : *μικροάνοδος* (petite hausse) *μικροπρόβλεψη* (mini-prévision),, *μικροτράπεζα* (petite banque), *μικροϊδιώτης* (investisseur de détail)
μακρο- (macro) : *μακροπρόβλεψη* (prévision à long terme), *μακροσταθερότητα* (stabilité à long terme)
μεγαλο (megalo) : *μεγαλοκαρχαρίας* (requin de la finance), *μεγαλοκαταθέτης* (gros déposant), *μεγαλομετοχοαπατεώνας* (gros-actionnaire-escroc)[13]
μεσο- (meso) : *μεσομακροπρόθεσμος*, (ayant un impact à moyen et long terme) *μεσομακροχρόνιος* (de moyen et longue durée)
νεο- (neo): *νεοεισερχόμενος* (nouvel entrant), *νεοεισαγόμενος* (nouvellement introduit), *νεοεπενδυτής* (nouveau sponsor), *νεοϊδρυόμενος* (nouvellement créé)
πολυ- (poly) : *πολυδιασπορά* (grande dispersion), *πολυσυγγραμικότητα* (multicolinéarité)
πρωτο (proto) : πρωτοεισάγω (introduire pour la première fois), *πρωτοεισαχθείς* (introduit pour la première fois), *πρωτοσορτάρω* (être vendeur pour la première fois)
ψευτο και ψευδο- (pseudo-): *ψευδομεταβλητή* (pseudo-variable), *ψευδοείδηση* (fausse nouvelle), *ψευτοάνοδος* (fausse hausse)
ψιλο- (psilo): *ψιλοκλειδώνω* (en train de clôturer), *ψιλοαρκούδα* (petit ours), *ψιλοσυγχωνευσούλα* (une toute petite fusion)

Nous allons présenter par la suite les traitements effectués afin de reconnaitre les termes de notre corpus qui n'ont pas été identifiés lors de l'analyse automatique.

3. Mots inconnus dans le CoBourse et requêtes de filtrage

3. 1. Mots inconnus

Le prétraitement de notre corpus a été effectué par le système Unitex. La première étape a été la segmentation automatique du CoBourse de 19.00.000 mots. Nous avons ainsi obtenu trois fichiers comportant : a. la liste des mots reconnus, b. la liste des mots non reconnus et c. la liste des mots (tokens) par fréquence et par ordre alphabétique[14]. Parmi les 14.450.259 occurrences, Unitex a identifié 151.753 mots simples et 23.746 mots composés à partir de 34.548 occurrences. La liste des mots inconnus qui contenait 156.493 mots différents comportait des coquilles, des fautes d'orthographe, des noms propres, des abréviations, des mots étrangers et des termes néologiques. Plus précisément, nous avons constaté que:

- une lettre ou plus manquent ou une lettre est en trop : *καταναωτικών (<καταναλωτικών)* (de consommation), et *κατανάλλωση (<κατανάλωση)* (consommation)
- deux lettres sont interverties ou une lettre remplace une autre : *κατα**κρ**ύλα (<κατρακύλα)* (forte baisse), *ελέ**ν**χους (<ελέγχους)* (contrôles)
- des mots sans espace existent : *ενγένει (< εν γένει)* (en général), *ενλόγωαγωγή (< εν λόγω αγωγή)* (traitement en question)
- une syllabe ou une lettre sont répétées : *κοκοκοκοκκινοοοοο* (rouge), *λεφτααααααάά* (argent)
- des mots ne sont pas accentués ou sont mal accentués: *απομειωση (< απομε**ί**ωση)* (réduction), *ανά**στρόφό** (< ανάστροφο)* (inverse)
- des mots étrangers existent : (souvent en caractères grecs ou en écriture hybride - caractères grecs et latins) : *split, haircut* (marge de sécurité), *swap, arbitrage, σπλιτ* (<split), *ντίλερ* (<dealer), ***sort**άρισμα* (action de vendre)

Il y avait également :

[13] Traduction littérale.
[14] Il faut noter que le grand nombre de mots inconnus est dû au sous-corpus A qui, rappelons-le, est constitué de messages publiés dans des débats publics dans deux forums sur internet, consacrés à la bourse. Ce genre de textes présente des similitudes avec le discours oral : de nombreux mots grecs transcrits en alphabet latin (*greeklish*), des abréviations, des coquilles et des fautes d'orthographe.

- des noms propres comme *Marfin*, *Vernikos*, des abréviations (sigles, acronymes) : *cds (*< crédit default swaps), *ETFs (<Exchange Trade Funds)*
- des mots dérivés par suffixation : *τιμ**ούλα*** (petit prix), *χαρτ**άκι*** (petite action), *λεφτουδ**άκια*** (du fric), *αναλυτ**αράς*** (grand analyste d'affaires)
- des mots construits à l'aide de préfixes soudés : ***υπερ**απορρόφηση* (absorption excessive), ***ψευτο**άνοδος* (fausse hausse)

Parmi ces mots inconnus figurent, comme nous avons déjà mentionné, des néologismes qui pourraient être des termes néologiques candidats. C'est pourquoi, nous avons procédé au filtrage du bruit pour diminuer le nombre de mots inconnus et faciliter l'identification de « vrais » termes néologiques.

3.2. Requêtes de filtrage

A partir du fichier des mots inconnus qui contenait 156.493 mots simples, nous avons implémenté des applications appropriées afin de traiter de manière différente les mots inconnus de diverses origines et repérer les termes néologiques attestés dans le corpus.

Il est connu que les noms propres sont fortement représentés dans les textes. Spriet et al. (1996) a démontré que, parmi un pourcentage de 6% d'erreurs relevés dans un corpus donné, 58% des mots inconnus étaient des noms propres. Afin de reconnaitre les noms propres, nous avons appliqué au corpus le graphe de reconnaissance automatique des noms propres du grec, élaboré par Mavropoulos (2012).

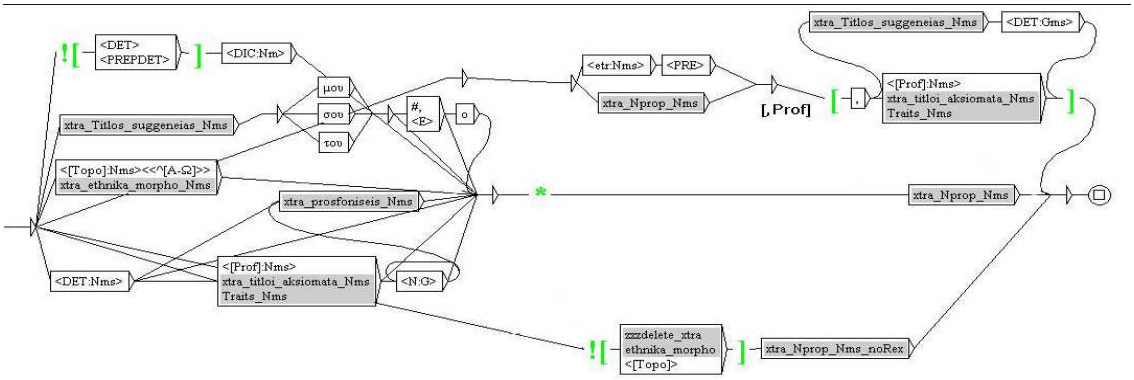


Figure 1 : Exemple d'un graphe de reconnaissance de noms propres (MAVROPOULOS, 2012)

De la même façon, nous avons extrait les abréviations du corpus, créant un graphe qui reconnait automatiquement des sigles et des acronymes des mots simples commençant par au moins une lettre majuscule. Ainsi, nous avons reconnu environ 70.000 noms propres et abréviations, ce qui représente 45% de mots inconnus.

Par la suite, nous reconnaissions les mots non accentués, ce qui représentait 16% de mots inconnus (24.675 formes). Pour le faire, nous avons inclut dans le dictionnaire du grec moderne (DELAF) les formes de tous les mots non accentués, augmentant ainsi sa taille de 2.029.797 à 3.858.936 formes fléchies[15].

Un grand nombre des mots restants étaient des mots écrits en alphabet latin, c'est-à-dire des emprunts (*buy*, *hedging*), des mots en greeklish (*metoxi*/action, *agora*/marché) ou en graphie hybride (*shortάρισμα*)[16]. 62% des termes anglais se réfèrent essentiellement aux outils économiques se terminant en *–ing* comme par exemple *hedging, rehedging, bookbuilding, churning, decoupling, scalping, trading.* Pour distinguer les mots anglais des autres mots écrits en alphabet latin nous avons utilisé le dictionnaire anglais DELAF[17]. Ensuite, nous avons supprimé les mots écrits en greeklish grâce à un script perl.

Il faut noter aussi que des noms et des adjectifs anglais ont servi de base pour la contruction de verbes grecs en *–άρω* comme en attestent les exemples : *short > shortarω*, *option> optionαρω*.

On obtient ainsi une liste de noms, verbes et adjectifs. Pour leur reconnaissance, nous avons créé deux graphes, un pour les verbes et l'autre pour les noms et adjectifs. Le graphe des verbes décrit une liste de 5.867 terminaisons caractéristiques, basée sur le fichier contenant la description des classes morphologiques, utilisé par le programme GFF[18].

[15] Pourtant il n'y a pas dédoublement des entrées car, il y avait déjà des mots monosyllabiques sans accent (conformément aux règles d'accentuation du grec moderne) grâce à un script perl.

[16] Ces mots sont détectés par le système comme des erreurs, mais comme ils appartiennent au langage particulier utilisé, au langage de messages aux blogs, sms etc, ils méritent une étude plus approfondie.

[17] Nous avons regroupé les emprunts reconnus dans un fichier à part pour une étude ultérieure. Nous constatons aussi la présence d´un grand nombre de mots anglais relevant du vocabulaire de la langue générale.

[18] Programme de Génération de Formes Fléchies.

A ce stade, dans la liste des mots non reconnus restaient des noms mais également des adjectifs et des participes passés[19]. Nous avons identifié les participes passés par l'application d'un graphe qui contenait les terminaisons caractéristiques des participes passés du grec moderne, mais aussi celles de la langue savante[20] ce qui nous a amené à en extraire 2.287 participes passés.

Les noms et les adjectifs ont été reconnus par un graphe (Figure 2) qui décrit les suffixes les plus productifs de la formation des termes nominaux : *-ότητα, -ποίηση, -λογία, -σμα, -ισμός*, etc.

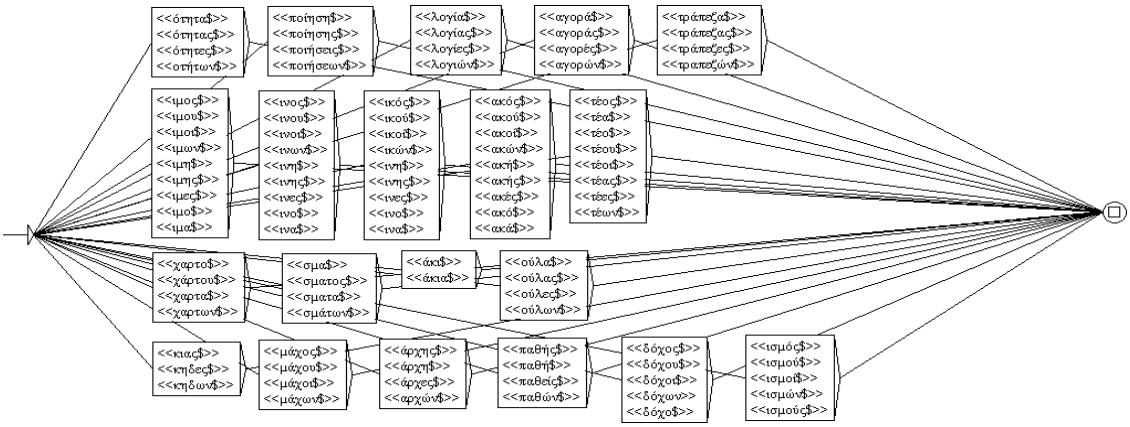


Figure 2 : Graphe de suffixes nominaux pour l'extraction des termes candidats

Comme résultat final nous avons obtenu deux listes de mots, la première contenant 2.158 formes verbales et la deuxième 1.678 formes nominales. Ces listes sont beaucoup moins volumineuses que les listes initiales des mots inconnus. Cela s'explique d'une part par la taille relativement limité[21] du corpus boursier et d'autre part par le fait que les mots appartenant à la langue générale ont été exclus. Après avoir fait une validation manuelle nous avons identifié au total 1.087 termes néologiques (413 formes verbales et 674 formes nominales).

4. Conclusions et perspectives

L'objectif de notre travail était d'identifier les termes spécifiques au corpus utilisé et qui ne sont pas présents dans les dictionnaires électroniques généraux et terminologiques du système Unitex. Nous avons développé diverses stratégies afin de repérer les termes néologiques dans le but d'enrichir le dictionnaire de termes boursiers. Nous avons extrait 1.087 termes néologiques parmi les mots inconnus (156.493 mots), en nous appuyant particulièrement sur leurs suffixes. Les procédures néologiques les plus significatives de notre corpus étaient l'emprunt et la dérivation par suffixation et préfixation.

Notre recherche démontre que le vocabulaire spécialisé de la bourse en grec moderne contient non seulement des noms et des adjectifs (674 formes de termes néologiques attestées) mais aussi un nombre important des verbes (413 formes verbales attestées). Ces informations seront intégrées à un outil d'annotation et de lemmatisation[22] qui est en cours de développement. Cet outil permettra la liaison et l'interopérabilité des dictionnaires DELAF grecs avec d'autres systèmes comme Antconc (ANTHONY, 2011) et Wordsmith Tools (SCOTT, 2011).

Bibliographie

ANASTASIADI-SYMEONIDI Anna (1986), *Η Νεολογία στην Κοινή Νεοελληνική. Epistimoniki Epetirida Filosofikis Scholis.* Thessaloniki: Aristotle University of Thessaloniki.

ANTHONY Laurence (2011), *AntConc* (Version 3.2.2) [Computer Software], Tokyo, Japan: Waseda University, http://www.antlab.sci.waseda.ac.jp.

BARONI Marco ; BERNARDINI Silvia ; FERRARESI Adriano ; ZANCHETTA Eros (2009), The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora, in : *Language Resources and Evaluation* 43(3), p. 209-226.

BLANCAFORT Helena ; RECOURCÉ Gaëlle ; COUTO Javier ; SAGOT Benoît ; STERN Rosa ; TEYSSOU Denis (2010), Traitement des inconnus : une approche systématique de l'incomplétude lexicale, in : *TALN 2010*, Montréal, Canada.

CARTONI Bruno (2006), Constance et variabilité de l'incomplétude lexicale. In *RECITAL 2006*, Leuven, Belgium, TALN 2006.

[19] Les mots mal orthographiés n'ont pas été traités et feront l'objet d'une étude ultérieure, basée sur leur degré de similitude avec des mots existants.

[20] Beaucoup de participes passés de la langue savante sont largement utilisés dans le grec moderne.

[21] Si on le compare avec de grands corpus annoncés (FERRARESI et al. 2008, 2010, BARONI et al. 2009, POMIKALEK 2009).

[22] La lemmatisation consiste à associer un lemme à chaque mot du texte. Si le mot ne peut pas être lemmatisé (nombre, mot étranger, mot inconnu), aucune information ne lui est associée.

CELLARD Jacques ; SOMMAELT Micheline (1979), *500 mots nouveaux définis et expliqués*, Paris-Gembloux, Duculot.

DINCĂ Daniela (2009), La néologie et ses mécanismes de création lexicale, in *Analele Universității din Craiova, Seria Lingvistică*, nr. 1-2, 2009, p. 79-91.

DISTER Anne ; FAIRON Cédrick (2004), Extension des ressources lexicales grâce à un corpus dynamique, *Lexicometrica*.

FAIRON Cédrick ; COURTOIS Blandine (2000), Extension de la couverture lexicale des dictionnaires électroniques du LADL à l'aide de GlossaNet, In *Actes du Colloque JADT 2000 : 5es Journées Internationales d'Analyse Statistique des Données Textuelles*, Lausanne.

FERRARESI Adriano ; ZANCHETTA Eros ; BARONI Marco ; BERNARDINI Silvia (2008), Introducing and evaluating ukWaC, a very large web-derived corpus of English, in : EVERT Stefan, KILGARRIFF Adam & SHAROFF Serge (éd.) *Proceedings of the 4th Web as Corpus Workshop (WAC-4) – Can we beat Google?* Marrakech.

FERRARESI Adriano ; BERNARDINI Silvia ; PICCI Giovanni ; BARONI Marco (2010), Web Corpora for Bilingual Lexicography: A Pilot Study of English/French Collocation Extraction and Translation, in: XIAO Richard (éd.), *Using Corpora in Contrastive and Translation Studies*, Newcastle, Cambridge Scholars Publishing.

GOUTSOS Dionysis (2010), The Corpus of Greek Texts: A reference corpus for Modern Greek, in : *Corpora 5 (1)*, p. 29-44.

KOCOUREK Rostislav (1991), La langue française de la technique et de la science. Vers une linguistique de la langue savante. Wiesbaden: Brandsletter.

KYRIACOPOULOU Tita ; TZIAFA Eleni (2011), Dictionnaires électroniques et terminologie: le cas du vocabulaire "boursier", *9èmes Journées Scientifiques du réseau Lexicologie, Terminologie, Traduction*, 15-16 Septembre 2011, Université Paris 13.

LAPORTE Éric (2009), Concordanciers et flexion automatique, in *Cahiers de Lexicologie*, 94 (1), p. 91-106.

LERAT Pierre (1993), *Les langues spécialisées*, Paris, PUF.

MATHIEU Yvette Yannick ; GROSS Gaston ; FOUQUERÉ Christophe (1998), Vers une extraction automatique des néologismes, in : *Cahiers de Lexicologie*, n° 72, p. 199-208.

MATHIOPOULOS Haris (1999), *Μικρό Εγχειρίδιο του Επενδυτή*, Athens, Estia.

MAUREL Denis (2004), Les mots inconnus sont-ils des noms propres?, in : *Actes des JADT* 2004.

MAVROPOULOS Athanasios (2012), Ένα σύστημα αυτόματης ανάλυσης κειμένων της Νέας Ελληνικής. Μέθοδοι αναπαράστασης των κύριων ονομάτων προσώπων, Thessaloniki, Aristotle University of Thessaloniki.

PAUMIER Sébastien (2003), *Unitex. Manuel d'utilisation*, Paris, Université Paris-Est Marne-la-Vallée, http://igm.univ-mlv.fr/~unitex/UnitexManual.pdf.

POMIKÁLEK Jan ; RYCHLÝ Pavel ; KILGARRIFF Adam (2009), Scaling to Billion-plus Word Corpora. Advances in Computational Linguistics, in : *Special Issue of Research in Computing Science Vol 41*, http://pics.cicling.org/2009/RCS-41/003-014.pdf.

RONDEAU Guy (1984), *Introduction à la terminologie*, Québec, Gaetan Morin.

SCOTT Mike (2011), *WordSmith Tools version 6*, Liverpool: Lexical Analysis Software.

SPRIET Thierry, BÉCHET Frédéric, EL-BÈZE Marc, DE LOUPY Claude & KHOURI Liliane (1996), Traitement automatique des mots inconnus; in : *Proceedings of TALN'96*, Marseille, p. 170-179.

WALTHER Géraldine ; SAGOT Benoît (2011), Problèmes d'intégration morphologique d'emprunts d'origine anglaise en français, in : *Proceedings of the 30th Lexis and Grammar Conference*, Nicosia, Cyprus.

A Local Grammar of verb predicates denoting Emotion in Greek: a Lexicon-Grammar approach

Voula Giouli, Aggeliki Fotopoulou

1. Introduction

This paper describes work aimed at a formal description of the Greek verb predicates denoting emotion. The focus is placed on verbs entering in constructions with *Subject* and *Object* arguments of the type $N_0VN_{1=hum}$ (SVO) *with* N_1 being the animate argument/object that assumes the role of the agent *Experiencer* of the emotion denoted by the verb and caused by the *Subject* N_0.

2. The Methodological Framework

The theoretical framework adopted within this study is that of Lexicon-Grammar (Gross, 1975). Being a model of syntax limited to the elementary sentences of a natural language, the theory argues that the unit of meaning is not located at the level of the word, but at the level of elementary sentences of the form *Subject – Verb – Object*. A set of distributional properties associated with words, i.e., types of prepositions, features attached to nouns in subject and object positions, etc. is also taken into account, resulting to a more fine-grained classification, and to the creation of homogenous word classes. Finally, transformation rules, construed as equivalence relations between sentences, further generate equivalent structures. All this information (argument structure, distributional properties and permitted transformational rules) is formally encoded in the so-called Lexicon-Grammar (LG) tables.

3. Previous work

Verbs denoting emotion (also referred to in the literature as « verbs of psychological state » (Levin, 1993) and more commonly known as *psych-verbs)* have been extensively treated in syntactic theory. Early studies (Lakoff, 1970) (Postal, 1971) have tried to give transformational accounts of the behavior of psych-verbs, whereas (Belletti et al., 1988) attempt an analysis of properties exhibited by psych-verbs in Italian on the basis of their thematic information and syntactic configurations.

Other theoretical studies have been engaged with the uniform and consistent description of the lexicon of emotions (Anscombre, 1995), (Ruwet, 1995), on the basis of their syntactic properties, lexical choices and semantic criteria.

Within the Lexicon-Grammar (LG) framework, the analysis of verb constructions that are semantically defined as denoting sentiment and their encoding in LG tables has been initiated by M. Gross (1975). French emotion predicates have also been treated by Mathieu (2000).

As far as Greek is concerned, verbal constructions (Antoniou, 1984) (Valetopoulos, 2005), noun (Gavrilidou, 2002) (Pantazara et al., 2008) (Fotopoulou et al., 2009), and adjective predicates (Valetopoulos 2005) denoting emotion have been treated within the LG framework. Building on previous efforts, this work aims to provide new evidence about verb predicates in the light of corpus data, the ultimate goal being the implementation of a complete grammar of emotions.

4. Methodological principles

Initially, a core lexicon of emotion verbs extracted from existing lexical resources and corpora was manually updated and extended; the semantic class of verb predicates denoting emotion was then defined based on a set of predefined criteria. A set of *appropriate lexical semantic tests* were employed as a formal device guiding the selection of a *core vocabulary of emotions* that covers the grammatical category of verbs, and the delineation of the semantic class of emotions (Giouli et al., 2012) leaving semantically and conceptually related verbs aside for future treatment.

The so-identified verbs were then classified on the basis of their syntactic and semantic properties. Verbs with more than one usage/meaning have been treated as separate lexical items, and their properties are assigned to each one thereof.

A set of features that are appropriate for the description of a syntactic category concerning grammatical (i.e., past participle form) or syntactic information (i.e., passive transformation, inchoative alternation, etc.) and lexical choice (i.e., preposition a verb is subcategorized for) has been applied to all verb predicates, and their linguistic validity was checked against corpus evidence. The syntactic classification of Greek verbs denoting emotion has been encoded in LG tables that describe formally from a linguistic point of view their argument structure, distributional properties and possible transformations. Each class is represented by a table that includes all lexical items of that class.

4.1. Lexicon-Grammar tables

This process resulted in a list of 339 verb predicates that were finally selected for further study guided by corpus evidence. Emotion verbs were identified to appear in *both transitive* (N_0VN_1) *and intransitive* (N_0V), ($N_0VPrepN_1$) constructions. A closer investigation of the *elementary sentences* of the form subject-verb-essential complements these verbs participate in, four major classes of Greek emotion verbs were identified and systematised in separate tables (VS1 - VS4) an overview of which is provided hereafter.

The first class (VS1) is defined by the structure $N_{0=hum}$ *V (E+PrepN1)* and encompasses verbs like *αγαλλιάζω* (=rejoice), *δυσανασχετώ* (=resent), ντρέπομαι (=be ashamed), *φρίττω* (=shudder). These verbs involve at least one human agent [+human] realized as N_0 in Subject position being the *Experiencer of the emotion.*

(1) *Οι* πολίτες/**Τα βιβλία αγανακτούν*
The citizens /*The books *resent*

(2) *Οι* πολίτες *αγανακτούν με* τα μέτρα λιτότητας
The citizens *resent (with)* the austerity measures

The second class (VS2), defined by the structure $N_{0=hum}V(E+N_1)$, encompasses verbs with a N_0 in *Subject* position that is obligatorily [+human], whereas N_1 is non-restrictive, in the sense that it can be either [+human] or [-human]. For example:

(3) Ο Λουκάς *φοβάται την Άννα/το σκοτάδι*
Luc is afraid of Anna/the darkness

*(4) **Η μουσική/*Το βιβλίο* φοβάται τον Γιάννη
**Music/*The book* is afraid of John

Classes (VS3) and (VS4) are defined by the basic structure $N_0VN_{1=hum}$, their shared property being the presence of a human noun (N_{hum}) in the N_1 (direct object) position. The object of these verbs (N_1) assumes the role of the *Experiencer*, i.e., the entity feeling the emotion denoted by the verb predicate, hence they are referred to as *Object-Experiencer* verbs.

5. Object-Experiencer verbs

Object-Experiencer verbs are systematized under two distinct classes on the basis of their syntactic, distributional and transformational properties. In this section, we will present the different properties of the verb predicates present in tables (VS3) and (VS4) (mainly distributional and transformational). In the remainder of this document we will present the transformational, syntactic and distributional properties that characterize each one of these two classes.

5.1. Transformational properties

Inchoative alternation

Class (VS3) comprises verbs that establish a homogeneous group with specific morpho-syntactic properties, as they have only [+active] morphology (*νευριάζω* (=pish), *ανησυχώ* (=worry), *ηρεμώ* (=calm), etc.) while lacking the [-active] one. These verbs occur in two syntactically distinct, yet semantically equivalent structures, allowing, thus, for the inchoative alternation. This equivalence is denoted as $\underline{N_0VN_{1=hum}} \Leftrightarrow \underline{N_{1=hum}V(E+PrepN_1)}$ and exemplified in (5) and (6) below:

(5) Ο Γιάννης/ο θόρυβος νευρίασε τη Μαρία
John / The noise *make* Maria angry
Η Μαρία *νευρίασε* (με τον Γιάννη/με το θόρυβο)
Maria *was angry* (at John / at the noise)
(6) Η Μαρία/άδικη κριτική *θύμωσε* τον Μάριο
Maria/the unfair criticism *made* Mario *angry*
Ο Μάριος θύμωσε με τη Μαρία
Mario *was angry* at Maria

Middle and Passive transformation

Class (VS4) comprises verbs that take part in either the *passive* or in the *middle* construction, which corresponds to the passive transformation with the passive verb form (denoted as V_{mp}); the object N_1 becomes the subject, while the subject N_0 appears as dependent to a phrase headed by a preposition; usually this prepositional phrase is omitted, especially in the middle construction. The definitional structure of this class is schematically

denoted as $\underline{N_0VN_{1=hum}} \Leftrightarrow : \underline{N_{1=hum}V_{mp}(E+PrepN_0)}$. Examples of this property are shown on the following sentences:

(7) Η Μαρία/οι εξετάσεις *αγχώνουν* τον Γιάννη
Maria / The exams *stress* John.
Ο Γιάννης αγχώνεται (Ε+με τις εξετάσεις)
John is stressed (E+because of the exams)

Genitive restructuring

Verbs pertaining to both classes (VS3) and (VS4) also enter the *genitive restructuring property* (Guillet el al., 1981):

(8) Η συμπεριφορά του Γιάννη θύμωσε τη Μαρία=
Ο Γιάννης θύμωσε τη Μαρία *με τη συμπεριφορά του*
(John made Maria angry with his behavior=
John's behavior made Maria angry)

(9) Το μωρό αγχώνει τη Μαρία με το κλάμα του=
Το κλάμα του μωρού αγχώνει την Μαρία
(The baby is making Maria anxious (because of its crying)=
The baby's crying is making Maria anxious)

Syntactic causative constructions

Another property shared by verb predicates that belong to both classes (VS3) and (VS4) is that they are semantically equivalent to a syntactic *causative construction*:

(10) Η συμπεριφορά *του Γιάννη θύμωσε τη Μαρία =*
Η συμπεριφορά του Γιάννη *κάνει τη Μαρία να θυμώσει*
(=John's behavior makes Maria angry)

(11) Η αδιαφορία του Γιάννη νευρίασε τη Μαρία
(John's indifference irritated Maria) =
Η αδιαφορία του Γιώργου έκανε τη Μαρία να νευριάσει
(John's indifference made Maria irritate)

5.2. Syntactic properties

One basic property verbs falling in class (VS4) exhibit lies in the fact that depending on semantic context, their argument structure may be on occasions reduced by one argument, omitting thus the Object. This is not true for verbs pertaining to class (VS3). Examples:

(12) Οι εξετάσεις αγχώνουν τον Γιάννη ($N_0VN_{1=hum}$)
Exams stress John

(13) Οι εξετάσεις αγχώνουν (N_0V)
Exams stress (=cause stress)

(14) Ο σεισμός ανησύχησε τους κατοίκους
The earthquake worried the inhabitants

(15) *Ο σεισμός ανησύχησε
*The earthquake worried

Apart from the basic arguments (Subj, Obj) emotion predicates also involve a *cause* that triggers the state denoted and experienced by the direct *Object* N_1*;* this may be realized via a variety of forms, ranging from phrases headed by prepositions (either με or για) or a clausal complement:

(16) Ο Γιάννης με άγχωσε *με τις ερωτήσεις του*
John has made me anxious *with his questions*

(17) Η Μαρία αγχώνεται *για το μέλλον της*
Maria is anxious about her future

(18) Αγχώθηκα *μην κάνω λάθος*
I was anxious *not to make some mistake*

(19) Ο Γιάννης αγχώνεται *μήπως ξεχάσει κάτι*
John is worried

(20) Οι γονείς αγχώνονται *πώς θα τα καταφέρουν*
Parents are anxious how to manage

(21) Αγχώνομαι για το τι θα γίνει
I am anxious about what will happen
(22) αγχώνομαι ότι μπορεί να χάσω τη δουλειά μου
I am worried that I might lose my job
(23) Ο Μάριος αγχώνεται να είναι καλός μαθητής
Mario is anxious to be a good student

5.3. Distributional properties

In the tables developed, the *distributional properties* of the verb predicates under study have been encoded, based on the semantics of their arguments.

The N_{hum} (human noun) property indicates that in a given argument position (like N_0, subject), verbs select a noun denoting an animate entity, the typical case being a noun phrase that can be replaced by a proper name. Collective nouns (N_{coll}) are also considered as $N_{hum,}$: τάξη (=class), κυβέρνηση (=government), etc.

The non-restrictive N_0 in subject position is either a N_{hum} (human noun), an abstract noun, a collective noun ($N_{0=coll}$), a concrete noun ($N_{concrete}$), or a clausal complement (P_{comp}) introduced with *το γεγονός ότι* (=the fact that), *το να*, *το ότι* (=the fact that):

(24) Η *Μαρία*($N_{0=hum}$) απογοήτευσε τους γονείς της
Maria let her parents down.
(25) Η *τάξη*($N_{0=coll}$) όλη στενοχώρησε τον δάσκαλο
The whole class made the teacher sad
(26) Οι *απόψεις*($N_{abstract}$) της με τρομάζουν
Her stance frightens me
(27) *Το γεγονός [ότι έπρεπε να βγάλω λόγο]*(P_{comp}) με άγχωσε
The fact that I had to make a speech made me anxious

Moreover, the distributional properties of prepositions used to denote the cause triggering the emotion denoted by the verbal predicate have been encoded in the LG tables: *με, για* and από are selected invariably by most verbs.

6. Conclusions

We have hereby presented work aimed at the formal description and classification of the Greek verbs denoting emotion. The analysis of the empirical data extracted from corpora aims at the development of a Language Resource that will be applicable for NLP applications ranging from Word sense Disambiguation to text understanding, Natural Language Generation, sentiment analysis, etc. The classification of the selected verb predicates was performed on the basis of the following axes: (i) predicate argument structure; (ii) selectional restrictions imposed over Subject and Object complements of the predicates; and (iii) transformation rules.

Future work concentrates on extending the set of syntactic and semantic properties of the verbs, and to integrate other grammatical categories (i.e., nouns, adjectives, adverbs) and multi-word expressions.

References

ANSCOMBRE, Jean-Claude (1996), Noms de Sentiment, Noms d'Attitude et Noms d'Abstraits. Les noms abstraits, Actes du Colloque de Dunkerque. Lille: Presses universitaires du Septentrion.

ANTONIOU Jeanne **(**1984), Syntaxe et métaphore des verbes psychologiques en grec. Doctorat de 3e cycle, Université Paris 7.

GIOULI, Voula; FOTOPOULOU, Aggeliki. (2012), Emotion verbs in Greek. From Lexicon-Grammar tables to multi-purpose syntactic and semantic lexica**.** In *Proceedings of* the 15th EURALEX International Congress, University of Oslo 7-11 August, 2012.

FOTOPOULOU Aggeliki ; MINI Marianna ; PANTAZARA Mavina ; MOUSTAKI Argiro (2009) La combinatoire lexicale des noms de sentiments en grec moderne. *Le lexique des émotions*, Iva Novacova & Agnes Tutin (eds), ELLUG, Grenoble.

GAVRILIDOU Zoi. (2002), Détermination des noms de sentiment en grec moderne. Langages 145, Paris : Larousse.

GUILLET**, A. ;** LECLERE**, C**. (1981), La restructuration du sujet. Langages 65. Paris, France

GROSS Maurice (1975), Méthodes en syntaxe, Paris: Hermann.

LAKOFF George (1970), Global Rules. In *Language* 46.

MATHIEU Yianique Y. (2000), Les verbes de sentiments. De l'analyse linguistique au traitement automatique, Paris: CNRS Editions.

PANTAZARA Mavina ; FOTOPOULOU Aggeliki; MOUSTAKI Argiro; MINI Marianna (2008), La description des noms de sentiments du grec moderne. In *Lingvisticae Investigationes* Vol. 31:2(2008), John Benjamins, Paris, France.

POSTAL P. M. (1971), Cross-Over Phenomena. New York: Holt, Rinehart & Winston.

RUWET N (1995), Les verbes de sentiment forment-ils une classe distincte dans la grammaire? H.B. Shyldkrot and L. Kupferman (eds). Tendances récentes en linguistique générale et française. Amsterdam-Philadelphia: Benjamins.
VALETOPOULOS Fridericos (2005), Ce que vous pensez des autres: La grammaire locale de la jalousie et de l'admiration. Lidil 32. Grenoble: Université Stendhal de Grenoble, 67-82.

Italian Verb-Particle Constructions: predicative element(s) and syntactic structure(s)

Daniela Guglielmo

Abtract: This paper is a contribution to the Predication Theory and syntactic-semantic analysis of Italian *Verb-Particle Constructions* (hereinafter VPCs) of idiomatic type like *buttare giù una lettera* (cf. 'to write down a letter'), *mandare avanti un'azienda* (cf. 'to carry on a business'). We aim at providing - within the Lexicon-Grammar framework (Gross M., 1975, 1981, 1998) and Operator-Arguments Grammar (Harris Z., 1968, 1976, 1978) – an original proposal regarding the predicative structure of Italian idiomatic VPCs arguing that they can be split into different syntactic types on the basis of two main issues: the predicative element of the construction (the so-called 'operator' in Harris's theory) and the pattern of variation which they exhibit. Our descriptive study is based on a lexicon-grammar database of 300 idiomatic VPCs with N_0 *V Part* N_1 sentence structure.

0. Introduction

We will take into account idiomatic Verb-Particle Constructions [1](hereinafter VPCs) of Italian language like *buttare giù una lettera* (cf. 'to write down a letter'), *mandare avanti un'azienda* (cf. 'to carry on a business'), *fare fuori un dolce* (cf. 'to eat up a cake'), *tirare su una parete* (cf. 'to pull up a wall'), showing that behind the same 'surface' structure **N_0 V Part N_1** they display different predicate-argument relations as well as different pattern of variation, so that a unique representation form and lexicon-grammar treatment cannot be called for all of them.

1. Compositional vs. idiomatic VPCs

The sentence structure under study is of the transitive form **N_0 V Part N_1** like the following examples:

(1)

a. *Max mette su un'attività* (cf. Max starts up a business)

b. *Eva fa fuori il gelato* (cf. Eva eats up an ice-cream)

where *mette su* and *fare fuori* are transitive verb-particle (VPart), *Max* and *Eva* are the first argument (N_0) while *un negozio* and *il gelato* are the second arguments (N_1). VPCs in (1) are '**idiomatic**' (or non compositional) because the meaning of the compound it is not analysable from the meanings of the two elements (*V* and *Part*) while in the following sentences:

(2)

a. *La ciminiera butta fuori fumo* (cf. 'The chimneys throws out smoke')

b. *Bob mette fuori la spazzatura* (cf. 'Bob puts out the garbage)

the verb-particle combinations *butta fuori* and *mette fuori* are '**compositional**' because the meaning of the compound is a function of the meaning of the two elements, verb and particle. Compositional VPCs like (2.a) are defined **redundant** (Schwarze 1985, Hampe 2002), while VPCs like (2.b) are defined **directional** (Jackendoff 2002, Bolinger 1971, Simone 1997, Masini-Iacobini 2006).

In the present paper we will partially revise such a previous simplified and traditional dichotomy (compositional vs. idiomatic) by identifying a novel in-between type of VPC in Italian language, which shares the non-literal interpretation and the lexical restrictions of idiomatic type (1) and the flexible syntactic status of compositional type (2) and that will be defined 'semi-fixed'. The proposal will be that in the semi-fixed type the particle plays a predicative role.

2. Lexicon-Grammar tables of Italian VPCs

Focusing only on **N_0 V Part N_1** uses of the type (1), a corpus of 300 idiomatic verb-particle was taken into account.[2] Data were divided into 10 separate lexicon-grammar classes (*V+fuori, V+su, V+giù, V+via, V+dentro, V+dietro, V+indietro, V+ avanti, V+sotto, V+ other particles*) on the basis of the locative particle on the right of the verb, as in Machonis (2009a, 2009b) classifications of English phrasal verbs.. Below we provide an extract of the table *V + giù* of Italian language, which encodes an example of the arguments in object and prepositional

[1] *A large bibliography of VPCs is available at* http://ling.univkonstanz.de/pages/home/dehe/bibl/PV.html *while an essential bibliography of Italian verb-particle constructions is available at the link* http://verbisintagmatici.caissa.it/Bibliography.html.

[2] *The study is corpus based. Data were collected from different sources: 1) monolingual dictionaries; 2) bilingual dictionaries; 3) on line dictionaries and website 4) by looking up in Google; 5) Linguistic articles cited; 6) native speaker competence; 7) VPCs extracted from* Italian Spoken language corpus LIP *(500.000 words) by Guglielmo (2010).*

positions (i.e. *N_1, Prep N_2*) the semantic class associated with N_1 (e.g. *buildings, telephones, texts, value, food*), the Italian paraphrase of the VPC and, finally, the corresponding English *phrasal verb*[3]:

N_0=:N hum	N_0= N anim	N_0=: Che F	Verb	Part	Ex. of N_1	Prep N_2	N_1=: N hum	N_1=: N -hum	N_1= :N concrete	N_1=: N abstract	Semantic Class of N_1	Paraphrase	Phrasal verb
+	+	-	buttare	giù	un muro	-	-	+	+	-	<buildings>	Demolire	to knock down
+	-	-	buttare	giù	il cell.	-	-	+	+	-	<telephones>	Riagganciare	to ring off
+	-	-	buttare	giù	una lettera	-	-	+	+	-	<texts>	Abbozzare	to write down
+	-	+	buttare	giù	i prezzi	del 20%	-	+	-	+	<values>	Ridurre	to drive down
+	-	+	buttare	giù	Max	-	+	-	-	-	-	Deprimere	to bring down
+	-	-	buttare	giù	un dolce	-	+	-	-	-	<food>	Ingoiare	to gulp down

Table 1: Transitive verb+particle uses followed by the particle "giù" (cf. down)

3. Difference in Predication

A more detailed analysis of data collected in lexicon-grammar tables of VPCs pointed out that transitive uses exhibit a considerable variety: even though their 'surface' sentence structure is the same (*N_0 V Part N_1*) they differ a lot in transformational behavior and syntactic cohesion degree. The goal of this work is to provide a systematic description of such a difference within the theoretical orientation to the grammar as a mathematical characterization of natural language, as outlined by Harris Z. (1968).[4] We consider, in fact, that a proper understanding of VPCs syntax and semantics necessitates the identification of element (or the sequence of elements) that in the construction plays the role of predicate or 'operator' (because 'it says something about its arguments'), as well as a deeper analysis of the syntactic variation degree of the pattern. Drawing from these main aims, we marked out from the initial database of transitive VPCs (table 1) four novel constructions types (depending on the predicative element identified):

Type 1. **Redundant**: constructions in which the predicative element is represented only by the verbal head as the particle is emphatic and can also not occur (e.g. *Buttare via un'occasione* ←→ *buttare un'occasione*, (cf. 'to throw away an opportunity');

Type 2. **Semi-fixed**: constructions in which the predicative role is played by the particle, (e.g. *mettere dentro il ladro*, cf. 'to send down the thief', 'to imprison') because the verbal head is variable (e.g. *mettere dentro il ladro* ←→ *sbattere dentro il ladro* ←→ *buttare dentro il ladro*) and/or it can also be 'missing' (e.g. *mettere dentro il ladro* ←→ *dentro il ladro*!);

Type 3. Fixed: constructions in which the predicative element is represented by verb and particle as a whole (e.g. *fare fuori il gelato*, lit. *to do out an ice cream*, cf. 'to eat up') because both of them are fixed, cannot vary and can occur only together (e.g. **fare il gelato*, **fuori il gelato*);

Type 4. **Frozen**: constructions in which all the sequence *verb-particle-N_1* plays the role of predicate because also N_1 is constrained ($N_1 = C_1$) while N_0 is free. This is the case of "frozen sentences" embedding a verb-particle (e.*g. mandare giù un boccone amaro*, cf. 'to send down a bitter pill', 'to swallow hard', *mettere su famiglia*, cf. 'to start a family'). Indicating with PRED between square brackets the predicative element of *N_0 V Part N_1* structure, we formalized the four 'new' VPCs types, identified in the current work, into the table below:

[3] *As suggested by Machonis (2008), in fact, the different uses of the same 'ambiguous' idiomatic Verb particle (like the lemma* buttare giù *in the table12) can be distinguished by associating each selected N_1 with a specific semantic class - or 'object class' within G. Gross' approach (Gross G. 1994) - and by adding such a semantic information into the LG tables. On the basis of such a formalization of the lexical restrictions on the arguments, a* Lexicon-Grammar Disambiguation Model *concerning Italian idiomatic VPCs was pointed out and applied on LIP corpus by Guglielmo (2010).*

[4] *The syntactic difference between* buttare giù un muro (cf. to knock down a wall) *and* buttare giù una lettera *(cf. "to write down a letter"), in fact, does not emerge from the simplified table 1.*

Type	DEFINITION	PREDICATIVE STRUCTURE	EXEMPLE	uses
1	Redundant	N_0 $[V]_{PRED}$ Part N_1	*Eva ha* [***buttato***]$_{PRED}$ ***via*** *un'occasione*	44 (15%)
2	Semi-Fixed	N_0 V $[Part]_{PRED}$ N_1	*Il poliziotto* ***ha messo*** [***dentro***]$_{PRED}$ *il ladro*	90 (30%)
3	Fixed	N_0 [V Part] $_{PRED}$ N_1	*Bob* [***ha fatto fuori***]$_{PRED}$ *il gelato*	106 (35%)
4	Frozen	N_0 [V $PartN_1$] $_{PRED}$	*Bob* [***ha mandato giù un boccone amaro***] $_{PRED}$	60 (20%)

Table 2: idiomatic VPCs transitive types distribution **Tot. 300**

This table shows that Italian idiomatic VPCs family does not represent a homogeneous class of constructions: they are situated along a *continuum* which ranges from more flexible to more cohesive constructions. In particular we consider the type 4, i.e. frozen VPCs, a subset of italian frozen sentences (Vietri, 1983) with which they share: (i) the ambiguity of meaning; (ii) the frozenness of the argument in object position, (iii) the attitude to block transformations. On the basis of these properties we argue that they need to be listed in dictionaries as complex, multi-word lexical entries and, for NLP applications, they need to be locate in texts as a block. In the next paragraph we will focus on the two central types of constructions, i.e. *semi-fixed* and *fixed*, which cover the large set of idiomatic VPCs (i.e. 65%). In particular we will describe the criteria used to distinguish between them, in term of different transformational properties which they accept.

4. Pattern of variation

Even though semi-fixed and fixed VPCs share (i) the non-compositionality of the meaning, (ii) the "same" surface form *N_0 V Part N_1*, and (ii) the lexical restrictions on N1 arguments, they differ a lot in predicate-argument requirements as well as in transformational behaviours. We used the following three test of variation to distinguish between fixed and semi-fixed: (a) syntactic decomposition (or 'factorization'); (b) variability of the head verb; (c) verbless use.

A) We call "**Syntactic Decomposition**" the factoring operation (also defined in mathematics 'factorization') of a "full" VPCs like:

(1) *Max mette dentro il ladro* (cf. "Max sends down the thief")

N_0 V Part N_1

into two sub-structures:

(1.1) Max mette # (1.2) Dentro il ladro

N V Part N

We argue that the "full" VPC in (1), i.e. *Max mette dentro il ladro*, with 'apparent' *N_0 V Part N_1* sentence form, must be regarded as a 'complex structure', resulting from the application (#) of a causative verb (*mettere*) together with its causative argument (*Max*) on the basic syntactic structure (1.2), i.e. *Dentro il ladro* (cf. 'down the thief') We formalize the syntactic decomposition property with the syntactic formula **N_0 V Part N_1 ←→ N_1 essere Part** indicating the synonymic relationship between a 'full' VPC and a support verb particle construction: *Max mette dentro in ladro ←→ il ladro è dentro*. We termed idiomatic VPCs passing such a syntactic decomposition test "**semi-fixed"** since they are less cohesive or 'assembled' and more flexible and 'decomposable'. They have a plus sign in the tables [+] under the property **N_1 essere Part** since they can be reduced to SVCs, as seen in (1.3). On the other hand, VPCs which do not pass the syntactic decomposition test - with a minus sign [-] in tables under the property **N_1 essere Part** - were defined **"fixed"** because of a higher syntactic cohesion property. Given the sentences (2) *Bob fa fuori il gelato* (cf. 'Bob eats up an ice cream') and (3) *Eva ha messo su un negozio* (cf. 'Eva has set up a shop'), for instance, the syntactic decomposition of them into the following substructures is not possible:

(2.1) Bob fa # (2.2) * fuori il gelato

(3.1) Eva ha messo # (3.2) * su un negozio

This suggests that *fare fuori* and *mettere su* display a higher syntactic and lexical cohesion and they must be considered 'fixed' VPCs, not decomposable. *Fare* (lit. "to do") and *mettere* (lit. "to put") in fact, falling into the idiomatic verb particle uses (2) and (3) lose their original causative value and they are not related with support-verb sentences:

(2.3) Bob fa fuori un gelato ←→ * il gelato è fuori

(3.3) Eva ha messo su un negozio ←→ *il negozio è su

In fact, while semi-fixed VPCs accept the 'factitive' paraphrase **N_0 cause N_1 to be Part**, as in (1.4)[5]:

(1.4) Max mette dentro il ladro ←→ Max fa che il ladro sia dentro

fixed VPCs, like *fare fuori* and *mettere su*, never accept it, as it is easily tested by (2.4) and (3.4):

(2.4) Bob fa fuori un gelato ←→ *Bob fa che il gelato sia fuori

(3.4) Eva ha messo su un negozio ←→ * Eva fa che un negozio sia su

B) Moreover, with regards to the larger pattern of variation involving semi-fixed VPCs, we noted that the causative head verb forming the combination (i.e. *mettere* in *mettere dentro il ladro*) can be easily substituted by other synonymic verbs without affecting the syntax and semantic of the construction, like in the following network of "alloconstructions", or "paraphrastic equivalence class" in harrisian terms (1.5):

Max (mette + porta + sbatte + butta + manda + spedisce) DENTRO il ladro

where the particle is the semantic and syntactic "centre" of the constructions network (i.e. the operator) selecting the human argument, i.e. DENTRO (Num), while the head-verb is variable into a finite range of possibilities: the verb *mettere* can be replaced by other causative motion verb, such as *portare* (cf. to take), *sbattere* (cf. to shunt), *buttare* (cf. to throw), *mandare* (cf. to send), *spedire* (cf. to send). This is the reason for which, in our view, semi-fixed VPCs are associated with the predicative structure **N_0 V [Part]$_{PRED}$ N_1** with a fixed slot for the particle (playing the role of predicate) and a semi-free slot for the verb[6]. The **variability of head verb** was formalized by the syntactic formula **N_0 V_x Part N_1 ←→ N_0 V_y Part N_1** which is, as demonstrated in (1.5) well accepted by semi-fixed uses, but, conversely, completely rejected by fixed VPCs like *fare fuori* and *mettere su*, as tested respectively by:

(2.5) Bob fa fuori un gelato ←→ *Max (manda + porta +…) fuori un gelato

(3.5) Eva mette su un negozio ←→ * Eva (porta + manda + tira +..) su un negozio

C) Finally, semi-fixed VPCs pass the third diagnostic test, i.e. "**verbless use**" (V=: E), by accepting a 'small' sentence consisting only of a predicative particle plus its selected argument, with no head verb in pre-particle position, as in (1.6): *Max mette dentro il ladro* ←→ *dentro il ladro*!

The following formula represents the equivalence between a full verb-particle use and the corresponding verbless use: **N_0 V Part N_1 ←→ Part N_1.** Semi-fixed VPCs have a plus sign [+] in the tables under the propriety **Part 1** indicating the relative synonymy with **N_0 V Part N1** while fixed VPCs like *fare fuori* or *mettere su* have a minus sign [-], because they do not pass such a variation test:

(2.6) Max fa fuori il gelato ←→ *fuori il gelato!

(3.6) Eva mette su un negozio ←→ *su un negozio!

Conclusions

In this work we assumed that the traditional semantic dichotomy involving Italian VPCs (compositional vs. idiomatic) does not suffice to account to their syntactic behaviour. Our approach distinguished, in fact, on the basis of some transformational criteria (*optional particle usage, frozenness of the N_1 argument, factorization, substitution of the head verb, verbless usage*) at least four types of VPCs (i.e. redundant, semi-fixed, fixed, frozen) situated along a *continuum* which ranges from more flexible to more "blocked" constructions. The main lexicon-grammar implication of this study is that, since every type exhibits differences in predication and in the pattern of variation, a parallel, independent LG taxonomy needs to be called for each type. As we demonstrated so far, the predicate and 'lexical entry' of Italian idiomatic VPCs corresponds to: (i) the head verb for redundant constructions, the combination verb plus particle for fixed constructions, (iii) all the sequence verb plus particle plus constrained noun for frozen VPCs; (iv) only the particle for semi-fixed VPCs. In particular, with regards to semi-fixed VPCs, we stressed the hypothesis that the underlying structure should not be considered the full or 'extended' N0 V Part N1 structure, as in the case of fixed VPCs, but the embedded, minimal Part N structure, that

[5] *This gives rise to one of the main hypothesis carried out in this work, that idiomatic VPCs are syntactically different and that semi-fixed VPCs, entering into N_0 V Part N_1 transitive structure, have a nature essentially 'causative', with the post verbal NP (i.e. N_1) acquiring the property denoted by the particle (the propriety of being down, up, out or whatever) as observed for English compositional VPCs by Bolinger (1971) and more explicitly proposed in Svenonius (1994). In particular the causative transitive structure (N_0 V Part N_1) is related, via a paraphrastic equivalence relation (←→) with the support verb structure* N_1 essere Part, *which is of a 'resultative' nature, and such a CAUSATIVE ←→ RESULTATIVE relation needs to be encoded into the Lexicon-Grammar.*

[6] *A 'semi-fixed' slot for the verb means that when we encode each lexical item we discover lexical restrictions on the causative motion verbs accepted as variants of 'be', like in the case of 'be Prep' structures analysed by Vietri (1996).*

we consider the predicative kernel of the construction and that we defined ‘verbless particle construction” (Cappelle 2005, Guglielmo 2012).

References

Bolinger, D. (1971), The phrasal verbs in English, Harvard University Press, Cambrige.

Cappelle, B. (2005), Particle Patterns in English. A comprehensive coverage, PHD Dissertation, Leuven.

Gross, G. (1994), Classes d’object et description des verbs, Langages, New York: Accademic Press.

Gross, M. (1975), Methodes en syntaxe, Paris, Hermann.

Gross, M. (1981), Les bases empirique de la notion de prédicate sémantiques, Languages 63, Paris, Larousse.

Gross, M. (1998), La foncion sémantiques des verbes supports, Travaux de linguistique 37.

Guglielmo D. (2010), Parlare con i Verbi Sintagmatici, in Proceedings of the 3th International Conference on Spoken Communication (GSCP), Vol. I, M. Pettorino, A. Giannini, F. Dovetto (eds.) Napoli, Università di Napoli L’Orientale (http://opar.unior.it/336/1/La_comunicazione_parlata_3_-_vol._I.pdf

Guglielmo D. (2012), Verbless particle constructions in Italian: formal analysis and empirical investigations, paper accepted at International Conference of the Society of Italian Language (=SLI), sett. 2012, Siena.

Hampe, B. (2002), Superlative Verbs: a corpus based study of semantic redundancy in English verb particle constructions, Gunter Narr Verlag, Tubingen.

Harris, Z. S. (1968) Mathematical Structure of Language, (Interscience tracts in pure and applied mathematics 21), N.Y. : Wiley.

Harris, Z. S., (1976) Notes du Cours de Syntaxe, Maurice Gross es., Paris: Editions du Seuil

Harris, Z.S., (1978), Operator Grammar of English, Linguisticae Investigationes II:1, 55-92

Iacobini, C., Masini, F. (2006), The emergence of verb-particle constructions in Italian: locative and actional meanings, in Morphology, 16.

Jackendoff, R. (2002), English particle constructions, the lexicon and the autonomy of syntax in Dehè N. et alii, (eds.), Verb-Particle explorations, Mouton de Gruyter, New York

Machonis, P. (2008), Disambiguating Phrasal verbs, in Linguisticae Investigationes 31.1

Machonis, P. (2009a), Compositional phrasal verbs with ‘up’: direction, aspect, intensity, in Linguisticae Investigationes 32.2

Machonis, P. (2009b), Nooj dictionary of English phrasal verbs: http://www.nooj4nlp.net/

Simone, R. (1997), Esistono I verbi sintagmatici in Italiano? In De Mauro T.& Lo Cascio V. (eds.), Lessico e Grammatica. Teorie Linguistiche e applicazioni lessicografiche (SLI 36), Roma, Bulzoni.

Svenonius, P. (1996). "The verb-particle-alternation in the Scandinavian languages". Ms. University of Tromsø.

Schwarze, C. (1985), Uscire e “andare fuori”: struttura sintattica e semantica lessicale, in Franchi de Bellis, A. Savoia (eds.) Sintassi e morfologia della lingua italiana d’uso. (SLI 24), Roma, Bulzoni.

Vietri, S. (1983), On the study of idiomatic expression in italian, in Proceedings of the XVII SLI Conference, De Belllis & Savoia L., (eds.) Bulzoni, Roma.

Vietri, S. (1996), The syntax of Italian verb essere Prep, in Linguisticae Investigationes XX:2, John Benjamins Publishing Company, Amsterdam/Philadelphia.

Adverbiaux de conviction personnelle dans un corpus parallèle grec-français

Fryni Kakoyianni-Doa, Stavroula Voyatzi, Eleni Tziafa

Résumé : Le présent article est consacré à l'étude descriptive et comparative d'un sous-ensemble d'adverbiaux de phrase français et grecs qui se réfèrent à la conviction personnelle du locuteur vis-à-vis de l'information transmise, tels que *à mon avis* (*κατά τη (γνώμη + άποψή) μου*), *selon moi* (*κατ' εμέ*), etc. Plus particulièrement, il est examiné si et dans quelle mesure les propriétés morphosyntaxiques et sémantiques (cf., pour le français, MOLINIER, 1984 ; 2000, GROSS, 1986 ; 1990 ; pour le grec, VOYATZI, 2006, KAKOYIANNI-DOA, 2008) ainsi que les similarités et divergences proposées pour ces adverbiaux dans un travail précédent (KAKOYIANNI-DOA, 2008), sont attestées dans un corpus réel.

Abstract: This paper aims at describing and comparing a subset of utterance-level French and Greek adverbials which express the personal opinion of the speaker towards the information transmitted, e.g. *à mon avis* (*κατά τη (γνώμη + άποψή) μου*) 'in my opinion', *selon moi* (*κατ' εμέ*) 'for me', etc. It is examined, in particular, whether and to what extent the morphosyntactic and semantic properties (cf. for French, MOLINIER, 1984; 2000, GROSS, 1986 ; 1990, and for Greek, cf. VOYATZI, 2006, KAKOYIANNI-DOA, 2008) as well as the similarities and differences proposed for these adverbials in a previous work (KAKOYIANNI-DOA, 2008), are attested in a real corpus.

1. Introduction

Si le comportement des adverbes pose encore d'épineux problèmes à ceux qui se penchent sur leur analyse linguistique, c'est parce que d'une part, la classe adverbiale est une classe difficile à cerner et à définir par un ensemble de propriétés homogènes, et que d'autre part, malgré la présence de travaux descriptifs récents, les résultats de ces derniers sont peu appliqués. Or l'étude des rapports de la langue et de son usage à partir de corpus réels s'avère essentielle et complémentaire à l'analyse linguistique. Le présent article est consacré à l'étude d'un sous-ensemble d'adverbiaux de phrase dans un corpus parallèle français-grec. Plus particulièrement, seront examinés des adverbiaux se référant à la source de l'information, en partie décrits par NØLKE (1993), MOLINIER ; LEVRIER (2000), BORILLO (2004), COLTIER ; DENDALE (2004) et par KAKOYIANNI-DOA (2008) dans une étude de perspective comparative entre français et grec moderne. Comme ils contiennent un morphème déictique de première personne (*à **mon** avis* (*κατά τη (γνώμη + άποψή) **μου***), *selon **moi*** (*κατ' **εμέ***), nous les appelons « adverbiaux de conviction personnelle ».

L'article comprend trois parties. La première décrit le cadre méthodologique adopté pour l'analyse de ces adverbiaux ainsi que leurs principales propriétés morphosyntaxiques et sémantiques. La deuxième partie présente le corpus parallèle français-grec à partir duquel nous avons travaillé pour vérifier les données précédemment étudiées. Enfin, nous exposons les nouveaux résultats issus du corpus concernant les adverbiaux de « conviction personnelle ». Seront ainsi présentées, dans une perspective comparative, la fréquence des différents types d'emplois ainsi que les nouvelles entrées (variantes).

2. Adverbiaux de conviction personnelle en français et en grec

2.1. Cadre méthodologique : classification et recensement

Pour le premier recensement des adverbiaux de « conviction personnelle » grecs (VOYATZI, 2006 ; KAKOYIANNI-DOA, 2008), nous nous sommes inspirées du modèle de classement des adverbiaux français aussi bien pour ce qui était de la classification syntaxico-sémantique des adverbes monolexicaux (ou simples) (MOLINIER, 1984; MOLINIER ; LEVRIER, 2000) que du classement morphosyntaxique des adverbes polylexicaux (ou complexes) (GROSS, 1986 ; 1990). La figure 1, présentée en annexe, illustre les différents niveaux de classement des adverbiaux grecs (VOYATZI & KAKOYIANNI-DOA, 2009) et français ainsi que les classes et sous-classes établies.

Ces adverbes une fois recensés, nous avons ici procédé à leur classement. Une première répartition, effectuée au niveau morphosyntaxique, nous permet de distinguer les adverbes monolexicaux (ou simples) des adverbes polylexicaux (ou complexes). Une deuxième distinction, s'appuyant sur la notion de compositionnalité, permet de différencier les adverbes productifs et notamment ceux qui sont dérivés d'adjectifs, des adverbes non compositionnels, ici appelés 'adverbes (semi-)figés'. Une troisième répartition est opérée au niveau fonctionnel, et permet de séparer les adverbes intégrés à la proposition (ceux qui sont rattachés au prédicat verbal ou à tout autre constituant de la phrase) des adverbes de phrase (ceux qui permettent de modifier une phrase entière). Pour ce qui est des adverbes de phrase, ils sont divisés en deux sous-classes syntaxico-sémantiquement homogènes : d'une part, celle des adverbes conjonctifs servant à établir un lien entre des unités du discours (ceux-ci correspondraient aux 'connecteurs logiques') et, d'autre part, la sous-classe des adverbes disjonctifs qui expriment la prise de position du locuteur vis-à-vis de son énoncé ou l'attitude du locuteur par rapport au contenu propositionnel. Enfin, la dernière distinction est effectuée au niveau sémantique : elle permet d'établir un certain nombre de sous-catégories plus fines et homogènes également présentées en annexe.

Le premier recensement nous a donné une quarantaine d'adverbiaux de ce type dans les deux langues (cf. Tableau 1, section 2.3). Les variantes sont par contre plus nombreuses en grec moderne.

2.2. Définition et propriétés

Les adverbiaux de « conviction personnelle » se présentent à l'intérieur de la classe des adverbiaux d'énonciation (ou disjonctifs de style[1]) qui comprennent les quatre sous-classes sémantiques suivantes :

(i) adverbiaux indiquant la disposition psychologique ou morale du locuteur vis-à-vis de l'interlocuteur : *sincèrement* (*ειλικριν(ά + ώς), en toute sincérité* (*με (κάθε + πάσα) ειλικρίνεια),*

(ii) adverbiaux exprimant un commentaire du locuteur sur la formulation de l'énoncé : *simplement (απλ(ά + ώς), en termes simples (με (απλά λόγια + απλές κουβέντες)),*

(iii) adverbiaux concernant la source de l'information : *selon moi* (*κατ' εμέ*), *de source sûre (από έγκυρη πηγή)* et

(iv) adverbiaux d'individuation : personnellement (προσωπικ(ά + ώς), quant à moi (όσο για μένα).

Les adverbiaux que nous examinons dans le présent article font partie de ceux concernant la source de l'information ; ils servent à représenter le contenu propositionnel comme une opinion résultant d'une réflexion, d'une impression ou d'un sentiment personnel de la part du locuteur. Nous identifions les adverbiaux de « conviction personnelle » par les propriétés de structure suivantes :

- présence dans ces formes syntaxiquement complexes de substantifs tels que *avis (άποψη + γνώμη), point de vue (σκοπιά), jugement (κρίση)* accompagnés du déterminant possessif de première personne *mon (δικ(ός-ή-ό) μου, μου), notre (δικ(ός-ή-ό) μας, μας),* ou présence du pronom personnel tonique de première personne *moi (εμένα, μενα, εμέ), nous (εμάς, μας),*

- présence dans des formes phrastiques de verbes d'opinion *penser (νομίζω)* et *croire (πιστεύω, θεωρώ)* à la première personne du présent *à ce que je pense (απ' ό,τι νομίζω (Ε + εγώ)).*

Ces adverbiaux présentent en français et en grec moderne des structures homogènes dans la mesure où ils contiennent un morphème déictique de première personne. Sémantiquement, ils indiquent que le locuteur présente un point de vue subjectif, une constatation personnelle ; ce faisant, il informe l'interlocuteur que la source de l'information vient de lui-même et que celle-ci n'est donc pas une vérité absolue :

(1) ***À mon avis****, il ne fera pas beau demain*

(1a) ***Κατά τη γνώμη μου****, δε θα έχει ωραίο καιρό αύριο*

le locuteur informe que son opinion peut ne pas être partagée ou peut être contredite, que son jugement n'est pas obligatoirement juste. Il peut même signaler explicitement la possibilité d'un démenti et, dans ce cas, l'adverbial semble jouer une fonction pragmatique d'atténuation (BORILLO, 2004) :

(2) ***À mon avis****, il ne fera pas beau demain mais je peux me tromper*

(2a) ***Κατά τη γνώμη μου****, δε θα έχει ωραίο καιρό αύριο αλλά μπορεί και να κάνω λάθος*

Si ces adverbiaux dans les deux langues refusent la cooccurrence avec des verbes à la première personne (ou d'autres formulations) dénotant des sensations (exemple 3), des affects ou des impressions (exemple 4) pour lesquels le locuteur est le seul juge

(3) ****(À mon avis + Selon moi)****, je ne me sens pas bien*

(3a) ****(Κατά τη γνώμη μου + Κατ' εμέ)****, δεν αισθάνομαι καλά*

(4) ****(À mon avis + Selon moi)****, j'ai l'impression que tu exagères*

(4a) ****(Κατά τη γνώμη μου + Κατ' εμέ)****, έχω την εντύπωση ότι υπερβάλλεις*

dans les deux langues, ces adverbiaux peuvent être suivis de verbes d'opinion tels que *croire (πιστεύω, θεωρώ)* ou *penser (νομίζω)* malgré l'impression de pléonasme :

(5) ***(À mon avis + Selon moi)****, je (crois + pense) que tu as du talent*

(5a) ***(Κατά τη γνώμη μου + Κατ' εμέ)****, (νομίζω + πιστεύω) ότι έχεις ταλέντο*

Enfin, ils peuvent se présenter en début de phrase, à la fin ou en incise :

(6) ***(À mon avis + Selon moi + À ce que je pense)****, il serait un très bon candidat*

(6a) ***(Κατά τη γνώμη μου + Κατ' εμέ + Απ' ότι νομίζω)****, θα ήταν ένας πολύ καλός υποψήφιος*

(7) Il serait un très bon candidat ***(à mon avis + ?selon moi + ?à ce que je pense)***

(7a) Θα ήταν ένας πολύ καλός υποψήφιος ***(κατά τη γνώμη μου +?κατ' εμέ + ?απ' ότι νομίζω)***

(8) Il serait, ***(à mon avis + selon moi + à ce que je pense)****, un très bon candidat*

(8a) Θα ήταν, ***(κατά τη γνώμη μου + κατ' εμέ + απ' ότι νομίζω)****, ένας πολύ καλός υποψήφιος*

[1] Cf. Molinier (2000).

2.3. Représentation dans les tables du Lexique-Grammaire

Une partie des adverbiaux de « conviction personnelle » sont représentés dans les tables du Lexique-Grammaire (désormais LG) des adverbes polylexicaux (ou complexes) figés pour le français (GROSS, 1990) et pour le grec (VOYATZI, 2006). Ils sont ainsi répartis dans plusieurs classes différentes, chaque classe correspondant à une classe morphosyntaxique des adverbes polylexicaux (ou complexes) figés. Pour ce qui est du français, seuls les adverbiaux *pour moi, à mon sens, à mon sentiment* et *à mon avis* (sous-structure de *à mon humble avis)* sont présents dans les tables LG. Nous présentons dans le Tableau 1 ci-dessous, l'ensemble des adverbiaux étudiés dans les deux langues (cf., MOLINIER ; LEVRIER, 2000 ; GROSS, 1990 ; VOYATZI, 2006 ; KAKOYIANNI-DOA, 2008).

FRANÇAIS		**GREC MODERNE**	
Entrée adverbiale	**Classe**	**Entrée adverbiale**	**Classe**
pour moi *d'après moi* *selon moi*	PC	*για (εμένα + μενα)* *κατ' εμέ*	GPC
à mes yeux *à mon avis* *à mon point de vue* *à mon sens* *à mon sentiment* *de mon point de vue* *selon mon intuition* *selon mon jugement*	PDETC	*στα δικά μου μάτια* *κατά τη δική μου (γνώμη + άποψη)* *κατά τη δική μου αντίληψη* *κατά τη δική μου εκτίμηση* *από (τη δική μου σκοπιά + δικής μου σκοπιάς)* *?κατά τη δική μου διαίσθηση* *κατά τη δική μου κρίση*	GPAC
		στα μάτια μου *κατά τη(ν) (γνώμη + άποψη) μου* *κατά την αντίληψή μου* *κατά την εκτίμησή μου* *?από τη σκοπιά μου* *?κατά τη διαίσθησή μου* *κατά την κρίση μου*	GPDETC
à mon humble avis	PAC	*κατά την ταπεινή μου (γνώμη + άποψη)* *κατά την προσωπική μου (γνώμη + άποψη)* *κατά την προσωπική μου αντίληψη* *κατά την προσωπική μου διαίσθηση* *κατά την προσωπική μου κρίση*	GPAC
à ce que je crois *à ce que je pense* *si vous voulez mon avis* *si vous me demandez mon avis*	PF	*απ' (ό,τι + ?όσο) πιστεύω (Ε + εγώ)* *απ' (ό,τι + ?όσο) (νομίζω + θεωρώ) (Ε + εγώ)* *(εάν + αν) θέλετε τη(ν) (γνώμη + άποψή) μου* *(εάν + αν) ζητάτε τη(ν) (γνώμη + άποψή) μου*	GPF
Total : 16 entrées		**Total : 25 entrées (18 adverbiaux)**	

Tableau 1. *Répartition des « adverbiaux de conviction personnelle » dans les tables LG*

Les adverbiaux grecs exprimés sous forme de groupe nominal prépositionnel ont fait l'objet d'un dédoublement dans les tables LG (classes GPAC et GPDETC). Ceci est dû aux propriétés générales du déterminant possessif du grec moderne (TONNET, 2006 : 187-188). Ainsi, parallèlement, à *κατά τη γνώμη μου/selon le avis mon* (à mon avis) qui contient la variante 'réduite' et, formellement parlant, simple, du déterminant possessif ayant les propriétés d'un clitique et la position postnominale, nous pouvons également avoir *κατά τη δική μου γνώμη/selon le mien mon avis* (à mon avis à moi), et *κατά τη δική μου τη γνώμη/selon le mien mon l'avis* (redoublement du déterminant défini désigné dans la grammaire générative comme 'determiner spreader', cf. ALEXIADOU ; WILDER, 1998). On relève enfin la variante discontinue et emphatique *κατά τη γνώμη τη δική μου/selon l'avis le mien mon*. Par ailleurs, la présence du possessif simple fait obligatoirement intervenir la variante à deux accents (outre l'accent principal, un accent supplémentaire est ajouté sur la dernière syllabe pour les vocables de trois syllabes ou plus accentués sur l'antépénultième) : *κατά την **εκτίμησή** μου/selon le sentiment mon*

(à mon sentiment) *vs κατά τη δική μου **εκτίμηση**/selon le mien mon sentiment* (à mon sentiment à moi). Ces propriétés sont représentées explicitement dans la table GPAC des adverbiaux (semi-)figés grecs (VOYATZI, 2006 : 263-265, 267-269, 299-301).

Notons enfin, pour ce type d'adverbiaux grecs, l'alternance du déterminant possessif complexe

δικ(ός-ή-ό) μου/mien mon (à moi) avec l'adjectif προσωπικ(ός-ή-ό) (personnel), ce qui donne lieu à des formes telles que κατά την προσωπική μου εκτίμηση/selon le personnel mon sentiment (à mon sentiment).

3. Étude expérimentale

3.1. Le projet Corpus Parallèle français-grec

La présente étude s'inscrit dans le cadre du projet de recherche « Corpus Parallèle français-grec », initié par Fryni Kakoyianni-Doa et entièrement financé par l'Université de Chypre (2012-2013). Ce projet projette la création d'un corpus électronique parallèle bilingue français-grec de 1.000.000 de mots. Plus particulièrement, il envisage, dans un premier temps, la création d'une base de données destinée à une analyse contrastive des éléments adverbiaux pour la paire linguistique français <> grec. Ce corpus, constitué de textes parallèles de genres divers, vise non seulement l'applicabilité déjà connue du corpus de textes (WICHMANN *et al.*, 1997 ; PARTINGTON, 1998 ; SINCLAIR, 2004) mais aussi entend appliquer les notions de « parallélisme » et de « comparabilité » en linguistique appliquée. La constitution de ce corpus, s'appuyant sur les dernières tendances de la linguistique de corpus, viendra combler l'absence d'une description exhaustive et statistique des éléments adverbiaux ainsi que les manques et besoins de ressources authentiques pour le domaine de l'enseignement des langues étrangères. Le corpus, constitué de textes appartenant à des genres différents, comprendra des textes politiques, littéraires, journalistiques et éducatifs.

3.2. Le corpus d'étude

Pour constituer un premier corpus d'étude adapté à la problématique du présent article, nous avons porté une attention particulière à un corpus parallèle existant pour la paire français-grec : le corpus *Europarl* (KOEHN, 2005) suffisamment étendu pour permettre au travail d'extraction de produire des résultats crédibles. Librement disponible sur Internet, il comprend les actes du Parlement Européen recueillis entre 1996 et 2011, et inclut les versions d'une vingtaine de langues européennes parmi lesquelles le français et le grec. Le corpus français comprend une cinquantaine de millions de mots et le grec, une trentaine de millions de mots. Quoique l'ensemble du corpus ne soit pas toujours aligné phrase par phrase, nous l'avons choisi pour la qualité à la fois de l'alignement (au niveau des textes et des paragraphes) et de la traduction, ainsi que pour l'adéquation du contenu avec le domaine adverbial (plus spécifiquement, avec les concepts concernés par les synonymes à traduire). Nous présentons ci-dessous un extrait du corpus aligné au niveau des « adverbes de conviction personnelle » :

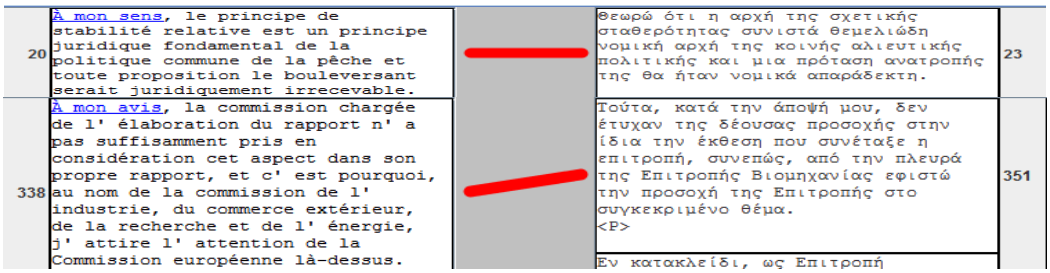


Figure 2. *Extrait du corpus Europarl français-grec : alignement au niveau des « adverbes de conviction personnelle »*

Dans cet extrait, la phrase (20) du corpus français contient l'adverbial *à mon sens* lequel est « traduit » au corpus parallèle grec par le verbe d'opinion *θεωρώ* (je pense) à la première personne du présent. Ensuite, dans la phrase (338) du corpus français, l'adverbial *à mon avis*, occupant la position initiale, est « traduit » au corpus parallèle grec par l'adverbial *κατά την άποψή μου* (à mon avis) présenté en incise.

Pour l'identification des fréquences des adverbiaux, nous avons opéré un pré-traitement spécifique des unités polylexicales (les fréquences de mots simples sont automatiquement générées par Unitex). À cet effet, nous avons conçu des listes d'adverbes polylexicaux accompagnés de leurs fréquences, en utilisant les concordances produites après application de graphes sous Unitex (PAUMIER, 2003 ; 2006). Les fichiers obtenus initialement au format .html ont été par la suite convertis en .txt (utf8). Puis, nous avons extrait leur fréquence grâce à l'application des clusters/ngrams compris dans le programme AntConc (ANTHONY, 2011). Nous avons ainsi construit des ngrams de 2 à 6 mots pour la langue française et de 2 à 10 mots pour la langue grecque, accompagnés de leur fréquence, dans le cas qui nous occupe, la liste des adverbes. Dans ces listes, nous avons pu identifier les « adverbiaux de conviction personnelle » moyennant deux graphes (un pour chaque langue) qui ont été spécialement conçus pour leur reconnaissance. Cette opération nous a permis d'obtenir les résultats chiffrés qui sont exposés en annexe (cf. Tableau 2).

4. Résultats

Les résultats quantitatifs issus de l'alignement des textes et de leur analyse automatique avec Unitex ont donné 8.958 occurrences d'« adverbiaux de conviction personnelle » français dont 4.581 occupent la position initiale dans

la phrase. Pour ce qui est du grec, on compte 13.609 occurrences d'adverbiaux dont 4.583 se trouvent en tête de phrase. Bien que le corpus grec (3M mots) soit manifestement moins important que le corpus français (5M mots), nous observons cependant une présence de ces adverbiaux nettement plus élevée (+34%) que celle en français. Le Tableau 2, présenté en annexe, fournit les détails de la fréquence de ces adverbiaux dans les deux langues. La répartition des adverbiaux selon leur structure lexicale interne est illustrée dans le Tableau 3 ci-dessous :

		FRANÇAIS	**GREC MODERNE**
Prép PRO	PC/ GPC	4.329	1.539
Prép Poss N	PDETC/ GPDETC	3.808	8.346
Prép Poss Adj N	PAC/ GPAC	31	109
Prép Pfigée	PF/GPF	8	2
personnellement/ προσωπικά		782	3.496
	Total	**8.958**	**13.492**

Tableau 3. ***Répartition des « adverbiaux de conviction personnelle » selon leur structure lexicale interne***

Cette étude nous a également permis de recenser de nouvelles structures. Ainsi, pour le français, la préposition *selon* combinée avec les substantifs *avis, opinion, point de vue* y est amplement utilisée. L'emploi du nom *opinion* dans des constructions telles que *à mon opinion, selon mon opinion* est très courant. Par ailleurs, nous rencontrons des structures comprenant la coordination de deux modifieurs adjectivaux du type *à mon humble et (modeste + partial + très subjectif)* avis.

Pour ce qui est du grec moderne, les nouvelles données sont plus nombreuses. Nous avons compté 8.334 formes comprenant le déterminant possessif 'réduit' (ou simple) contre 60 comprenant le possessif complexe. Nous avons relevé une nouvelle construction *κατά την πεποίθησή μου/selon la conviction ma* n'ayant pas de correspondant en français. De plus, une nouvelle structure morphosyntaxique est repérée : *Prép PRO Adv=: για (εμένα + μένα) προσωπικά/pour moi personnellement*, qui devrait faire l'objet de la classe GPCA des adverbes polylexicaux (semi-)figés. Enfin, la variation du groupe nominal au sein de ces adverbiaux est particulièrement riche : présence d'un modifieur adverbial (*κατά την ταπεινή* ***πάντα*** *άποψή μου/selon le humble* ***toujours*** *avis mon*), variation du modifieur adjectival (*κατά τη* ***μη έγκυρη*** *άποψή μου/selon le* ***non valide*** *avis mon*), coordination de modifieurs adjectivaux (*κατά την* ***ταπεινή και φτωχική*** *μου γνώμη/selon le* ***humble et pauvre*** *mon avis*), et constructions plus complexes (*κατά την* ***ταπεινή και ίσως αδιάφορη στο ευρύ κοινό*** *άποψή μου/selon le* ***humble et peut-être indifférent au large public*** *avis mon*).

Conclusion et perspectives

Si dans la recherche linguistique de nos jours, l'intuition linguistique ne peut plus donner de résultats scientifiques crédibles ou quantitativement vérifiables, les corpus électroniques en linguistique, à la fois bases de recherche et outils de description, sont un outil indéniable pour les chercheurs actuels. Nous avons vu que pour les « adverbiaux de conviction personnelle », le corpus *Europarl* nous a permis de réaliser une meilleure description de leurs propriétés fondée sur la réalité linguistique. Notamment, nous avons pu relever, pour le français et pour le grec moderne, la fréquence de ces adverbiaux, la variabilité de modifieurs adjectivaux, ainsi que de nouvelles structures et combinaisons. En outre, les données tirées de ce corpus serviront de base pour une recherche plus approfondie, nous permettant de soulever également toute ambiguïté possible. Enfin, l'alignement des textes de ce corpus nous permettra de relever les correspondances effectuées pour les adverbiaux de conviction personnelle dans les deux langues ainsi que les cooccurrences et combinatoires avec les différents types de prédicats.

Références bibliographiques

ALEXIADOU Artemis ; WILDER Chris (1998), Adjectival Modification and Multiple Determiners, in : ALEXIADOU Artemis et WILDER Chris (éds.), *Possessors, predicates and movement in the determiner phrase*, Amsterdam/Philadelphia, John Benjamins Publishing Co.

ANTHONY Laurence (2011), *AntConc* (Version 3.2.2) [Computer Software], Tokyo, Japan, Waseda University, http://www.antlab.sci.waseda.ac.jp (dernière consultation le 20 mai 2012).

BORILLO Andrée (2004), Les « adverbes d'opinion forte » *selon moi, à mes yeux, à mon avis, … :* point de vue subjectif et effet d'atténuation, *Langue française 142*, Paris, Larousse, p. 31-40.

COLTIER Danielle ; DENDALE Patrick (2004), La modalisation du discours de soi : éléments de description sémantique des expressions *pour moi, selon moi* et *à mon avis, Langue française 142*, Paris, Larousse, p. 41-57.

GROSS Maurice (1990), Grammaire transformationnelle du français. 3-Syntaxe de l'adverbe, Paris, ASSTRIL.

KAKOYIANNI-DOA Fryni (2008), *Adverbes de phrase français et grecs : étude contrastive et perspectives didactiques*, Thèse de doctorat, Toulouse, Université de Toulouse-Le Mirail, p. 229-237.
KOEHN Philipp (2005), Europarl: A Parallel Corpus for Statistical Machine Translation, in : *Proceedings of the 10th Machine Translation Summit*, Phuket, Thailand, p. 19-86.
MOLINIER Christian (1984), *Étude syntaxique et sémantique des adverbes de manière en –ment*, Thèse de doctorat d'État, Toulouse, Université de Toulouse-Le Mirail.
MOLINIER Christian ; LÉVRIER Françoise (2000), *Grammaire des adverbes. Description des formes en –ment*, Genève-Paris, Droz.
NØLKE Henning (1993), *Le regard du locuteur*, Paris, Éditions Kimé.
PARTINGTON, Alan (1998), *Pattern and Meanings*. Using Corpora for English Language Research and Teaching. Amsterdam, John Benjamins Publishing Company.
PAUMIER Sébastien (2003), *De la reconnaissance de formes linguistiques à l'analyse syntaxique*, Thèse de doctorat, Paris, Université de Marne-la-Vallée.
PAUMIER Sébastien (2006), *Manuel d'utilisation du logiciel Unitex*, Paris, Université de Marne-la-Vallée, http://www-igm.univ-mlv.fr/~unitex/manuelunitex.pdf (dernière consultation 20 mai 2012).
SINCLAIR, John (2004), *How to Use Corpora in Language Teaching*. Amsterdam, John Benjamins Publishing Company.
TONNET Henri (2006), *Précis pratique de grammaire grecque moderne*, Paris, Langues & Mondes − L'Asiathèque.
VOYATZI Stavroula (2006), Description morphosyntaxique et sémantique des adverbes en vue d'un système d'analyse automatique des textes grecs, Thèse de doctorat, Paris, Université de Marne-la-Vallée.
VOYATZI Stavroula ; KAKOYIANNI-DOA Fryni (2009), Le dictionnaire électronique des adverbes du grec moderne, *Actes de la 30^{e} Rencontre Annuelle du Département de Linguistique*, Faculté des Lettres, Université Aristote de Thessalonique, p. 149-161.
WICHMANN Anne *et al*. (1997), *Teaching and Language Corpora*, London, Longman.

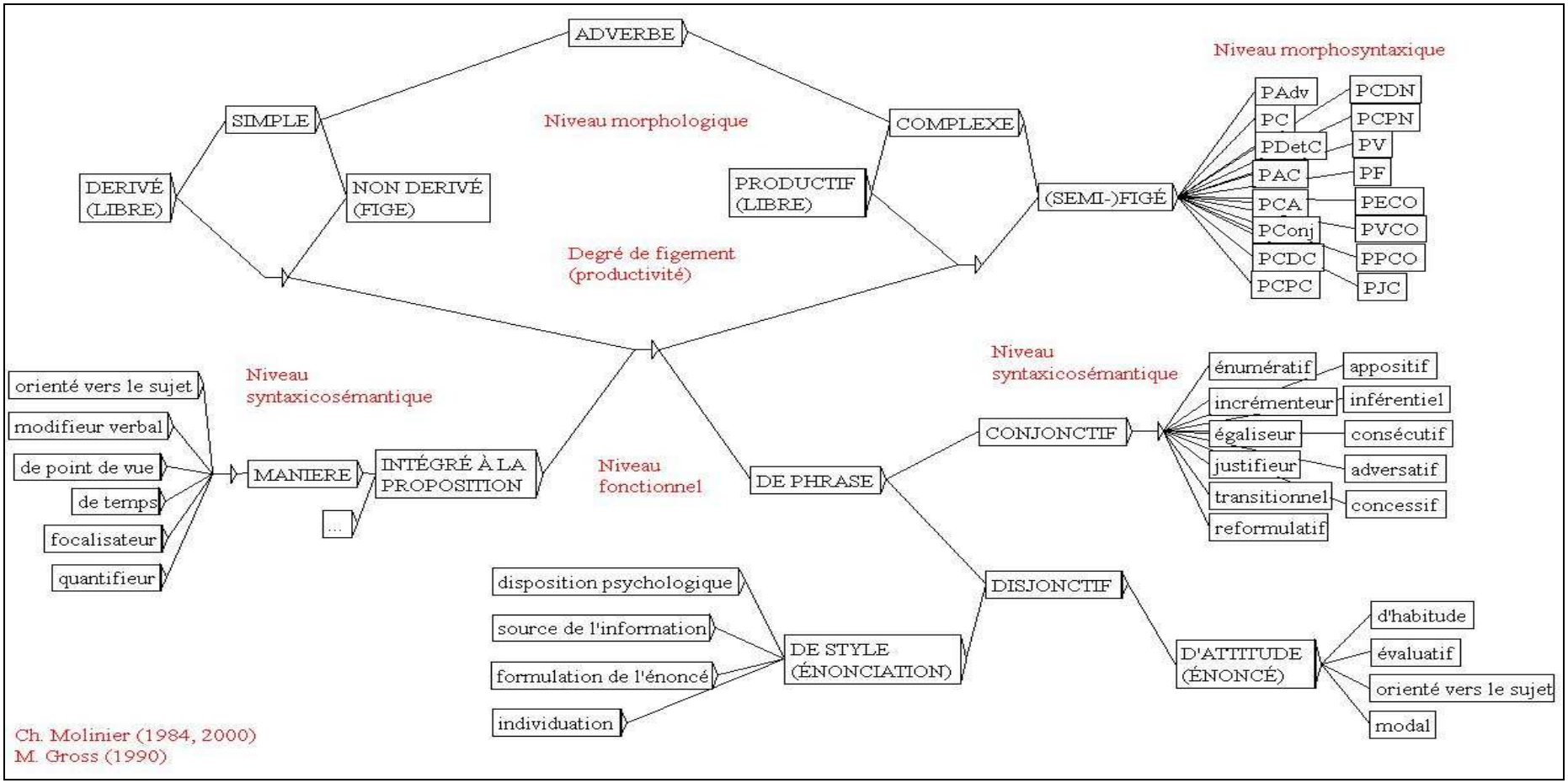


Figure 1. *Classification des adverbiaux français et grecs*

FRANÇAIS		GREC MODERNE	
2.369	1.785 pour moi 584 Pour moi	**4.846**	2.963 κατά τη γνώμη μου 1.687 Κατά τη γνώμη μου 148 κατά την γνώμη μου 48 Κατά την γνώμη μου
1.876	1.459 à mes yeux 406 À mes yeux 11 A mes yeux	**3496**	2.415 προσωπικά 1.081 Προσωπικά
1.690	Selon moi	**1.903**	1.108 κατά την άποψή μου 795 Κατά την άποψή μου
782	Personnellement	**1341**	766 για μένα 208 Για μένα 273 για εμένα 94 Για εμένα
1.107	986 À mon avis 121 A mon avis	**982**	672 κατά τη γνώμη μας 310 Κατά τη γνώμη μας
582	544 À mon sens 38 A mon sens	**474**	295 κατά την άποψή μας 179 Κατά την άποψή μας
270	160 d'après moi 110 D'après moi	**198**	110 κατ' εμέ 73 Κατ' εμέ 11 κατ' εμένα 4 Κατ' εμένα
203	127 de mon point de vue 76 De mon point de vue	**62**	32 κατά την άποψη μου 30 Κατά την άποψη μου
31	23 à mon humble avis 8 À mon humble avis	**52**	26 εμένα προσωπικά 26 μένα προσωπικά
18	15 à mon point de vue 3 À mon point de vue	**39**	20 για μένα προσωπικά 19 για εμένα προσωπικά
6	à mon opinion	**39**	22 κατά τη δική μου άποψη 17 Κατά τη δική μου άποψη
4	Selon mon opinion	**35**	21 κατά την εκτίμησή μου 14 Κατά την εκτίμησή μου
4	à mon sentiment	**27**	14 κατά την προσωπική μου άποψη 13 Κατά την προσωπική μου άποψη
4	selon mon point de vue	**26**	19 για εμένα προσωπικά 7 Για εμένα προσωπικά
4	à ce que je pense	**23**	20 κατά την κρίση μου 3 Κατά την κρίση μου
2	selon mon avis	**13**	8 κατά την αντίληψή μου 5 Κατά την αντίληψή μου
2	selon mon opinion	**12**	10 κατά την ταπεινή μου γνώμη 2 Κατά την ταπεινή μου γνώμη
2	Selon mon point de vue	**9**	7 κατά την προσωπική μου γνώμη 2 Κατά την προσωπική μου γνώμη
2	Si vous me demandez mon avis	**9**	5 Κατά τη δική μου γνώμη 4 κατά τη δική μου γνώμη
		5	3 κατά τη δική μου αντίληψη 2 Κατά τη δική μου αντίληψη
		4	κατά την πεποίθησή μου
		4	2 Κατά τη άποψή μου 2 κατά τη άποψή μου
		3	κατά τη δική μου εκτίμηση
		3	κατά την δική μου άποψη
		2	από ό τι νομίζω
		2	κατά την προσωπική μου εκτίμηση
	Total : 8.958		**Total : 13.609**

Tableau 2. *Fréquence d'emploi des « adverbiaux de conviction personnelle » dans le Corpus Parallèle français-grec*

Co-occurrence des *ADV(Instr)* figés dans les constructions *V ADV(Instr)* libres en polonais

Agnieszka Kaliska

1. Introduction

Les ADV(Instr) figés constituent une sous-classe des adverbes polonais délimités à la base des critères sémantiques et syntaxiques proposés par GROSS (1990) pour l'analyse de *l'adverbe généralisé* en français et adaptés pour l'analyse du polonais (KALISKA, à paraître).

Les ADV(Instr) figés constituent une classe relativement homogène du point de vue morphologique : ils ont la forme d'un N décliné à l'Instrumental (ex. *cichcem* 'en cachette') sans qu'il existe forcément un Nominatif attesté correspondant (respectivement : *°cichc*). Le N base peut donc être un mot possible non existant mais reconstructible selon le paradygme de déclinaison.

Notre étude sera, pour partie, lexicale. Nous présenterons une liste des ADV(Instr) figés que nous avons dressée à la base du dictionnaire électronique de la langue polonaise *Uniwersalny Słownik Języka Polskiego* (*USJP ;* 100 000 entrées lexicales). Les unités regroupées ont été ensuite caractérisées en termes de propriétés de co-occurrence – il s'agit notamment des co-occurrences V ADV(Instr) libres – que nous avons identifiées à la base du Corpus national de la langue polonaise (*Narodowy Korpus Języka Polskiego* ; 1,5 milliards de mots), disponible en ligne : http://nkjp.uni.lodz.pl.

2. Délimitation des *ADV(instr)* figés

Le principal critère de délimitation des ADV en lexique-grammaire est de nature sémantique. À savoir, l'*adverbe* n'est pas un *objet* sélectionné du prédicat, ex.: (*Luc* / *E) *est entré* (*en cachette* / E) (*dans le bureau* / *E). Parallèlement en polonais : (*Luc* / *E) *wszedł* (*cichcem* / E) (*do biura* / *E). Autant les arguments (sujet et objet indirect) sont obligatoires et essentiels pour le sens de la phrase, autant les ADV *en cachette* et *cichcem* fournissent des caractéristiques accessoires et sont donc facultatifs.

Le critère intuitif de non sélection est confirmé en lexique-grammaire par des traits syntaxiques qui ne sont tout de même ni nécessaires ni suffisants (GROSS, 1990 : 20). C'est la conséquence de l'extrême hétérogénité des ADV qui se trouvent regroupés sous l'étiquette *adverbe généralisé* (à côté des ADV proprement dits, il y a des syntagmes et des propositions subordonnées circonstanciels, voire des exclamations et des incises figées).

La distinction entre ADV(Instr) et N objets en polonais se vérifie d'une manière générale par la forme des pronoms interrogatifs. Les ADV(Instr) figés n'acceptent pas la question par *Czym ?* (fr.: *De* / *Avec quoi ?*) bien qu'elle corresponde au cas Instrumental dans les paradygmes réguliers. On peut, par contre, leur faire correspondre la question par *Jak ?* (fr.: *Comment ?*) incompatible avec les objets : (**Czym* / *Jak*) *Luc wszedł do biura ? – Cichcem*[ADV]. En revanche : (*Czym* / **Jak*) *poruszył Luc ? – Głową*[OBJ].

Les ADV(Instr) ayant la forme d'un N à l'Instrumental auquel correspond néanmoins une forme de Nominatif attestée vérifient également cette propriété : (**Czym* / *Jak*) *Luc leży na plaży ? – Plackiem*[ADV]. L'ADV *plackiem* (fr.: 'à plat – en parlant des humains allongés') appartient au même paradygme de déclinaison que le Nominatif *placek* (fr.: 'mince couche de pâte'). Les deux sont liées par une relation sémantique de métaphore qui a permis de construire à partir du N régulier l'ADV figé ayant un sens autre que le N base et des propriétés de co-occurrences différentes.

2.1. Distinction entre *ADV(Instr)* libres, *N* objets et *N* prédicats

S'il n'y a pas de doutes quant à la nature nominale de tout élément qui apparaît sous forme d'Accusatif, de Datif ou de Génitif car seuls les N prédicats et les N objets prennent ces formes dans le discours, il est plus problématique de transiger quant à l'appartenance catégorielle des éléments ayant la forme de l'Instrumental. Ainsi certains ADV(Instr) peuvent être facilement confondus avec la catégorie objet lorsqu'ils sont des N déclinés issus de paradygme de déclinaison de substantifs correspondants (ex. *widelcem* 'à la fourchette'). D'autres expriment la manière d'effectuer une activité exprimée par un V qui, d'un côté, est sémantiquement plein et, de l'autre, pourrait être considéré comme classifieur pour un type sémantique de prédicat, ex. *Max jedzie* (*pociągiem* / *samochodem* / E)[ADV] *do Krakowa* (fr.: 'Max va à Cracovie (en train / en voiture / E') ou *Max płynie* (*motylkiem* / *kraulem* / *żabką* / E)[ADV] (fr.: 'Max nage (le papillon / le crawl / la brasse)') . Les ADV cités sont facultatifs et acceptent la double forme du pronom interrogatif : (*Czym* / *Jak*) *Max* (*je makaron* / *jedzie do Krakowa* / *płynie*) ?.

Du point de vue sémantique, ils sont pourtant très divers. la forme *widelcem* (fr.: 'à la fourchette') a la sémantique *concrète* (c'est un ustensile de table). Les ADV styles de nage ont une référence événementielle car ni le crawl ni le papillon ni la brasse ne sont des objets concrets. Les ADV qui renvoient aux moyens de transport ont une référence concrète car, à chaque fois, la manière qu'ils désignent est en relation avec un objet concret désigné par le même mot. Ils connottent quand même toute une série d'activités connexes lorsqu'ils renvoient aux manières de déplacement. La catégorie d'ADV(Instr) est donc très diversifiée. On y trouve autant des instruments à référence concrète que des moyens sémantiquement moins évidents. Nous considérons ce fait comme

caractéristique dans l'analyse de l'ADV(Instr) et la forme du pronom interrogatif, notamment la possibilité de lui faire correspondre la question par *Jak ?* (fr.: *Comment ?*), est pour nous une propriété de vérification dans la distinction entre ADV(Instr), N prédicats et N objets.

Cependant, si les moyens de transport forment des ADV(Instr) libres et appartiennent à la même sous-catégorie que les ustensiles de table, les ADV(Instr) styles de nage sont figés car, souvent, ils résultent d'une extension sémantique de N base et ont donc perdu les propriétés de co-coccurrence et de sélection d'origine au profit des autres, propres à la sémantique des styles de nage.

3. Construction du corpus des *ADV(Instr)* figés

La première étape dans la construction d'une liste des ADV(Instr) figés consistait à relever les unités qui n'ont pas de forme de Nominatif attestée. De tels ADV se trouvent répertoriés dans l'index *a tergo* que fournit le dictionnaire *USJP.* L'index *a tergo* nous a permis de relever les lexèmes finis en *-em, -ą, -ami* qui sont des terminaisons de l'Instrumental. À cette étape nous n'avons pu relever que les ADV(Instr) figés qui constituent eux-mêmes des entrées de dictionnaire. Les autres, ceux qui sont figés par leurs forme et sens tout en faisant partie des paradygmes de déclinaison des N base existants, n'ont pas d'entrées lexicales distinctes et n'apparaissent pas dans l'index *a tergo*. Cependant, leurs propriétés de co-occurrence sont souvent tellement limitées que le dictionnaire les considère comme constituants des expressions figées ou, pour être plus exact, semi-figées dont il dresse une liste facile à réviser. C'est à partir de cette liste que nous avons relevé les constructions V C(Instr) qui, paradoxalement, correspondent aux critères de délimitation des ADV figés de phrases libres. Notamment, les ADV figés des phrases libres « *affectent des phrases quelconques d'une manière accessoire pour le sens* » et, « [d]*ans la quasi-totalité des cas, ils sont facultatifs* » (GROSS, 1986 : 72).

Beaucoup d'expressions ont dû être éliminées de notre corpus. Par exemple, bien que les unités *wzrokiem* (fr.: 'd'un regard') et *ramieniem* (fr.: 'd'un bras') aient une construction adverbiale, leur omission redonne au V son sens primaire. Ainsi, ni l'activité de *zmierzyć wzrokiem* (fr.: 'regarder quelqu'un avec attention pour le juger') n'inclut celle de *zmierzyć* (fr.: 'mesurer'), ni l'activité de *otoczyć ramieniem* (fr.: 'soutenir quelqu'un') n'inclut celle de *otoczyć* (fr.: 'entourer'). Les deux co-occurrences sont donc figées et, par conséquent, sémantiquement opaques. Les conséquences de figement se vérifient à travers d'autres propriétés : l'expression *zmierzyć wzrokiem* ne pourrait être considérée comme cohyponyme de *zmierzyć linijką* (fr.: 'mesurer à la règle') qui inclut le sens de déterminer une grandeur.

D'autres formes qui n'ont pas trouvé leur place dans le tableau, bien qu'elles soient des ADV(Instr), ont des caractéristiques des *objets internes* et sont toujours accompagnés d'un modifieur adjectival, ex. *iść* (*szybkim* / *wolnym* /...) *krokiem* (fr.: 'marcher d'un pas (lent / rapide /...)'). De tels ADV constituent une classe ouverte dans la mesure où l'on peut toujours, du moins en théorie, décrire la façon d'effectuer une action à l'aide d'une construction à objet interne modifié par un adjectif qui convient.

4. Corpus d'analyse : *ADV(Instr)* figés en polonais

Ci-dessous nous présentons le corpus des ADV(Instr) figés que nous avons consitué à partir des données du dictionnaire *USJP* selon les critères expliqués ci-haut. Les ADV(Instr) figés constituent une classe relativement peu nombreuse et plutôt homogène. Les formes issues de N masculins constituent une majorité indiscutable.

La colonne 2. comporte une liste de co-occurrences verbales identifiées à la base d'un échantillon de 50 000 segments du sub-corpus équilibré, complétées ensuite par des co-occurrences que fournit le dictionnaire *USJP* (elles sont séparés d'un point-virgule). Les chiffres indiquent le nombre de co-occurrences relevées sur corpus.

Certaines co-occurrences V C(Instr) fournies par le corpus n'ont pas été prises en compte lorsque nous avons pu juger par intuition ou d'après le contexte que le C(Instr) n'était pas l'ADV mais il était un objet ou un N prédicat (ex. le N attribut dans les constructions avec le V *być* (fr.: 'être')).

La colonne 3. fournit l'information sur un type sémantique des V de co-occurrence si seulement les données permettent d'en déterminer un. Le type sémantique tel que [déplacement] en parlant des V de déplacement peut correspondre aux *propriétés de sélection* d'un ADV ; autrement dit, un ADV se combinera avec des V de déplacement plutôt qu'avec d'autres types de verbes.

ADV(Instr) figé	V de co-occurrence	Type sémantique du V
boczkiem	- ; wymknąć się, zerkać, siedzieć	déplacement en parlant des humains autres activités
bykiem	- ; patrzeć	variantes de *regarder*
martwym bykiem	- ; leżeć	leżeć (fr.: ‘être allongé’)
chyłkiem	przemykać 54 wymykać się 24 pomykać 6 wymknąć się 17 przemknąć 8 wysunąć się 8 wycofać się 16 wycofywać się 7 ruszyć 10 opuszczać 7 uciekać 11 opuścić 11 iść 13 wracać 6 uciec 5 wrócić 7 udać 8 jeść 8 ; uciekać, wymykać się, przemykać, przekradać się, załatwić, wycofać się	déplacement en parlant des humains autres activités
cichaczem	wymknąć się 11 wynieść 9 ; wymknąć się, zakraść się, zaglądać, wyjechać	déplacement en parlant des humains activités diverses
cichcem	wymknąć się 6 wynieść 6 wyjść 5 robić 5 ; uciec, przemknąć, handlować, popijać	déplacement en parlant des humains activités diverses
ciszkiem	- ; wymknąć się, zakraść się, zaglądać, wyjechać	activités diverses
ciurkiem	lać 12 spływać 11 lecieć 14 płynąć 7 ; płynąć, spływać, lać się, kapać, oglądać	« traversement » en parlant des liquides activités diverses
delfinem	pływać 6 ; pływać	déplacement sur l’eau en parlant des humains
duszkiem	wypić 98 wychylić 23 wypijać 22 wychylać 8 pić 9 ; wychylić	absorption de liquides
falą	ruszyć 7 uderzyć 5 iść 5 ; iść, płynąć	déplacement
galopem	pomknąć 8 lecieć 10 ruszyć 26 gnać 5 pędzić 7 puścić się 7 wracać 6 ; pędzić, załatwić, zrobić	déplacement activités diverses
głową	- ; pracować	*pracować* (fr.: ‘travailler’)
krzyżem	upadać 12 leżeć 59 paść 13 kłaść się 7 położyć się 6 ; leżeć	variantes d’*être allongé*
mimochodem	napomknąć 9 wspomnieć 36 rzucić 6 dorzucić 24 wspominać 24 zauważyć 20 rzucać 8 spytać 8 zapytać 7 dodać 6 powiedzieć 19 pisać 5 ; spojrzeć, wspomnieć, zapytać	parole variantes de *regarder*
motylkiem	- ; płynąć	déplacement sur l’eau en parlant des humains
okrakiem	siąść 9 usiąść 23 siedzieć 47 siadać 11 stanąć 8 stać 16 ; siąść, zjeżdżać, stać	variantes de *s’assoir* variantes d’*être debout*
pieskiem	- ; -	déplacement sur l’eau en parlant des humains
plackiem	leżeć 35 paść 12 padać 11 ; padać, leżeć	variantes d’*être allongé*
półgębkiem	uśmiechnąć się 15 uśmiechać się 12 wspominać 6 odpowiadać 8 mówić 14 ; bąkać, mówić, odpowiadać, jeść, uśmiechać, usiąść	parole autres activités : *sourire*, *manger*, *s’asseoir*

ADV(Instr) figé	V de co-occurrence	Type sémantique du V
półgłosem	nucić 14 naradzać się 6 zanucić 5 rzec 46 odezwać się 18 spytać 22 zagadnąć 5 powiedzieć 102 mruknąć 9 zapytać 39 zawołać 13 rzucić 5 recytować 5 rozmawiać 36 szepnąć 5 odczytać 5 odczytywać 5 powtarzać 21 wymieniać 6 mówić 87 wyznać 5 czytać 21 odmawiać 5 przeczytać 6 powtórzyć 5 modlić się 6 wyjaśnić 6 odeprzeć 5 śpiewać 5 zauważyć 7 odpowiedzieć 5 prowadzić dyskusje / rozmowy 5 opowiadać 5 ; zanucić, mówić, rozmawiać, opowiadać	parole
półuchem	- ; słuchać, wysłuchać	variantes d'*écouter*
rakiem	wycofywać się 9 wycofać się 6 ; pełzać, chodzić, cofać się, wycofać się	déplacement
społem	- ; zasiąść, udać się	activités diverses
strzałką	- ; USJP płynąć	déplacement sur l'eau en parlant des humains
jednym susem	przeskoczyć 13 dopaść 16 przesadzić 8 wskoczyć 10 skoczyć 13 wyskoczyć 12 znaleźć się 18 być 12 ; być, znaleźć się	déplacement
jednym tchem	wyrecytować 7 wyrzucić 27 wymieniać 57 wyliczać 20 wymienić 20 przeczytać 26 wypowiedzieć 14 czytać 14 opowiadać 5 mówić 18 powiedzieć 6 ; -	parole
tranzytem	przejeżdżać 6 przechodzić 5 jechać 5 ; jechać	déplacement
ukosem	przecinać 7 ; patrzeć, zerkać	déplacement variantes de *regarder*
ukradkiem	ocierać się 27 zerkać 19 przeżegnać 11 zerknąć 16 spoglądać 35 rozejrzeć 14 spojrzeć 30 wycierać 9 otrzeć 9, wytrzeć 6 obserwować 25 rozglądać 8 wsunąć 5 patrzyć 26 rzucać 7 śledzić 6 żegnać 5 schować 6 przeglądać 5 przyglądać się 7 chować 5 zaglądać 5 obejrzeć 5 patrzeć 10 sięgnąć 5 spotykać się 8 czytać 15 przychodzić 8 jeść 22 chodzić 7 wychodzić 6 oglądać 5 robić 11 zrobić 10 wyjść 5 dawać 6 widzieć 5 mówić 10 ; przemknąć się, spotykać się, spoglądać, zerkać	activités diverses
wierzchem	jeździć 10 jechać 10 ; jechać, jeździć	déplacement au cheval
wilkiem	patrzyć 21 patrzeć 17 ; spojrzeć, patrzeć	variantes de *regarder*
żabką	pływać 9 skakać 8 ; pływać	déplacement sur l'eau en parlant des humains
pełnym żaglem	- ; płynąć, iść	déplacement sur l'eau en parlant des navires

7. Conclusion

L'intérêt de la présente étude était de fournir une liste de co-occurrences verbales pour l'analyse des ADV(Instr) figés en polonais pour voir si'peut, d'après les données du corpus, identifier un type sémantique des V avec lesquels se combinent les ADV(Instr) figés. Dans la plupart des cas, une telle identification était possible. C'est ainsi que nous avons pu identifier, par exemple, des V de déplacement, des V de parole, etc., même si, d'une façon générale, cette *identification* n'était qu'une simple confirmation de l'intuition qu'a probablement le locuteur de la langue polonaise au sujet des co-occurrences V ADV(Instr). D'autre part, la liste des co-occurrences que nous avons dressée dépasse largement le nombre de co-occurrences que cite le disctionnaire et, pour cette raison, constitue une étape importante dans la construction du *lexique-grammaire* des ADV polonais.

Références bibliographiques

DUBISZ Stanisław (2004). *Uniwersalny słownik języka polskiego* (version électronique 1.0). Wyd. Naukowe PWN.
DE GIOIA Michele (2001). Avverbi idiomatici dell'italiano: analisi lessico-grammaticale. L'Harmattan Italia.

GROSS Maurice (1986). Lexique-grammaire et ADVs : deux exemples, *Revue québécoise de linguistique* vol. 15, n° 2, 1986, p. 299-310.
GROSS Maurice (1990). *Syntaxe de l'adverbe*. Paris, Asstril.
KALISKA Agnieszka K. (à paraître) Introduction à la syntaxe de l'*adverbe généralisé* en polonais.
KALISKA Agnieszka K. (à paraître). Tests morphosyntaxiques pour la délimitation de l'*adverbe généralisé* en polonais et en français.
PRZEPIÓRKOWSKI Adam, Bańko Mirosław, Górski Rafał, L., Lewandowska-Tomaszczyk Barbara (eds.) (2012). *Narodowy Korpus Języka Polskiego*. Warszawa: Wydawnictwo Naukowe PWN.

Les adverbiaux : de la phrase au discours

Michel Charolles, Béatrice Lamiroy

Les adverbiaux étant, par définition, peu soudés à leur phrase d'accueil, on peut s'attendre à ce qu'ils assument des fonctions à l'échelle du discours. C'est de fait ce qui se passe avec les adverbiaux connecteurs (Molinier et Lévrier 2000, Guimier 1996, Bonami et al. 2003) qui signalent des relations rhétoriques (Mann & Thompson 1986, 1988) ou de discours (Asher & Lascarides 2003) essentielles pour l'interprétation des liens sémantiques et pragmatiques entre les phrases/énoncés.

S'il est couramment admis que les connecteurs contribuent (avec les anaphores) à la cohésion et à la cohérence des discours, il n'en va pas de même avec certains adverbiaux dits cadratifs apparaissant en tête de phrase ou en zone préverbale (Charolles et Prévost 2003, Charolles et Péry-Woodley 2007, Vigier et Terran 2005). Ces adverbiaux jouent aussi, comme nous allons le rappeler dans la première partie, un rôle dans l'organisation des informations textuelles. Ils tirent ce pouvoir de leur capacité à indexer ou porter sur le contenu de phrases faisant suite à celle dans laquelle ils apparaissent. Les cadratifs instaurent des liens d'une autre nature que ceux signalés par les connecteurs, même si, comme nous le montrerons dans la partie suivante, la frontière entre cadratifs et connecteurs n'est pas étanche. Le fait que beaucoup de connecteurs soient d'anciens adverbiaux suggère en effet qu'il y a un continuum entre les adverbiaux cadratifs et les adverbiaux connecteurs, quoiqu'ils signalent des relations de direction opposée : descendantes pour les cadratifs et remontantes pour les connecteurs. Beaucoup de points restent à préciser sur ces phénomènes, notamment sur les aspects diachroniques, mais ici nous chercherons à comprendre (i) quels facteurs peuvent expliquer que certains adverbiaux jouissent d'un potentiel cadratif et (ii) comment il se fait que certains cadratifs se rapprochent des connecteurs.

1. Les adverbiaux cadratifs

1.1. Fonction des adverbiaux cadratifs

Les adverbiaux spatiaux et temporels en tête de phrase sont très souvent utilisés pour structurer les informations textuelles. Ces adverbiaux dits scéniques (Riegel et al. 2011) ou cadratifs sont d'un emploi très courant dans les articles scientifiques ou techniques où ils vont souvent de pair avec les alinéas ou d'autres marques typo-dispositionnelles comme c'est le cas dans (1) : [1]

> (1) Les départements où l'emploi tertiaire progresse fortement sont aussi bien souvent des départements où l'emploi industriel se maintient ou croît : Alpes, Midi méditerranéen, Bretagne et sud du Bassin Parisien. C'est surtout dans les départements de la grande banlieue parisienne que le dynamisme du tertiaire ne s'appuie pas sur des évolutions industrielles particulièrement favorables.
>
> **[Dans le sud-est et la grande banlieue parisienne**, départements déjà fortement tertiarisés, c'est l'ensemble du tertiaire qui évolue plus rapidement qu'au niveau national. La progression des services marchands y est inhabituellement forte**]$_{Csp1}$**.
>
> **[Dans l'ouest et le sud du Bassin Parisien**, les services non marchands ne jouent qu'un rôle négligeable. La croissance du tertiaire repose d'abord sur le dynamisme des services aux entreprises, et dans une moindre mesure sur celui des transports-télécommunications et des autres services marchands. La liaison entre tertiaire et industriel est donc plus forte que dans le sud-est**]$_{Csp2}$**.

Les fragments de texte structurés de cette manière n'apparaissent pas dans n'importe quel contexte : les adverbiaux détachés en tête de phrase exploitent, le plus souvent, une dimension des situations qui est amorcée dans le discours précédent. C'est ce qui se passe dans (1) où le rédacteur, après avoir asserté que les départements français se différencient en fonction du développement de secteurs économiques divergents (tertiaire *versus* industriel), illustre ensuite son propos en faisant état de données sur des régions qui présentent des caractéristiques divergentes de ce point de vue. L'espace sert de critère pour la répartition des informations textuelles, et cette dimension étant annoncée dans le paragraphe précédent, les deux SP adverbiaux (*dans le sud-est et la grande banlieue parisienne*, et *dans l'ouest et le sud du Bassin Parisien*) sont des sortes de topiques (en l'occurrence scéniques), quoique le topique de discours principal reste la répartition entre les secteurs économiques. Le fait que ces deux adverbiaux soient à même d'indexer d'autres situations que celle dénotée par la phrase en tête de laquelle ils figurent est tout à fait remarquable d'un point de vue l'inguistique. C'est ce pouvoir, relevé par Thompson (1985)[2] dans un article princeps sur les infinitives de but antéposées en anglais, qui explique qu'ils puissent être utilisés, à l'échelle du discours, pour structurer les informations textuelles.

[1] Extrait de Mabile S. & Jayet H. La redistribution géographique des emplois entre 1975 et 1982. In: *Economie et statistique*, N°182, Novembre 1985. pp. 23-35.

[2] Pour une discussion détaillée, cf. Charolles & Lamiroy (2002)

Le potentiel cadratif des adverbiaux antéposés n'est exploité, en général, que dans des fragments de discours qui s'y prêtent, où une dimension des situations (en général amorcée) peut servir de critère pour la répartition des informations arrivantes. Ce mode d'organisation des informations vient se surajouter aux relations référentielles signalées par les anaphores et aux relations de discours indiquées (ou non) par des connecteurs qui peuvent apparaître aussi bien entre des phrases faisant partie d'un même cadre ou articuler des cadres. On le voit bien dans (1) où l'on pourrait rajouter un connecteur contrastif comme *par contre* ou *à l'inverse* au début du troisième paragraphe, avant (ou après) *dans l'ouest et le sud du Bassin Parisien*.

Si les anaphores, les connecteurs et les cadratifs contribuent à la cohésion-cohérence des discours et à leur structuration, ils se différencient par le niveau auquel ils interviennent et par le type de relations qu'ils induisent. Les cadratifs, comme on vient de le relever, jouent sur la structure informationnelle, permettant de la mettre sous une forme qui soit la plus coopérative possible. Les cadratifs se différencient aussi des anaphores et des connecteurs par le fait qu'ils induisent des relations descendantes (orientées vers la suite du discours), alors que les connecteurs et les anaphores impliquent fondamentalement des liens avec le discours précédent (cf. Charolles 2005). Ce point mérite d'être souligné car les taxinomies de marques de cohésion du discours se limitent en général aux connecteurs et aux anaphores (cf. Reinhart 1980, Sanders et Spooren 2001). Le fait que les langues offrent toute une gamme d'expressions dont la fonction spécifique est de coder grammaticalement des relations avec le discours qui précéde plutôt qu'avec le discours qui suit n'a rien de surprenant quand on pense aux lecteurs/auditeurs qui doivent traiter (à toutes sortes de niveaux) et en quelques millisecondes ce qu'ils lisent ou entendent. On comprend que le plus urgent consiste à rattacher ce qui arrive avec ce qui a déjà été traité. Raison de plus pour s'intéresser comme nous le faisons ici, aux expressions qui, comme les cadratifs, pèsent sur le traitement du discours arrivant.

1.2. Affinité des adverbiaux cadratifs avec l'initiale de phrase

Avec les cadratifs, il semble bien que la zone préverbale joue un rôle crucial, comme d'ailleurs avec les connecteurs. Il ne s'agit pas ici de nier que les adverbiaux insérés ou en fin de phrase puissent aussi influencer l'interprétation des phrases suivantes. Crompton (2006) relève ce point à propos d'exemples comme (2) à (5) où l'on comprend que la seconde phrase fait allusion à une action accomplie au sommet, et cela quelle que soit la position (2, 3 et 5) du SP temporel, et même quand il est intégré syntaxiquement (1 et 4) :

(2) [i] On the fifth day they were at the summit. [ii] They took photographs.

(3) [i] They reached the summit on the fifth day. [ii] They took photographs.

(4) [i] The fifth day found them at the summit. [ii] They took photographs.

(5) [i] The summit was reached on the fifth day. [ii] They took photographs.

Le principe de pertinence (Sperber et Wilson 1986) suffit à expliquer les effets de ce type, mais toute la question est de savoir si ces effets ne s'imposent pas plus quand les SP sont détachés en position initiale.

Les données sur corpus présentées par Crompton sur l'anglais n'étayent pas cette hypothèse, mais elles ne portent que sur un nombre restreint d'emplois de SP tirés de divers extraits de textes qui ne se prêtent guère à l'exploitation de leur potentiel cadratif. Par ailleurs, ces données sont contredites par Charolles (2006) dans une étude sur 100 emplois dits narratifs du SN adverbial *un jour*. Pour chaque emploi ont été annotés des traits morphologiques (*un jour* modifié ou non), positionnels (*un jour* préposé, inséré ou postposé), aspectuels et temporels (*un jour* dans une phrase perfective précédée d'une phrase imperfective ou non), l'extension de la portée extraphrastique de *un jour* (étroite, moyenne ou large), la présence d'indices de clôture et leur fiabilité. L'étude fait ressortir deux grands types d'usage en discours qui s'opposent terme à terme. Dans le premier type, qui correspond aux emplois cadratifs, *un jour* modifié figure en tête de paragraphe ou de phrase, s' inscrit dans un contexte aspecto-temporel prototypique (phrase d'accueil perfective et phrase précédente imperfective), a une portée extraphrastique large, le cadre étant fermé par des indices fiables. Dans le second type, *un jour*, est non modifié, inséré, le contexte aspecto-temporel n'est pas protypique, la portée est étroite et les indices de fin de portée, quand il y en a, ne sont pas fiables. Ces deux grands types de configurations ne représentent respectivement que 12% et 13% des emplois étudiés, les autres configurations s'échelonnant entre les deux. Les pourcentages d'emploi clairement cadratifs sont assez peu nombreux, mais ils suggèrent que les cadratifs ont une prédilection pour la position initiale. Ces données sont confirmées par ailleurs par des données psycholinguistiques tirées de deux expérimentations (avec l'autoprésentation segmentée) montrant que les lecteurs ne traitent pas de la même façon des SP spatiaux antéposés et postposés (Argenti et Charolles 2011, Colonna, Sarda, Pynte et Charolles soumis).

1.3. Types de cadratifs

1.3.1. Propriétés définitoires

Les adverbiaux cadratifs ne sont pas définis par leur appartenance à une partie du discours : ils peuvent être:

- des SP spatiaux et temporels (cf. exemple (1)) ou autres, comme *selon X, grâce à X, avec X*
- des SN: *un dimanche, un matin, une fois*

- des ADV: *soudainement, parallèlement*
- des subordonnées à verbe conjugué : *Quand Paul viendra, … Si Paul venait, …*
- des subordonnées infinitives: *A voir le nombre de X, ...*
- des expressions (semi-)figées: *en un mot, tout bien considéré, autrement dit, tout compte fait, à propos de, en matière de, au sujet de X , côté X, quant à X*
- des corrélatifs: *d'un côté / de l'autre, d'une part / d'autre part, premièrement / deuxièmenent*

Les cadratifs sont définis par trois critères :

- syntaxique : leur fonction d'adjoint et leur position en tête de phrase
- prosodique : leur prosodie incidente (indiquée par la virgule à l'écrit)
- sémantique : leur indépendance avec leur proposition d'accueil.

Ces critères permettent d'exclure :

- les compléments antéposés, ex. *De sa sœur, Paul n'avait plus aucune nouvelle*
- les adverbiaux (non détachés) dans les inversions locatives : *Du plafond pendaient des guirlandes*
- les adverbiaux évaluatifs du genre de *heureusement/malheureusement*, ex. *Heureusement, Marie n'était pas là* ou *Par chance / par hasard, Paul était là*
- les adverbiaux modaux, ex. *Peut-être objectera-t-il que nous aurions dû* ou *Probablement, Paul a oublié ses clés*
- les adverbiaux liés au sujet comme :
  - certains adverbes en *-ment* (Molinier et Gévrier 2000: 117-147): *Intelligemment, Paul n'a rien dit*
  - les gérondifs : *En sortant de chez lui, Paul …*
  - les constructions absolues : *Le chapeau de travers, la marquise …*

La frontière entre les adverbiaux et les constituants intégrés à la phrase étant une affaire de degré, on retiendra le principe que plus un ajout est indépendant par rapport à sa phrase d'accueil, plus il sera susceptible de fonctionner comme un cadratif. Ainsi, dans *Sur le pont d'Avignon, on danse beaucoup*, le complément de lieu est *a priori* mieux à même de servir de cadre que dans la phrase à dislocation avec reprise anaphorique : *Sur le pont d'Avignon, on y danse*. De même, une subordonnée antéposée comme : *Si Paul gagne au loto, je me fais moine* jouirait d'un potentiel cadratif supérieur à celui qu'elle aurait dans *Si Paul gagne au loto, il s'achètera une grosse voiture*. Autrement dit, pour qu'un élément puisse assumer une fonction discursive, en l'occurrence cadrative, il doit être syntaxiquement le plus autonome possible par rapport au prédicat de sa phrase d'accueil. Au point d'ailleurs que les adverbiaux cadratifs peuvent constituer à eux seuls une phrase graphique, procédé largement exploité dans la presse :

> (6)"**Printemps 1927.** Le matelot mécanicien Jack Holeman vient d'être affecté sur le San Pablo, une vieille canonnière dont la mission est d'impressionner les Chinois … " (Début de la présentation du film La canonnière du Yang-Tsé, magazine TV)

1.3.2. Typologie

Les SP spatiaux et temporels se prêtent *par excellence* aux emplois cadratifs : toutes les situations étant ancrées dans le temps et dans l'espace, il est toujours possible (même quand il n'y a pas de SP explicite, cf. Erteshik-Shir 2007) de les repérer par rapport à ces critères. L'idée de repère étant associée à celle de point de référence, les SP spatiaux jouissent en la matière d'un avantage remarquable. Ils prennent très facilement une valeur temporelle (*A Vienne, nous allions au concert chaque semaine*), l'inverse étant beaucoup plus difficile.

Des emplois spatiaux, on passe sans difficultés aux mondes représentés (*Dans ce film / le roman de Luc*), aux institutions ou organisations attachées à des lieux (*A l'ONU, on vote à l'unanimité*) puis à des formes de repérage encore plus abstraits, permettant de distinguer une espèce (*Chez les fourmis, …*), un individu particulier (*Chez Mozart, ….*) ou encore un domaine de connaissances (*En botanique, … / En tagalog, …*) parmi d'autres.

De ces emplois, fondés sur une métaphore spatiale qui reste relativement sensible à l'intuition, on passe à des usages où le repérage porte sur la source de l'énonciation. Il en va ainsi avec des SP comme *selon X, suivant X, d'après X* qui, étymologiquement, sont tous d'origine spatiale ou temporelle. Les SP médiatifs/évidentiels (Aikenvald 2004) qui indiquent la source des informations rapportées, sont souvent exploités pour structurer les textes, le fait de savoir qui se trouve à l'origine de telle ou telle information ayant des conséquences importantes sur le crédit que l'on peut accorder à celles-ci. Ces emplois sont légion notamment dans les travaux académiques où les auteurs sont amenés à rappeler les analyses défendues par des auteurs :

> (7) "**Selon d'Alembert**, la notion de civilisation doit être envisagée dans sa dimension essentiellement "scientifique" alors que **pour Rousseau** elle doit être évaluée par rapport au politique et à la "vertu". Par ailleurs, **pour d'Alembert**, le "progrès" est toujours menacé par une résurgence de la "barbarie", et sa conception évoquant la "fragilité de la civilisation" (p. 293) infirme l'idée qu'il y aurait déjà eu comme une "religion du progrès." (Les études philosophiques, 1981, p.366)

Parmi les adverbiaux potentiellement cadratifs portant sur l'énonciation, il faut également signaler les SP indiquant quel est le thème du discours, avec des introducteurs lexicalisés, de sens spatial, ex. *du côté de X, côté X…* (8), ou non, ex. *en matière de X, à propos de X, concernant X, quant à X, …*(9) :

(8) "Amours et aventures dirigeaient ma vie depuis toujours. Ça allait continuer. C'est exactement ce qui s'est passé. **Côté amour**, mon beau tatoueur assurait la permanence…." (E. Hanska, Les amants foudroyés)

(9) "Je n'ai plus le feu sacré.

- Tiens, **à propos de feu**, tu sais depuis quelque temps, sur la place du marché, il y a un cracheur de feu qui vient faire son numéro..." (E. Hanska, Les amants foudroyés)

D'autres adverbiaux servent exclusivement à organiser la forme du discours. On trouve dans cette dernière catégorie des SP dont le complément nominal qualifie ou catégorise (10) les segments de discours suivants, un grand nombre d'adverbiaux corrélatifs (11) et les ordinaux sériels (12) :

(10) « **En résumé,** de quelque façon qu'on se retourne. il est impossible de découvrir à l'empirisme géométrique un sens raisonnable. »(Poincaré, La science et l'hypothèse)

(11) « **D'un côté** il était un type important, le seul témoin de la tuerie, et on le ménageait, on le questionnait, on lui demandait son nom. **D'un autre** il n'était qu'un vieux tas de fringues récalcitrant et on le secouait, on le menaçait. (Fred Vargas Le marchand d'éponges 2010)

(12) Quelques raisons peuvent être avancées même si aucune d'entre elles n'est suffisante cependant. **Premièrement**, et comme Boutan le faisait remarquer lui-même, le fait qu'il ait travaillé avec un seul spécimen lui interdisait de généraliser ses résultats. (…). **Deuxièmement,** il est important de rappeler que Boutan était isolé dans son travail sur l'intelligence animale et que ses études s'inspiraient avant tout de travaux étrangers. (Hallais La philosophie d'après Corpus Philo Centre Canguilhem)

Parmi les adverbiaux mentionnés ci-dessus, certains comme *en X, chez X* sont syntaxiquement et sémantiquement intégrables à leur phrase d'accueil (*C'est en 1930 / en Italie que … C'est chez les fourmis que …/ C'est dans le film de Luc que…)*, tandis que d'autres le sont moins facilement (?? *C'est selon X que*) et d'autres pas du tout (* *C'est quant à X que*). Les SP énonciatifs qui renvoient à la façon dont le rédacteur ou le locuteur agence son discours et qui sont aussi souvent les plus lexicalisés sont les moins intégrables, avec des cas intermédiaires comme *à propos de X :*

(13) a. A propos de Marie, est-ce que tu sais si elle va venir ?

b. On m'a dit à propos de Marie qu'elle était malade

c. ? C'est à propos de Marie qu'on m'a dit qu'elle était malade

Sur l'ensemble des adverbiaux susceptibles d'assumer une fonction cadrative, il y a donc une gradation allant des plus intégrables aux moins intégrables, avec à une extrêmité les spatiaux et les temporels et, à l'autre, les organisateurs métadiscursifs :

<table>
<tr><td colspan="7">Adverbiaux cadratifs</td></tr>
<tr><td colspan="3">Enoncé</td><td></td><td colspan="3">Enonciation</td></tr>
<tr><td colspan="2">Scéniques</td><td>Représentations</td><td>Domaines Abstraits</td><td>Source</td><td>Thème de discours</td><td>Organisateurs métadiscursifs</td></tr>
<tr><td>Spatial</td><td>Temporel</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>En Allemagne,…</td><td>Au XVIII^e siècle, …</td><td>Dans le film de X, …</td><td>En botanique, …</td><td>Selon X, …</td><td>Quant à X, …</td><td>En résumé, …</td></tr>
<tr><td colspan="7">+ Intégrables ⟷ - Intégrables</td></tr>
</table>

2. Cadratifs et connecteurs

2.1. Des cadratifs aux connecteurs

Certains adverbiaux se prêtent à un processus de grammaticalisation qui les amène à fonctionner en tant que connecteurs : ainsi, *autrement* ou *simplement*, des adverbes de manière au départ, sont devenus des connecteurs indiquant respectivement l'hypothèse négative et l'opposition (Lamiroy et Charolles 2005, Charolles et Lamiroy 2007) :

(14a) Tu aurais dû te comporter **autrement**

(14b) Fais-le, **autrement** tu auras des ennuis = 'si tu ne le fais pas, tu auras des ennuis'

(15a) Elle était habillée très **simplement**

(15b) J'irais bien au cinéma. **Simplement** je n'ai pas le temps = 'mais je n'ai pas le temps'.

Il en va de même de l'adverbe *ailleurs*, au départ un adverbe déictique (dans les emplois en situation) et anaphorique (dans les emplois textuels) de sens spatial qui se prête à des emplois intégrés (*Paul habite ailleurs*), ou comme ajout cadratif (*Ailleurs, Paul a appris que …*) auquel cas il peut parfaitement enbcadrer plusieurs phrases :

(16) "Cet été-là, François, 11 ans, croit à l'imminence de la fin du monde. Il pressent les orages, devine les chagrins, tente de les conjurer par une prière étrange sur une plage d'Afrique du Nord.(…). **Ailleurs**, sur une route, un homme et une femme roulent. La radio annonce de mauvaises nouvelles : attentats de terroristes et de contre-terroristes... " (Présentation du film L'été de tous les chagrins, magazine TV)

Précédé de *par*, *ailleurs* donne *par ailleurs* qui accepte des emplois comme complément d'un verbe:

(17a) Paul pensait que la fuite devait venir du robinet mais l'eau s'échappait **par ailleurs.**

Par ailleurs est cependant le plus souvent employé pour signaler un changement de thème de discours. Si les phrases reliées par *par ailleurs* partagent une même orientation argumentative, la locution signale que les arguments avancés dans la suite ne relèvent pas du même thème de discours que ceux qui précèdent, par ex.

(17b) Paul ne pourra pas venir. **Par ailleurs**, j'ai un article à finir, il vaudrait donc mieux repousser le repas.

Mais il se prête aussi à des emplois où il ne signale qu'un changement de thème de discours (17c) et même, dans des textes plus anciens, à des usages dans lesquels il garde un sens spatial et indique un changement de lieu (17d) :

(17c) « (Daniel …) Son rêve : faire la couverture du Monde du muscle. Pour réussir, Daniel se dope. **Par ailleurs**, Paul est amoureux de Myriam, une jeune femme originaire de Lens et supportrice du club de football de la ville... »

(17d) « Mais il est remarquable que dans tous ces lieux « défrichés », où sur un sol propice prospérèrent les cultures, maintes ruines de l'époque gallo-romaine témoignent que ces prétendus « gains » de la campagne cultivée ne lui ont été que des restitutions.

Par ailleurs pourtant, il est des pays dont les défrichements médiévaux transformèrent totalement les aspects. » (Roupnel Histoire de la campagne française, 1932)

D'ailleurs se prête plus facilement que *par ailleurs* à des emplois intégrés avec un sens spatial (18a), mais il est n'est plus employé dans le français actuel que comme connecteur pour introduire un argument destiné à justifier une assertion précédente (18b) :

(18a) Le ciel était très sombre à l'ouest mais le bruit du tonnerre venait **d'ailleurs**.

(18b) Paul et Robert se détestent. **D'ailleurs** ils ne se disent même plus bonjour.

Parmi les connecteurs justificatifs (*car, parce que, puisque, …*), *d'ailleurs* occupe une place singulière en ce qu'il présente le contenu de la phrase au sein de laquelle il apparaît comme relevant d'un autre ordre d'idées, et comme apportant après coup un argument supposé décisif et de nature à clôre définitivement la discussion (Ducrot 1980).

L'observation des emplois comme adverbiaux auxquels se prêtent *ailleurs*, *par ailleurs* et *d'ailleurs* ou *parallèlement* (Sarda et Charolles 2011) en français contemporain montre qu'ils ne sont pas tous aussi avancés sur l'échelle de grammaticalisation qui mène aux connecteurs :

- *Parallèlement*, qui est le seul à pouvoir régir des compléments prépositionnels (*parallèlement à SN*) accepte très facilement des emplois absolus (spatiaux et surtout temporels) et va jusqu'à des emplois (peu nombreux) comme cadratif de l'énonciation (*parallèlement à ce que je viens de dire*) où il annonce un glissement thématique et marque une relation d'analogie.
- A l'opposé, *d'ailleurs*, lorsqu'il est employé comme ajout, ce qui est le plus souvent le cas, est devenu un connecteur : il sert essentiellement à marquer une relation sémantique et pragmatique de justification entre la phrase dans laquelle il apparaît et une ou plusieurs phrases précédentes.
- *Par ailleurs* occupe une position intermédiaire entre *ailleurs* et *d'ailleurs*. Il signale, comme *parallèlement*, un changement de topique de discours et il apparaît souvent dans des contextes argumentatifs où il introduit un argument supplémentaire, et d'un ordre différent, ce qui est moins

nettement le cas avec *parallèlement*. *Par ailleurs* ne peut cependant, contrairement à *d'ailleurs*, instituer une relation de justification :

*(17e) Paul et Robert se détestent. * **Par ailleurs**, ils ne se disent même plus bonjour.*

2.2. Différences entre cadratifs et connecteurs

Comme nous venons de le voir, certains adverbes et certains SP cadratifs peuvent donc, en se détachant de leur phrase d'accueil, devenir des connecteurs, quoique ces marqueurs, rappelons-le, instituent des relations en principe opposées : les cadratifs projetant vers le contexte suivant et les connecteurs vers le contexte précédent. Le passage à la fonction de connecteur ne s'observe qu'avec les ajouts susceptibles d'apparaître en tête de phrase, qui sont les plus externes à la prédication, et il ne se produit qu'avec des adverbes ou des SP :

- dont le sens est foncièrement relationnel (comme *autrement, ailleurs, parallèlement,* …).
- qui peuvent prendre une valeur abstraite susceptible de s'appliquer à l'énonciation
- qui se prêtent à des emplois absolus anaphoriques favorisant leur lexicalisation comme mot outil dont l'interprétation sollicite le contexte précédent.

Lorsqu'un ajout adverbial est devenu un pur connecteur et n'a plus que cette valeur, ce qui est le cas de *mais,* il ne peut plus apparaître qu'en tête de phrase, il peut être détaché par la ponctuation de la phrase précédente mais pas de celle dans laquelle il figure, et il n'a plus de potentiel cadratif :

(19a) Paul et Robert se détestent **mais** ils se disent bonjour.

(19b) Paul et Robert se détestent. **Mais** ils se disent bonjour.

(19c) * Paul et Robert se détestent. **Mais,** ils se disent bonjour.

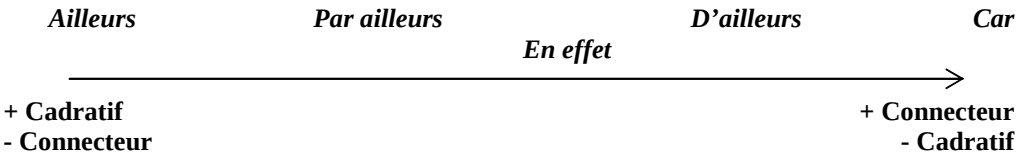
Ce comportement que l'on trouve avec *mais* vaut également pour *d'ailleurs,* qui est proche de *car*, même si *d'ailleurs* peut occuper d'autres positions que l'initiale de phrase. Si *d'ailleurs* n'est donc pas devenu un pur connnecteur, il est pourtant très proche de *car,* du fait qu'il a perdu tout potentiel cadratif. L'établissement de la relation de discours codée par un vrai connecteur tel que *car* ou *mais* ne peut en effet être retardé : le traitement interprétatif doit être effectué dès que le connecteur est énoncé et le calcul de la relation qui implique un retraitement de E1 au vu du contenu de E2 doit être mené jusqu'au bout. Autrement dit, un vrai connecteur ne peut pas "passer au-dessus" d'un énoncé antiorienté, ce qui se vérifie avec *car* et *d'ailleurs* dans (20a-b). On notera le contraste avec *en effet* justificatif (Charolles 2011) en (20c), qui n'a pas perdu entièrement son potentiel cadratif :

(20a) Paul et Robert se détestent. ***Car** ils se disent bonjour, mais ils ne peuvent plus de supporter.

(20b) Paul et Robert se détestent. ***D'ailleurs,** ils se disent bonjour, mais ils ne peuvent plus de supporter.

(20c) Paul et Robert se détestent. **En effet,** ils se disent bonjour, mais ils ne peuvent plus de supporter

Certains adverbiaux occupent donc une position intermédiaire entre les cadratifs purs et les connecteurs, ce que l'on peut schématiser comme suit :



Conclusions

Les ajouts adverbiaux peuvent assumer, quand ils sont détachés en tête de phrase, des fonctions à l'échelle du discours. Ils permettent de répartir les informations dans des rubriques en fonction du sens de l'adverbial et ainsi de structurer les informations communiquées. Ils tirent ce pouvoir du fait qu'ils sont à même d'étendre leur portée au-delà de leur phrase d'accueil. Les adverbiaux cadratifs ont, tout comme les connecteurs, une fonction procédurale : ils influent sur la compréhension des phrases faisant suite à celle-ci.

Quoique les adverbiaux cadratifs portent sur le discours ultérieur et soient orientés vers le contexte en aval, leur apparition en tête de phrase et leur exploitation à des fins cadratives sont le plus souvent préparées par le contexte précédent, de sorte qu'ils entetiennent des liens avec les phrases précédentes et suivantes. Cela les prédestine déjà à se rapprocher des connecteurs. Mais ce rapprochement, comme nous l'avons vu dans la seconde partie, est favorisé par d'autres facteurs, notamment par le fait que l'adverbial inclut un N relationnel qui impose et renforce le lien avec le contexte précédent. Le rapprochement peut-être plus ou moins avancé, pour arriver aux « purs » connecteurs comme *mais* et *car* qui ne sont pas détachables, qui ne peuvent appraître qu'en position interphrastique, et ne peuvent mettre relation que leur phrase hôte avec une une phrase ou une séquence de phrases précédentes. Les rapprochements plus ou moins avancés entre cadratifs et connecteurs, que l'on peut observer à partir d'emplois contemporains, appellent bien évidemment des études diachroniques, en lien avec les études sur la grammaticalisation.

Références bibliographiques

AIKENVALD Alexandra (2004), *Evidentiality*, Oxford, OUP.

ARGENTI Anne-Marie & CHAROLLES Michel (2011), Position's effect (preposed/inserted) of evidential prepositional phrases in *selon X*, Poster, Colloque *AMLAP*, Paris 1-3/09/2011.

ASHER Nicholas & LASCARIDES Alex (2003), *Logic of Conversation*, Cambridge, Cambridge University Press.

BERRENDONNER Alain (1983), Connecteurs pragmatiques et anaphores, *Cahiers de Linguistique Française*, 5, p. 215-246.

BONAMI Olivier ; GODARD Danièle ; KAMPERS-MANHE Brigitte (2003), "Adverb Classification", in CORBLIN Francis & DE SWART Henriet (éds), *Handbook of French semantics*, Stanford, CSLI.

CHAROLLES Michel & LAMIROY Béatrice (2002), Syntaxe phrastique et transphrastique : du but au résultat, in NØLKE Henning & ANDERSEN Hanne (eds), *Macrosyntaxe et macrosémantique*, Bern, Peter Lang, p. 383-419.

CHAROLLES Michel & PRÉVOST Sophie (eds) (2003), Adverbiaux et topiques, *Travaux de Linguistique*, 47.

CHAROLLES Michel (2005), Framing Adverbials and their Role in Discourse Cohesion: From Connection to Foward Labelling, in AURNAGUE Michel ; BRAS Myriame ; LE DRAOULEC Anne & VIEU Laure (eds) *SEM-05 Proceedings*, Biarritz, p. 13-30.

CHAROLLES Michel (2006) Un jour (one day) in narratives, in KORZEN Iorn & LUNDQUIST Lita (eds), *Comparing Anaphors. Between Sentences, Texts and Languages*, Copenhagen, Copenhagen Studies in Language, 34 : p. 11-26.

CHAROLLES Michel & PÉRY-WOODLEY Marie-Paule, (eds) (2005), Les adverbiaux cadratifs, *Langue Française*, 148.

CHAROLLES, Michel & LAMIROY Béatrice (2007). Du lexique à la grammaire : Seulement, simplement, uniquement. *Cahiers de Lexicologie*, 90, 1, p. 93-117.

CHAROLLES Michel. (2011), Les emplois justificatifs de *en effet, i*n NEVEU Franck, BLUMENTHAL Peter & LE QUERLER Nicole eds., *Au commencement était le verbe. Syntaxe, sémantique et cognition*, Peter Lang, p. 29-52.

COLONNA Savéria ; SARDA Laure ; Joël PYNTE & CHAROLLES Michel (soumis), Effect on comprehension of preposed / posteposed spatial adeverbials.

CROMPTON Peter (2006), The effect of position on the discourse scope of adverbials, *Text and Talk*, 26-3, p. 245-279.

DUCROT Oswald (1980), *Les mots du discours*, Paris Editions de Minuit.

ERTESCHIK-SHIR Nomi (2007), Information Structure. The Syntax-discourse interface, Oxford, OUP.

GUIMIER Claude (1996), Les adverbes du français. Le cas des adverbes en –ment, Paris, Ophrys.

LAMIROY Béatrice & CHAROLLES Michel (2005), Utilisation d'un Corpus pour l'évaluation d'hypothèses linguistiques. Etude de *autrement,* in CONDAMINES Anne (éd.), *Sémantique et Corpus*, Paris, Hermès, p. 109-147.

MANN William & THOMPSON Sandra (1986), Relational Propositions in Discourse, *Discourse Processes*, 9, p. 57-90.

MANN William & THOMPSON Sandra (1988), Rhetorical Structure Theory: Toward a functional theory of text organization, *Text* 8/3, p. 243-281.

MOLINIER Christian & LÉVRIER Françoise (2000), *Grammaire des Adverbes*, Genève: Droz.

REINHART Tania (1980), Condition for text coherence, *Poetics to day*, 1/4, p. 161-180.

RIEGEL Martin ; PELLAT Jean-Christophe & RIOUL René (2011), *Grammaire méthodique du français*, Paris, PUF

SANDERS Ted & SPOOREN Wilfred (2001), Text representation as an interface between language and its users, in SANDERS Ted ; SCHILPEROORD Joost & SPOOREN Wilfred, Wilfred, (eds), *Text Representation: Linguistic and Psycholinguistic Aspects*, Amsterdam, Benjamins, p. 1-26.

SARDA Laure & CHAROLLES Michel (2011), *Parallèlement*: de l'espace au temps puis à l'énonciation ? in SERVET Marie-Hélène & BOISSIÉRAS Fabienne (ed.), *Hiérarchisation, énonciation*, Bibliothèque de L'information grammaticale, 66. Peeters, p. 127-156.

THOMPSON Sandra (1985), Grammar and written discourse : Initial vs. Final purpose clauses in English, *Text*, 5, 1-2: p. 55-84.

VIGIER Denis & TERRAN Elise (eds) (2005). Les adverbiaux cadratifs et l'organisation des textes, *Verbum*, XXVII, 2.

Les groupes prépositionnels adnominaux complexes en roumain

Alexandru Mardale

Résumé : Dans cette contribution, nous examinons certaines constructions faisant apparaître deux types de GP complexes adnominaux, en roumain. Nous montrerons, d'une part, qu'il s'agit de GP formés de deux prépositions lexicales exprimant la séparation d'un point ou le lieu de provenance (type A – *plecarea de la mare* « le départ de la mer »). Ces GP s'attachent généralement à des noms déverbaux et peuvent fonctionner comme arguments ou comme adjoints de ces derniers. D'autre part, nous montrerons qu'il s'agit de GP formés d'un élément fonctionnel (à savoir de) et d'une préposition lexicale exprimant le temps ou le lieu (type B – *ședința de la prânz* « la réunion de midi », *ziarul de pe birou* « le journal sur le bureau »). Ces GP apparaissent avec différents types de noms et fonctionnent exclusivement comme adjoints de ces derniers. Nous verrons également que ces GP sont soumis à de nombreuses contraintes d'apparition (notamment l'interprétation spécifique) et que par ailleurs ils représentent une construction particulière au sein des langues romanes.

Introduction

Le roumain présente au moins deux types de groupes prépositionnels (GP) complexes (c.-à-d. formés de deux et, plus rarement, de trois prépositions simples) adnominaux (c.-à-d. qui peuvent s'attacher à des noms).

Les deux types en question sont illustrés par les exemples suivants :

(i) coborâtul de pe deal[1] (type A)
descendu-le de sur colline
« la descente sur la colline »

(ii) copilul de pe stradă (type B)
enfant-le de sur rue
« l'enfant dans la rue »

Dans cette contribution, nous nous proposons d'examiner les conditions d'apparition et de proposer une analyse pour ces constructions, en insistant sur le second type, car il nous semble particulier dans la famille des langues romanes.

Type A

Voici d'autres exemples que nous considérons comme illustrant ce type de construction :

(1) a. venirea de la Paris
venir-la de à Paris
« l'arrivée de Paris »

b. un vânt din deşert (din, forme contractée de de et în)
un vent de-en désert
« un vent du désert »

c. ieşirea de sub anonimat
sortir-la de sous anonymat
« la sortie de l'anonymat »

d. plimbatul pe sub poduri
promené-le sur sous ponts
« la promenade sous les ponts »

e. o excursie pe la mănăstiri
une excursion sur à monastères
« une excursion aux monastères »

f. statul pe lângă prieteni
resté sur près amis
« le fait de rester à côté des amis »

g. alergatul printre obstacole (printre, forme contractée de p(r)e et între)
couru-le sur-entre obstacles
« la course d'obstacles »

[1] Le roumain présente un phénomène particulier à l'égard de l'emploi des prépositions et de l'article défini (qui est enclitique), à savoir la chute de ce dernier. Ce phénomène a lieu quand le nom précédé d'une préposition ne comporte pas d'autre constituant, et cela pour une interprétation définie (exemples (i) et (ii) ci-dessus). En revanche, lorsqu'un nom précédé d'une préposition comporte un autre constituant (quelle que soit sa nature), l'article défini est obligatoire : (i') *coborâtul de pe deal**ul** mare* descendu-le de sur colline-le grand « la descente sur la grande colline », (ii') *copilul de pe strad**a** mea* enfant-le de sur rue-la ma « l'enfant dans ma rue ».

Parmi les propriétés de ce type de construction, notons d'abord que les prépositions pouvant l'introduire (en tant que P1) sont *de* (1a-c) et *pe* (1d-g). *De* est considérée ici comme une préposition ablative, exprimant la séparation d'un point ou d'un lieu de provenance (cf. aussi *plecarea de la Paris* « le départ de Paris »). *Pe* peut exprimer l'approximation et/ou la localisation (cf. aussi *jucatul pe sub masă* « le jeu sous la table »). Il n'y a pas de contrainte quant à l'apparition de la deuxième (ou de la troisième) préposition (P2 ou P3). Ainsi, différentes prépositions peuvent y apparaître (*la* « à », *în* « en », *sub* « sous », *peste* « au delà », *lângă* « près de », *între* « entre », etc.).

Notons également que ces GP complexes s'attachent en général à des noms déverbaux et, de ce fait, ils se retrouvent avec la même forme dans la structure verbale dont ils proviennent :

(2) a. venirea (Mariei) de la Paris <= Maria vine de la Paris
« l'arrivée de Marie de Paris » « Marie arrive de Paris »

b. plimbatul (Mariei) pe sub poduri <= Maria se plimbă pe sub poduri
« la balade de Marie sous les ponts » « Marie se balade sous les ponts »

En ce qui concerne le complément nominal introduit par P2, nous retenons qu'il s'agit principalement d'un GD (c.-à-d. nom avec déterminant), comme en (3) ci-dessous, mais dans certains cas il peut également se réaliser comme GN (c.-à-d. nom sans déterminant), comme en (4) ci-dessous :

(3) a. un vânt din deşert**u**l african / *din deşert african
« un vent du désert africain »

b. o excursie pe la mănăstiri**le** din Moldova / *pe la mănăstiri din Moldova
« une excursion aux monastères de Moldavie »

(4) alergatul printre obstacole naturale / *obstacole**le** naturale
« la course d'obstacles naturels »

Parallèlement à cette distinction, nous observons que les constructions ayant un constituant de type P2+GD en position adnominale peuvent alterner avec un adverbe (comme en (5) ci-dessous), tandis celle qui ont constituant de type P2+GN ne le peuvent pas (6) :

(5) a. venirea de la Paris => venirea de acolo
« l'arrivée de Paris » « l'arrivée de là-bas »

b. mersul pe sub poduri => mersul pe aici
« la marche sous les ponts » « la marche (par) ici »

(6) alergatul printre obstacole => *alergatul pe acolo
« la course d'obstacles » « la course par là-bas »

Le contraste que nous venons de décrire peut s'expliquer également si l'on prend en compte le type de dénotation du complément adnominal. Plus précisément, les constituants réalisés comme GD (cf. (5)) dénotent des lieux (donc un type spécifique d'individu), d'où leur possibilité d'alterner avec des adverbes. En revanche, les constituants GN (cf. (6)) dénotent des propriétés (ne sont donc par référentiels), d'où l'impossibilité d'alterner avec des adverbes.

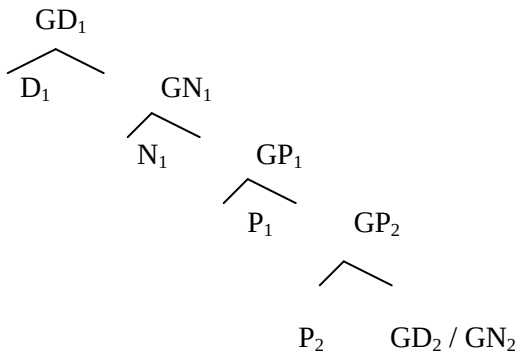
Par ailleurs, soulignons que certains des GP complexes ici en question ont la possibilité d'apparaître avec leur forme simple (c.-à-d. sans P1), comme en (7) ci-dessous. Dans de tels cas, l'interprétation de la construction n'est plus la même, en ce sens que l'on passe d'une lecture de type provenance (avec préposition complexe, (7a-b)) à une lecture de type but (avec préposition simple, (7a'-b')) :

(7) a. venirea de la Paris ≠ a'. venirea la Paris
« l'arrivée de Paris » « l'aller à Paris »

b. mersul pe sub poduri ≠ b'. mersul sub poduri
« la marche sous les ponts » « la marche au dessous des ponts »

Compte tenu des différentes propriétés examinées jusqu'ici, on peut proposer une analyse des GP complexes de type A qui est à représenter selon la structure suivante :

(8)



a. *o sosire de la Paris* « une arrivée de Paris »
b. *o viață pre între străini* « une vie parmi les étrangers »

Sur la base d'une telle représentation, on peut – pour conclure sur les GP complexes de type A – avancer les éléments d'analyse suivants :

(i) ils sont introduits par deux prépositions ordinaires (c.-à-d. lexicales) simples. Autrement dit, P1 lexical prend comme complément un GP introduit par une autre P2 lexicale. Cette dernière peut prendre comme complément soit un GD, soit un GN ;

(ii) le statut du GP adnominal complexe est différent selon le type de nom auquel il s'attache : si ce dernier est un nom déverbal (8a), le GP fonctionne en tant qu'argument ; en revanche, s'il ne s'agit pas d'un nom déverbal (8b), le GP fonctionne comme adjoint.

Type B

Les exemples suivants illustrent ce type :

(9) a. vecinul de la parter
voisin-le de à parterre
« le voisin du rez-de-chaussée »

b. fotografia de pe raft
photographie-la de sur étagère
« la photo sur l'étagère »

c. Revoluția de la 1848
révolution-la de à 1848
« la Révolution de 1848 »

d. răscoala din 1907 (din = de + în)
émeute-la de-en 1907
« l'émeute de 1907 »

e. floarea din grădină
fleur-la de-en jardin
« la fleur du / dans le jardin »

f. priza de sub masă
prise-la de sous table
« la prise sous la table »

g. castelul dintre munți (dintre = de + între)
château-le de-entre montagnes
« le château entre les montagnes »

Il s'agit en l'occurrence d'adjoints prépositionnels adnominaux dont la propriété saillante, qui les distingue de ceux du type précédent, est d'être introduits par *de*, comme seul élément admis en tant que P1 :

(10) a. *vecinul pe la parter
voisin-le sur à parterre

b. *evenimentele pe la 1848
événements-les sur à 1848

Notons que de apparaît également lorsque son complément n'est pas un GP2, mais un GAdv :

(11) a. casa de la mare => casa de acolo
« la maison à la mer » « la maison là-bas »

b. cartea de pe raft => cartea de aici
« le livre sur l'étagère » « le livre d'ici »

c. evenimentele din 1989 => evenimentele de atunci
« les événements de 1989 » « les événements de cette époque-là »

d. progresele din prezent => progresele de acum
« les avancées du présent » « les avancées de maintenant »

Une autre différence avec le type précédent, qui vient plus exactement de leur statut d'adjoints, est que ces GP complexes n'ont pas de préférence quant à la nature du nom auquel ils s'attachent. Ils peuvent, par conséquent, apparaître avec différents types de noms : relationnels (9a), iconiques (9b), d'événement (9c), déverbaux (9d), d'objets (9e-g), etc.

Du point de vue sémantique, ces GP adnominaux expriment la localisation dans le temps ou dans l'espace :

(12) a. cărțile de pe masă = cărțile care se află pe masă
« les livres sur la table » « les livres qui se trouvent sur la table »

b. haina din dulap = haina care este în dulap
« la veste de l'armoire » « la veste qui est dans l'armoire »

c. revoluția din 1989 = revoluția care a avut loc în 1989
« la révolution de 1989 » « la révolution qui a eu lieu en 1989 »

d. orașul dintre lacuri = orașul care se găsește între lacuri
« la ville entre les lacs » « la ville qui se trouve entre les lacs »

De même, on note l'existence de certaines constructions en *de* qui sont ambiguës, pouvant donner lieu à une analyse de Type A ou de Type B. Ainsi, des exemples comme ceux donnés en (13) ci-dessous peuvent recevoir soit la lecture ablative (c.-à-d. en exprimant la provenance), soit la lecture adnominale (c.-à-d. en exprimant la localisation) :

(13) a. vinul din Italia
« le vin d'Italie »

b. vinul din pivniță
« le vin dans la cave »

c. florile de la munte
« les fleurs de la montagne »

Dans de telles situations, le contexte aide à ôter l'ambiguïté potentielle. Ainsi, dans (14) le GP complexe reçoit une interprétation de type A, tandis que dans (15) nous avons une interprétation de type B :

(14) Vinul din Italia se vinde bine peste tot.
« Le vin (qui a été produit / qui provient) d'Italie se vend bien partout »

(15) A adus mai multe vinuri din călătorie. Vinul din Italia a fost mai apreciat decât cele din alte părți.
« Il a apporté plusieurs vins de son voyage. Le vin (qui a été apporté) d'Italie a été plus apprécié que ceux (apportés) d'ailleurs »

Par ailleurs, il est important de souligner que tous les GP de type B ne peuvent pas être introduits par *de*. Ce dernier ne peut pas apparaître dans les situations suivantes :

(i) quand le nom auquel ils s'attachent dénote un événement complexe (c.-à-d. qui comporte des arguments lexicalement réalisés) :

(16) a. organizarea Jocurilor Olimpice **la** Londra **în** 2012 / ***de la** Londra **din** 2012
« l'organisation des JO à Londres en 2012 »

b. semnarea acestei convenții **la** Paris / ***de la** Paris
« la signature de cette convention à Paris »

En revanche, si le nom-tête dénote un événement simple (c.-à-d. sans argument réalisé), on doit utiliser *de* si le GP adnominal a une interprétation spécifique :

(17) a. atentatele **de la** 11 septembrie / ***la** 11 septembrie
« les attentats du 11 septembre »

b. revoluția **din** (de + în) 1907 **din** România / *<u>în</u> 1907 **în** România
« la révolution de 1907 en Roumanie »

Enfin, si le GP adnominal donne lieu a une lecture générique qui décrit un certain type d'événement (c.-à-d. sans décrire une relation entre un événement et une localisation spécifique), *de* est exclu :

(18)a. studiul **la** bibliotecă /***de la** bibliotecă
« l'étude à la bibliothèque »

b. dansul **pe** masă / ***de pe** masă
« la danse sur la table »

(ii) quand le nom auquel ils s'attachent est un indéfini non spécifique dans des contextes intensionnels (19a) ou génériques (20). En effet, la présence de *de* dans ces constructions entraîne une lecture spécifique (19b), tandis que son absence est corrélée à une lecture non spécifique (19c) :

(19) a. *Vrea o casă **la** mare*. (lecture : une maison qui soit (construite) à la mer)
« Il / elle veut une maison à la mer »

b. *Vrea o casă **de la** mare*. (lecture : une certaine maison qui est à la mer)
« Il / elle veut une (certaine) maison à la mer »

c. *Vrea o casă anume **la** mare*.
« Il / elle veut une certaine maison à la mer »

(20) *O casă **la** mare e mai scumpă decât una **la** munte*.
« Une maison à la mer coûte plus cher qu'une maison à la montagne »

(iii) quand le nom auquel ils s'attachent est un défini générique au pluriel :

(21) a. *Casele **la** mare sunt în general locuințe de vacanță*.

« Les maisons à la mer sont généralement des résidences secondaires »

b. *Casele **de la** mare sunt în general locuințe de vacanță*. (même traduction que (21a))

Dans un tel contexte, *de* peut en effet apparaître (comme en (21b)), la lecture générique (donc non spécifique) étant préservée. L'absence de *de* est par ailleurs corrélée à une contrainte supplémentaire : il faut que le sens exprimé par le GP adnominal soit une condition essentielle pour la généralisation exprimée. Ceci explique pourquoi *de* ne peut pas être absent dans l'exemple suivant ((22a) vs. (22b)) :

(22) a. *Clădirile **în** New York sunt foarte înalte.
« Les immeubles à New York sont très hauts »

b. Clădirile **din** (de + în) New York sunt foarte înalte.
« Les immeubles de New York sont très hauts »

(iv) *de* peut également ne pas apparaître lorsque le nom auquel il s'attache est un indéfini apparaissant en position prédicative (c.-à-d. après la copule), dans une structure de type définitionnel ou caractérisant ((23) vs. (24)) :

(23)a. Burdwan este o localitate **în** India.
« Burdwan est une ville en Inde »

b. Burdwan este o localitate **din** (de + în) India. (même traduction que (23a))

Mais :

(24)a. *Indranil est un prieten **în** India.
« Indranil est un ami en Inde »

b. Indranil este un prieten **din** (de + în) India.
« Indranil est un ami d'Inde »

Pour conclure sur les conditions d'apparition de *de* avec le type B, la généralisation que l'on peut faire est qu'il doit en effet apparaître dans les structures ayant une interprétation spécifique.

Un autre aspect qui retient l'attention lorsqu'on examine ce genre de structure est le statut (catégoriel) et la fonction de *de*. En effet, nous avons vu que *de* du type B est différent de celui du type A examiné précédemment, en ce sens qu'il n'a pas les propriétés d'une préposition ordinaire.

Par ailleurs, on rappelle que *de* du type B est traditionnellement considéré comme une sorte de copule nominale dont le rôle est similaire à celui du relatif *care* « qui / que » par lequel il peut d'ailleurs être paraphrasé (cf. aussi (12) ci-dessus) :

(25) un vecin **de** la etajul 1 = un vecin **care** locuieşte la etajul 1
« un voisin du premier étage » « un voisin qui habite au premier étage »

On note également que ce *de* ne peut pas apparaître dans les structures verbales correspondantes, mais – comme nous l'avons vu jusqu'ici – uniquement pour introduire des adjoints en contexte adnominal :

(26) a. Am pus cartea **pe** raft / ***de pe** raft. Cartea **de pe** raft / ***pe** raft a fost interesantă.
« J'ai mi le livre sur l'étagère. Le livre sur l'étagère a été intéressant »

b. Şi-a construit o casă **la** mare / ***de la** mare. Casa **de la** mare / ***la** mare l-a costat foarte mult.
« Il s'est fait construire une maison à la mer. La maison à la mer lui a coûté très cher »

Dans une perspective comparative romane, soulignons que ce genre de construction ne se rencontre pas dans d'autres langues de la même famille :

(27) a. J'ai mis le bol **sur** la table. Le bol **sur** la table / ***de sur** la table est encore chaud. (français)
b. J'ai une réunion **dans** une semaine. La réunion **dans** une semaine / ***de dans** une semaine me stresse.

(28) a. Ho messo la tazza **sul** tavolo. La tazza **sul** tavolo / ***di sul** tavolo è ancora calda. (italien)

b. Ho una riunione **fra** una settimana. La rinunione **de**lla settimana / ***di de**lla settimana mi fanno innervosire.

(29) Puse el bol **sobre** la mesa. El bol **sobre** la mesa / ***de sobre** la mesa todavía está caliente. (espagnol)

Il est toutefois attesté dans des langues comme le chinois et le tagalog (Rubin (2002)), et cela dans des conditions similaires à celles du roumain. Plus précisément, les éléments *de* du chinois et *na* du tagalog apparaissent exclusivement en contexte d'adjonction adnominale, pour introduire des GP dans des constructions à interprétation spécifique.

Les exemples (30a) et (31a) montrent en effet que *de* et *na* sont absents dans les structures verbales locatives, tandis que (30b) et (31b) montrent qu'ils sont obligatoires dans les structures nominales correspondantes :

(30) a. na yiben shu zai zhuozi-shang. (chinois)
celle-CL livre sur table-haut
« le livre est sur la table »

b. na yiben zai zhuozi-shang **de** shu
celle-CL sur table-haut DE livre
« le livre sur la table »

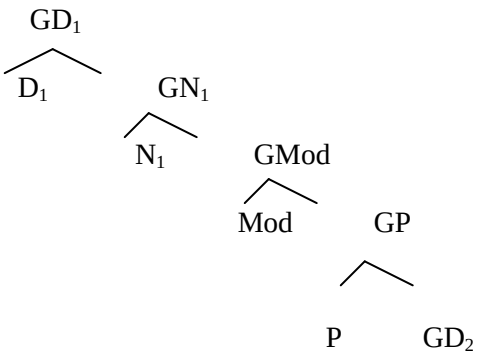
(31) a. Nasa probinsya ang bahay. (tagalog)
in-the province TOP maison
« la maison est en province »

b. (Binili niya) ang bahay **na** nasa probinsya.
acheté il TOP maison NA en-les provinces
« (il a acheté) la maison en province »

Pour répondre à la question du statut catégoriel et de la fonction de *de* apparaissant avec le type B, on peut s'appuyer sur l'analyse de RUBIN (*op. cit.*). Les propriétés spéciales de *de* en roumain (et, par extension, de *de* en chinois et *na* en tagalog) ont amené cet auteur à les considérer non pas comme des catégories lexicales (c.-à-d. comme de véritables prépositions), mais comme une sorte d'éléments fonctionnels (cf. aussi VAN RIEMSDJIK (1990)) dont l'occurrence est fortement contrainte. En effet, comme nous avons pu le constater jusqu'ici, *de* est le seul à pouvoir y apparaître (c.-à-d. qu'il ne peut pas alterner avec d'autres prépositions (cf. (10) ci-dessus)), il n'exprime aucun sens lexical et son occurrence est réservée au contexte adnominal, pour introduire des modifieurs (GP) locatifs ou temporels entraînant à une lecture spécifique de l'ensemble de la structure.

Par conséquent, *de* peut être considéré comme la réalisation d'une catégorie fonctionnelle, à savoir Mod (comme Modifieur). Selon cette analyse, les structures dans lesquelles il apparaît sont à représenter comme suit :

(32)



un cort de pe malul mării « une tente au bord de la mer »

En guise de conclusion

Nous avons examiné dans cette contribution des constructions faisant apparaître deux types de GP complexes adnominaux, en roumain. Nous avons montré, d'une part, qu'il s'agit de GP formés de deux prépositions lexicales exprimant la séparation d'un point ou le lieu de provenance (type A – *plecarea de la mare* « le départ de la mer »). Ils s'attachent généralement à des noms déverbaux et peuvent fonctionner comme arguments ou comme adjoints de ces derniers. D'autre part, nous avons vu qu'il s'agit de GP formés d'un élément fonctionnel (à savoir *de*) et d'une préposition lexicale exprimant le temps ou le lieu (type B – *ședința de la prânz* « la réunion de midi », *ziarul de pe birou* « le journal sur le bureau »). Ils apparaissent avec différents types de noms et fonctionnent exclusivement comme adjoints de ces derniers. Nous avons également constaté qu'ils sont soumis à de nombreuses contraintes d'apparition (notamment l'interprétation spécifique) et que par ailleurs c'est une construction particulière au sein des langues romanes.

Bibliographie

GALR (2005, 2008) = GUȚU-ROMALO Valeria (coord.), *Gramatica Limbii Române*, Vol. I *Cuvântul*, Vol. II *Enunțul*, Bucureşti, Editura Academiei Române.

GBLR (2010) = PANĂ DINDELGAN Gabriela (coord.), *Gramatica de bază a limbii române*, Bucureşti, Editura Univers Enciclopedic Gold.

MARDALE Alexandru (en prép.), Adnominal Prepositional Phrases, in DOBROVIE-SORIN CARMEN ; GIURGEA Ion (coord.), *The Essential Romanian Grammar*, Vol. 1, The Determiner Phrase, ms. accessible à http://www.linguist.univ-paris-diderot.fr/~essromgram/

VAN RIEMSDJIK Henk (1990), Functional Preposition, in PINKSTER H. ; GENÉE I. (eds), *Unity in diversity*, Dordrecht, Foris, p. 229-241.

RUBIN Edward (2002), *The Structure of Modifiers*, ms. University of Utah, accessible à www.hum.utah.edu/linguistics/Faculty/rubin.htm

ȚENCHEA Maria (2011), Relations spatiales et temporelles exprimées par les compléments du nom et de l'adjectif en roumain et en français, in *L'expression de l'espace et du temps en français : quelles formes pour quels sens ?*, Université de Belgrade, les 23-26 mars 2011.

Enrichir le lexique-grammaire : Caractéristiques modales et énonciatives de *Quelle question !*

Christiane Marque-Pucheu

Résumé : Le classement des adverbes établi par M. Gross (1990) selon des critères morphologiques a permis d'intégrer des formes complexes comme *Quelle question !* Toutefois, il est contestable car il met sur le même plan (dans une même table) des exemples comme (*cuire) au charbon* et *(venir) quelle question*, alors que le premier est lié au verbe, contrairement au second qui est utilisable indépendamment d'un verbe (*Moi ! Quelle question !*). Or, comme l'élément *question* lui-même le suggère, l'analyse doit mettre en jeu des « actes de discours ». L'expression, proche d'un modal, réfute le type de phrase (la question) tout en véhiculant généralement un jugement négatif du locuteur. Présentant les caractéristiques d'un commentaire métalinguistique, *Quelle question !* introduit donc une appréciation qui se superpose à la modalité du quasi-(in)certain et rapproche l'expression des évaluatifs. Mais si habituellement, les évaluatifs ont pour argument la proposition (le contenu de la phrase dans laquelle ils se trouvent), dans le cas présent c'est la phrase précédente qui est en jeu, du type *Quelle question (inutile/ridicule) tu poses* ! Par rapport à un adverbe en *–ment* comme *évidemment*, avec lequel elle commute, des différences émergent, par exemple au regard de l'ordre linéaire.

Abstract: The classification of adverbs established by M. Gross based on morphological criteria enabled the integration of complex forms such as *Quelle question!* However, it is somewhat questionable in that it places examples like (*cuire) au charbon* and *(venir) quelle question* in the same classification table, whereas the first adverbial occurrence is clearly linked to the verb, whilst the second may be used independently (*Moi ! Quelle question !*). However as the very element 'question' suggests, its analysis has to include "speech". The expression, which is close to a modal, refutes the type of phrase (the question) whilst generally conveying a negative opinion on the part of the speaker. Bearing the features of a metalinguistic comment, '*Quelle question!*' introduces an appraisal which is superimposed over the almost (un)certain modality and can be compared to the evaluative. However, although the evaluative needs the clause (the content of the sentence in which it features), in this case it concerns the previous sentence, for example *Quelle question (inutile/ridicule) tu poses* ! Compared with '*-ment*' adverbs such as *évidemment*, for which the expression can be substituted, some differences emerge, e.g. with respect to the linear order.

Introduction*

Le classement des « adverbes » effectué par M. GROSS 1990 sur une base morphologique est venu régulariser une classe hétérogène, grâce à la structure générale *Prép Dét N Modif*, ce qui a permis d'intégrer des formes complexes[1]. Les classements fonctionnels, qui recourent à des tests syntaxiques (MOLINIER & LEVRIER 2000) dont ils dénoncent parfois les limites (NØJGAARD 1992, 1993, 1995), privilégient le niveau de discours sur lequel les adverbes portent (énoncé, phrase, etc.). D'autres, plus complexes, prennent en compte portée et position et la corrélation entre ces deux critères (BONAMI, GODARD & KAMPERS-MANHE 2004). Enfin, BONAMI & GODARD 2007 ajoutent un paramètre prosodique. À l'exception de M. GROSS 1990, ces études se limitent en général aux formes simples, essentiellement aux adverbes en *–ment*, tout comme les approches non classificatoires étudiant les variations de portée (GUIMIER 1996) ou celles qui recourent à des opérations énonciatives (GEZUNDHAJT 2000).

L'insertion d'une forme complexe comme *Quelle question !* dans les tables de M. GROSS 1990 soulève de nombreux problèmes dès lors qu'on sort de l'analyse strictement morphologique. En vertu de ce dernier classement, l'expression dans (*venir) quelle question !* cohabite dans ces tables avec des exemples comme (*cuire) au charbon* ou *(venir) à la hâte*[2]. Ce classement est contestable car il met sur le même plan, par exemple (*travailler…) bof* et (*venir) auparavant*, alors que le premier n'est pas lié au verbe, contrairement au second. Or, de même que *bof*, *quelle question !* est autonome, utilisable indépendamment d'un verbe (*Moi, Quelle question ! / Urgent ? Quelle question !*) et ne saurait passer pour un ajout au verbe, alors que *auparavant* ou *à la hâte* modifient *venir* et que *au charbon* modifie *cuire*. Si le travail de Gross 1990 représente une avancée considérable dans le listage des « adverbes » et un premier inventaire de leurs propriétés, il n'est pas sans défaut du point de vue syntaxique et distributionnel. En effet, les exemples précédents montrent que sont rangés dans une même classe des « adverbes » ayant une fonction (par rapport au verbe) et ceux qui n'en ont pas (sauf à parler d'exclamation, d'apostrophe, d'interjection, etc.). Si ce rapprochement est conforme au cadre méthodologique revendiqué, c'est-à-dire strictement morphologique, il neutralise la question du cotexte, donc le fait que *travailler* à gauche de *bof* n'a pas le même statut que *venir* à gauche de *auparavant*. Il laisse donc les dimensions énonciative et textuelle de côté. De même, le classement de MOLINIER & LEVRIER 2000 qui peut justifier un étiquetage de cette expression comme modale, puisqu'elle peut constituer une réponse à une question totale, présente rapidement des limites, de nombreux tests ne s'appliquant pas.

Notre hypothèse est que *Quelle question !* présente des propriétés syntaxiques particulières par rapport aux adverbes modaux habituels, ou dits « assertifs » chez BORILLO 1976, et que par ailleurs, l'interprétation de ce

* Merci à Danielle Leeman et à mes relecteurs anonymes pour leurs suggestions.

[1] N correspond au mot unique (*nuitamment*), Prép à une préposition et Modif à un modifieur ; ces derniers sont réalisés (*par une nuit sans lune*) ou non (*de nuit, la nuit, la nuit où Luc est parti*).

[2] Elle est étiquetée PDETC, où P, DET et C correspondent respectivement à *Prép*, *Dét* et *N Modif*. Mais dans *Quelle question ! Prép* n'est pas réalisé.

« connecteur » qui n'en est pas un, marquant l'impuissance à établir un lien logique avec ce qui précède, suppose la prise en compte du plan énonciatif, du fait même qu'il s'inscrit dans des actes de parole.

Les exemples sont extraits de la base de données Frantext, les occurrences s'échelonnant entre 1835 et 2010, du *Monde* et de *Libération*, ou proviennent encore du moteur de recherche Google ; le raisonnement s'appuie aussi sur des énoncés forgés selon notre intuition.

1. Des propriétés de modal

Éliminons d'abord les emplois (1) et (2) qui ne concernent pas notre sujet et où *Quelle question* apparaît respectivement dans une interrogative directe et une interrogative indirecte :

> 1) Marquis. *C'est admirable, positivement... changeant de ton. Voulez-vous me permettre, monsieur l'artiste, de vous adresser une question ?*
> *Marignan. **Quelle question ?*** (Meilhac Halévy 1967)

> 2) *On ne sait pas trop **quelle question** Tina Kieffer posera ce soir* (*Libération*, 2 mai 1996)

Dans un autre emploi (3-4), objet de cette étude, l'expression constitue le plus souvent une réponse à une question exprimée, soit dans un dialogue (3) :

> 3) - *Où va-t-on?*
>
> - ***Quelle question**, soupira Victor... à Orly, évidemment !* (Giraud 1966)

soit dans un monologue (4), par exemple dans un discours indirect libre où l'énonciateur se « dédouble » dans un jeu de question/réponse :

> 4) *Aimerais-je pouvoir étudier ? me demandèrent-ils. Étudier ? Le sens précis et pratique du mot m'échappait : en campagne, à l'école, on « apprend », bien ou non, mais avoir le nez dans les livres ? ! Pardi, **quelle question !** Sans que je m'en rende compte (au moins au début) ils me firent passer des tests...* (Szczupak-Thomas 2008)

Dans ces deux exemples, *Quelle question !* est une exclamative directe. Et le passage de l'emploi (1) à (3) ou (4) est indépendant du passage du dialogue au monologue puisque le premier emploi (interrogatif) s'observe nécessairement en dialogue, alors que le second peut figurer dans un dialogue ou dans un monologue. En effet, d'une manière générale, un emploi métacommunicatif n'implique pas nécessairement deux interlocuteurs (NØJGAARD 1992 : 19).

1.1. Définition et tests

L'expression a été définie en Introduction comme modale, car elle fonctionne comme une phrase (BORILLO 1976 : 76), pouvant constituer une réponse à une interrogation totale[3] :

> 5) Delomer : *Quoi ! si je n'avais rien, vous me rechercheriez avec le même empressement ! Vous me prendriez sans dot ?... Consultez-vous bien.*
> M. Jullefort : ***Quelle question !** Je n'ai pas besoin de me consulter, je vous donnerais avec la même tendresse une preuve de mon désintéressement.* (Mercier 1775)

Une restriction toutefois : NØJGAARD (1992 : 43) a suggéré que les modaux fonctionnaient plus exactement comme « semi-phrases ». Ainsi, *Quelle question !* ne représente pas nécessairement la phrase dans son ensemble, mais peut correspondre à un ajout. L'expression est d'ailleurs compatible avec *non* (6), *si* (7) ou *oui* (8), auxquels elle peut s'adjoindre :

> *6) Deux verres !*
> *- Ça doit être du bon, dit Pierrette qu'on avait oubliée à côté.*
> *- Du très bon mademoiselle, dit von Brautschich.*
> *- Il est pas sucré ?*
> *- Non ! **Quelle question** ! Je ne suis pas un barbare* (Déon 1960)

> *7) Il est pas sucré ?*
> *- Si ! **Quelle question** !*

> *8) Il est sucré ?*
> *- Oui ! **Quelle question** !*

[3] La table de GROSS 1990 la catalogue comme non conjonctif, au même titre que *(cuire) au charbon*, ce qui suggère une portée sur le verbe. Or notre étude montre que la portée est plus large.

1.2. Interprétation : la modalité du quasi-(in)certain

Quelle question ! introduit une modalité (logique) sur la valeur de vérité de l'énoncé-réponse. La modalité est celle du quasi-(in)certain : ainsi, en (6), les possibilités de validation de l'énoncé sont maximales[4]. Cette conviction est inférée explicitement ou non de ce que le locuteur sait ou voit, par exemple de lui dans (5) ci-dessus, ou du monde environnant (9) :

> 9) *Où peut-on dîner ? Elle avait feint de replonger le nez dans ses comptes. Elle releva la tête, visiblement exaspérée.*
> *- Mais où vous voudrez, partout,* ***quelle question !*** *Dans les brasseries de la rue d'Isly, à celle des Facultés. On vous servira jusqu' à minuit, une heure.* (Droit 1967)

Elle peut être inférée aussi de sa conception du monde dans l'exemple (4) que nous rappelons :

> 4) *Étudier ? Le sens précis et pratique du mot m'échappait : en campagne, à l'école, on « apprend », bien ou non, mais avoir le nez dans les livres ? ! Pardi,* ***quelle question !*** *Sans que je m'en rende compte (au moins au début), ils me firent passer des tests...* (Zczupak-Thomas 2008)

Quelle question ! fonctionne donc comme un modal : il constitue une semi-prophrase et sémantiquement, il concerne la valeur de vérité de la proposition contenue dans la question précédente. Un rapprochement pourrait être effectué avec *évidemment* (voir note 4) qui peut être associé aussi à *oui/non*. Mais *Quelle question !* ne possède pas toutes les propriétés syntaxiques des modaux (voir 2.1) ; de plus, l'analyse déjà donnée (modal assertif) ne rend pas compte intégralement du sens de l'expression (2.2).

2. Particularités

2.1. Particularités formelles

2.1.1. Une mobilité restreinte

Il convient d'abord de s'interroger sur le statut adverbial de *Quelle question !* Même si les tests ont prêté le flanc aux critiques (NØJGAARD 1992 : 75-78), ils servent de « garde-fou » aux linguistes. Si l'on prend comme critère définitoire de l'adverbe qu'il admet (sans pause) la place entre l'auxiliaire et le participe passé des temps composés, on n'a pas **J'ai quelle question invité Pierre*, comme on aurait *J'ai sans aucun doute invité Pierre* : *sans aucun doute* est analysable syntaxiquement comme un « adverbe », quoiqu'il n'en soit pas un morphologiquement. Cela posé, on observe que, de fait, *Quelle question !* ne peut s'insérer dans une phrase : si l'adverbe peut être employé de manière autonome (en réponse ou en commentaire : *Pierre doit être arrivé… - Sans aucun doute ! / Evidemment*), il peut aussi s'inscrire dans une phrase (*Sans aucun doute, il est déjà arrivé* / *Il est déjà sans aucun doute arrivé*), ce qui n'est pas le cas de *Quelle question !* (**Quelle question !, il est déjà arrivé / ?? *Il est déjà quelle question ! arrivé*)[5]. Seule une pause, matérialisée par les tirets, autorise la position à droite de l'auxiliaire :

> 10) *Ce que Dumas veut savoir, ce n'est pas tant ce que [les anciens taulards] ont vécu que ce dont ils se souviennent, et comment ils analysent leur passage à l'ombre ─ inattendu chez les quatre ─, et s'ils ont ─* ***quelle question !*** *─ changé*. (*Libération* 6 février 1998)

Contrairement aux modaux, y compris les formes complexes comme *bien sûr*, *Quelle question !* a donc une mobilité relative, la position de prédilection étant l'initiale détachée.

Le test de la focalisation ne s'applique pas davantage (MOLINIER & LEVRIER 2000 : 102) :

> 11) *Qui a offert des fleurs à Marie pour son anniversaire ?*
>
> ** C'est* ***quelle question*** *Luc qui a offert des fleurs à Marie pour son anniversaire*

sauf pause importante :

> *? C'est -* ***quelle question !*** *- Luc qui a offert des fleurs à Marie pour son anniversaire*

L'expression n'a donc pas l'autonomie d'un adverbe en *–ment* et présente les caractéristiques d'un commentaire métalinguistique, du type *Quelle question (inutile) tu poses* !

GEZUNDHAJT (2000 : 94) oppose deux fonctionnements de *évidemment*, dont l'un pourrait illustrer la spécificité de *Quelle question !* par rapport à la mobilité :

---

[4] BORILLO 1976 et GEZUNDHAJT (2000 : 229-232) divergent sur la nature de la modalité de *évidemment* qui commute avec *Quelle question !*. La première considère que cet adverbe relève d'une certitude plus ou moins forte, alors que la seconde opte pour la modalité du nécessaire. Le test de la négation qui discrimine nécessité et probabilité donne raison à BORILLO : *Il n'a pas forcément réussi vs *Il n'a pas (peut-être + évidemment) réussi*.

[5] Remarquons la variante *Cette question ! C'te question !* dans *Bien sûr qu'il est arrivé, cette question ! / Quelle question !*

a) Ces tableaux sont évidemment des faux

b) Évidemment, ces tableaux sont des faux

Elle conteste le rapprochement souvent effectué entre (a) et (b), arguant que dans (a), il s'agit d'une « déduction d'ordre inférentiel d'une propriété à partir des connaissances que l'énonciateur a du sujet dont il est question » ; dans (b), *évidemment* détaché en tête de P introduit « une concession accordée aux coénonciateurs et basée sur les idées que l'énonciateur attribue à ces derniers ». Dans (a), *évidemment des faux* commute avec *des faux évidents ;* dans (b) *évidemment* commute avec *Je vous l'accorde* ou *Je vous le concède* :

a') Ces tableaux sont des faux évidents

b') (Je vous l'accorde + Je vous le concède), ces tableaux sont des faux

L'expression *Quelle question !* correspond à l'emploi (b). Elle reste liée au dialogue, ce qui bloque sa mobilité. De plus, l'identité de fonctionnement avec b) suggère une portée phrastique.

2.1.2. Portée phrastique

Des tests révèlent la portée phrastique[6], comme la propriété pour un adverbe de pouvoir apparaître dans une construction *Adv que P* (a) ou *Adv, que P* (b), même si cette propriété n'est pas systématique pour les Adv en *–ment* et si elle ne peut permettre de définir la classe des modaux dans son entier :

a) Évidemment qu'il est parti

b) Évidemment, qu'il est parti

La construction (a), qui se caractérise par une pause marquée par une virgule avant la complétive, s'inscrit généralement dans un dialogue (question/réponse). Or, *Quelle question !* qui s'utilise souvent comme réponse à une question n'accepte pas pour autant d'entrer dans cette construction :

- Il est parti ?

*- * **Quelle question**, qu'il est parti*

La construction attestée est

***Quelle question,** (de demander) s'il est parti !*

qui, contrairement au cas précédent, met en jeu une interrogative indirecte et non une complétive. Remarquons qu'inversement *évidemment* refuse cette construction :

** Évidemment, s'il est parti*

Nous avons donc une distribution complémentaire qui mériterait d'être précisée dans une étude plus détaillée. Les tests à notre disposition pour les modaux doivent donc être adaptés.

2.1.3. Question partielle

Contrairement aux modaux en *–ment* (13), *Quelle question !* peut constituer une réponse à une question partielle sans reprise obligatoire du syntagme sur lequel porte la question :

12) Qui vient dîner ?

- ***Quelle question !** (Pierre + E)*

13) Qui vient dîner ?

*- (Pierre + *E) / (Pierre + *E) évidemment*

Autrement dit, *Quelle question !* met surtout l'accent sur la question comme type de phrase et non pas sur le détail de la réponse (*Pierre*), ce qu'indiquait la construction *Quelle question, de demander si P !*.

Si le rattachement de l'expression aux modaux (voir 1.) ne rend pas compte des propriétés formelles, il en occulte également les particularités sémantiques.

[6] La portée sur la phrase peut être justifiée grâce à une paraphrase du type *Adv, P = Cela est Adj* ou *Il est Adj que P, Adj* étant relié morphologiquement à *Adv* : *Évidemment, Paul viendra = Paul viendra, cela est évident / Il est évident que Paul viendra.* Cela ne vaut que pour les adverbes reliés à un adjectif, ce qui n'est clairement pas le cas de *Quelle question !*

2.2. Particularités sémantiques : un modal de quasi-(in)certitude

Quelle question ! peut correspondre aussi bien à *oui / bien sûr* (14) qu'à *non* (15) :

> *14) Dis toujours.*
> *- As-tu jamais attaché de l'importance à la présence de quelqu'un ?*
> *-* ***Quelle question !*** *Bien sûr*. (Romains 1929)

> 15) *tmen veu ?? - Mais non* ***quelle question*** *j'adore être pris pour un con […]*[7].

Pour s'en convaincre, intervertissons *non* et *oui* dans (14) et (15) :

> 14a) *- Dis toujours.*
> *- As-tu jamais attaché de l'importance à la présence de quelqu'un ?*
> *-* ***Quelle question !*** *Bien sûr que* ***non***. (Romains 1929)

> 15a) *tmen veu ?? - Mais oui* ***quelle question*** *j'adore être pris pour un con.*

Or MOLINIER & LEVRIER (2000 : 104) notent, à la suite de BORILLO (1976 : 80), que les modaux « ont tous, nécessairement, une orientation positive », ce qu'atteste l'absence des adverbes **incertainement et *impossiblement*, alors que la langue dispose des adjectifs *incertain* et *impossible*. Pour obtenir une orientation négative, la langue recourt au morphème *pas* (*certainement pas*, *évidemment pas)* qui se trouve alors sous la portée de l'adverbe. Sur ce point, *Quelle question !* se distingue en admettant une double interprétation, soit *Quelle question ! Oui, c'est vrai*, soit *Quelle question ! Oui, c'est faux*[8].

Cette différence se double d'une autre différence. Non seulement *Quelle question !* introduit un commentaire sur la valeur de vérité de la proposition contenue dans la question précédente, mais sa profération introduit aussi une appréciation subjective.

2.3. Particularités énonciatives

2.3.1. Un commentaire appréciatif

GEZUNDHAJT (2000 : 4) fait grief aux classifications syntaxiques et fonctionnelles de ne pas prendre en compte « l'énonciateur à l'intérieur de la situation énonciative ». Dans cet emploi de commentaire métalinguistique, *Quelle question !* introduit une appréciation qui se superpose à la modalité du quasi-(in)certain et rapproche l'expression des évaluatifs. Mais habituellement, les évaluatifs ont pour argument la proposition (le contenu de la phrase dans laquelle ils se trouvent). Or dans le cas présent, c'est la phrase précédente qui est en jeu. L'expression indique au destinataire qu'il doit l'interpréter comme *Quelle question (inutile + ridicule) tu as posée !*[9]. Dans l'exemple (16), *ridicule* peut fort bien être supprimé sans que l'interprétation soit modifiée :

> 16) Philippe. *Tu vas là-bas, tu vas tout vivre sans moi, tu vas te faire des copains... est-ce que je peux avoir confiance en toi ?*
> Senac. ***Quelle question ridicule*** *!*
> Philippe. *Mais non, parce que tu vas être tué, toi aussi, comme les autres.*
> Senac. *Et puis après !* (Montherlant 1929)

Dans cet emploi, *Quelle question !* est souvent précédé ou suivi de *mais :*

> *(17) - Est-il vraiment aussi honnête homme qu'on le dit ? demanda-t-il négligemment à M. d'Harville, qui se souvint alors de ce que Rodolphe avait raconté à Clémence à propos du notaire.*
> *-* ***Jacques Ferrand ? Quelle question ! Mais*** *c'est un homme d'une probité antique, dit M. de Lucenay.* (Sue 1843)

Il en découle un rapport particulier à la coénonciation.

2.3.2. Le rapport à la coénonciation

GEZUNDHAJT (2000 : 245) note que « l'adverbe évidemment renvoie à une opération de fermeture. Il marque une assertion forte sans possibilité de contradiction ». Cette analyse vaut pour *Quelle question !* Généralement, le dialogue coupe court après la profération de l'expression. S'il est poursuivi par le locuteur 2, le

---

[7] L'orthographe du site (consulté le 26 avril 2012) a été respectée (http://fr-fr.facebook.com/pages/tmen-veu-Mais-non-quelle-question-jadore-etre-pris-pour-un-con-D/136653106363523).

[8] Un exemple apparenté est fourni par *Tu parles !* qui peut s'analyser comme *Tu parles que (oui + non)* (MARQUE-PUCHEU 2011).

[9] Les équivalents en anglais ('what a question !') et en italien ('che domanda !' / 'bella domanda !') ont une fonction identique, mais présentent un figement plus fort.

locuteur 1 y met fin une nouvelle fois, témoin *et puis après* dans l'exemple (16). Soit il renchérit, ce qu'indique *un degré extrême* dans la dernière réplique (18) :

18)- *As-tu jamais attaché de l'importance à la présence de quelqu'un ?*

- ***Quelle question*** *! Bien sûr.*

- *Oui, mais quel degré d'importance ?*

- *Un degré extrême, parfois.* (Romains 1929)

soit il confirme sa question en paraphrasant *se plaire*, qui figure dans la question initiale, par *ne pas s'ennuyer* dans la dernière réplique (19) :

19) *Dorothy. - Vous pensez que vous allez vous plaire en Amérique ?*

Jimmy. - ***Quelle question*** *! Bien sûr, Dot !*

Dorothy. - Vous savez, ce n'est pas très drôle, Boston ! Peut-être vous allez vous ennuyer là-bas ?

Jimmy.- Je ne m'ennuierai pas, Dot (Bourdet 1931)

C'est donc davantage le type de phrase lui-même (l'interrogative) qui est contesté que la proposition contenue dans la question elle-même.

Conclusion

Le classement des adverbes chez M. GROSS (1990) intègre des formes complexes et exclamatives qu'illustre l'expression *Quelle question !*. L'emploi étudié apparait dans le cadre d'un dialogue ou, plus rarement, d'un monologue, dans un discours indirect libre. Il se caractérise par son sens polémique : il correspond souvent à une réfutation du type de P (question), la question allant de soi, qu'elle soit totale ou partielle. Dans le cas d'une question totale, l'expression peut correspondre soit à une approbation forte du contenu (parallèlement à la réfutation de l'acte de questionnement), signifiant alors *oui* :

20)- *Dis toujours.*

- *As-tu jamais attaché de l'importance à la présence de quelqu'un ?*

- ***Quelle question !*** Bien sûr. (Romains 1929)

soit à une mise en doute du contenu/réfutation du contenu proféré dans une question, et elle signifie alors *non* :

21) *Dire comment elle est devenue si rondelette ?*

- *Pardi, en se crevant de mangeaille comme vous et moi.*

- *Fort bien, et ce qu'elle mangeoit vivoit-il ? Ou non ?*

- ***Quelle question !*** *Pardi* ***non****, il ne vivoit pas. - quoi !* (Diderot 1762)

Cette parenté avec les modaux dissimule néanmoins des différences en terme de portée et de mobilité. L'expression se caractérise également par le fait qu'elle traduit une modalité appréciative, *ridicule* étant sous-entendu.

Le classement morphologique (GROSS 1990) pourrait être enrichi par des propriétés supplémentaires et la présente étude élargie à des adverbiaux à contenu phrastique ou à des exclamations.

Références

BONAMI Olivier ; GODARD Danielle ; KAMPERS-MANHE Brigitte (2004), Adverb Classification, in : CORBLIN Francis ; DE SWART Hélène. (eds), *Handbook of French semantics*, Stanford, CSLI, p. 143-184.

BONAMI Olivier ; GODARD Danielle (2007), Quelle syntaxe, incidemment, pour les adverbes incidents ?, *Bulletin de la société de linguistique* 102, p. 255-284.

BORILLO Andrée (1976), Les adverbes et la modalisation de l'assertion, *Langue française* 30, Paris, Larousse, p. 74-89.

GEZUNDHAJT Henriette (2000), *Adverbes en –ment et opérations énonciatives*, Berne, P. Lang.

GROSS Maurice (1990), *Syntaxe de l'adverbe*, Paris, Asstril.

GUIMIER Claude (1996), *Les adverbes du français. Le cas des adverbes en –ment*, Paris, Ophrys.

MARQUE-PUCHEU Christiane (2011), Entre statut phrastique et statut textuel : l'exemple des énoncés situationnels, *Discours* 2011, 9 disponible sur http://discours.revues.org/8553.

MOLINIER Christian ; LEVRIER Françoise (2000), *Grammaire des adverbes. Description des formes en –ment*, Genève, Droz.

NØJGAARD Morten (1992, 1993, 1995), *Les adverbes en français : essai de description fonctionnelle, tomes I, II, III*, Copenhagen, Munksgaard.

Do well, do good : fare bene, benino, benone, benissimo
Italian lexicography in disarray

Ignazio Mauro Mirto

Abstract: When Italian *bene* 'good/well' occurs post-verbally with *fare* 'do', several constructs with remarkably different argument frames are involved. This paper deals with three of them: (a) *Il latte fa bene ai bambini* 'Milk is good for children'; (b) *Fa bene il suo lavoro* 'She does her job well', and (c) *Faresti bene a non dire niente* 'You would do well to say nothing about it'. We discuss dictionary discrepancies concerning the lexical category of *bene* in (a), which we take to be a noun predicate, and draw a distinction between the adverbial uses in (b) and (c).

0. Introduction

Similarities and differences between the three sentences below invite a comparison:

(1) Leo fece bene a Giulio.
'Leo did good to Giulio.'

(2) Leo fece bene l'esercizio.
'Leo did the exercise well.'

(3) Leo fece bene a tacere.
'Leo did well to remain silent.'

All of the examples share the presence of the verb *fare* 'do' immediately followed by *bene* 'well / good' (unmarked linear order).[1] Their key differences are as follows: sentence (1) includes an indirect object (as the cliticization test shows: *Leo gli fece bene* 'Leo did good to him'[2]); sentence (2) contains a direct object (*l'esercizio*); finally, sentence (3) comprises a subordinate clause with an infinitive (*tacere*) – mandatorily introduced by the preposition *a* 'to' – the unexpressed subject of which is co-indexed with the subject of the main clause (*Leo*).

Our attention is focused on the occurrences of *bene* in the three examples. In Italian dictionaries, the classification of this word in lexical categories is uncontroversial in the uses illustrated in (2) and (3). In entries of dictionaries that by chance provide examples in which *bene* co-occurs with *fare* as in (2) and (3) *bene* is invariably classified as an adverb. However, regarding the use of *bene* in (1), a type in which the subject can also be inanimate,[3] a number of discrepancies emerge. On the one hand, the *DEVOTO-OLI Le Monnier* classifies the *bene* occurring in example (4a) as a noun. For our purposes, the subject of this example should be interpreted as in (4b):

(4a) Bevi un sorso, ti farà bene.
'Take a sip, it will do you good.'

(4b) (Ciò + Bere un sorso) ti farà bene.
'(This + Taking a sip) will do you good.'

On the other hand, the dictionary *Il grande italiano 2008* (GABRIELLI, 2007) places the example in (5) in an entry for adverbial uses of *bene*:[4]

(5) Una passeggiata ti farà bene.
'A walk will do you good.'

Both the sentences in (4b) and (5) belong to the type exemplified in (1), for they contain the indirect object clitic *ti* 'to you' (in these cases *fare bene* is semantically equivalent to *giovare* 'to be a help'). The divergence between the entries of the two dictionaries shows that the classification of this use of *bene* is to some degree baffling.

Grammars of Italian pay little attention to the differences outlined above for sentences such as those in (1)-(3). The relevant chapters of a number of them do provide a few examples, but neither compare nor analyze the uses

[1] Other constructs with *fare* + *bene* will be disregarded, e.g. *Un medico fa il bene dei pazienti* 'A doctor works for her/his patients' sake', *Tieniti lontano dal male e fa il bene* 'Turn away from evil and engage in good works' (Peter 3:11), and *A scuola, il bambino fa bene* 'The child is doing well at school'. The comparative form of *bene*, i.e. *meglio*, will be ignored as well. The tests carried out below often yield the same results when in (1), (2), and (3) the antonym *male* 'badly, evil' replaces *bene*. For space reasons they will not be shown.

[2] The head of the indirect object can also be [− animate]: *Questa lozione fa bene ai capelli* 'This lotion is good for the hair'.

[3] As (4b) and (5) illustrate. The subject of the former is an infinitive, whereas that of the latter is a noun phrase headed by an abstract noun. This property of the subjects of the clause type exemplified in (1) will be discussed in section 8, dedicated to selectional restrictions. Further investigation is needed to fully ascertain whether (1) and (5) illustrate the very same construction.

[4] As does the bilingual dictionary *IL BOCH* (French / Italian) for the following example: *L'aria del mare gli fa molto bene* 'Sea air is good for him'. Numerous inconsistencies are also found in the Italian *corpus* of *SketchEngine* (http://www.sketchengine.co.uk/).

shown above (see LEPSCHY and LEPSCHY 1981, LONZI 1991, SCHWARZE 2009/1995, SENSINI 1997, SERIANNI 1989).

This paper aims at making a detailed comparison between the occurrences of *fare bene* 'to do well / good' by carrying out tests that are able to formally distinguish the three uses. In our opinion, such tests bear out what follows: (i) the examples in (1) to (3) instantiate three distinct clause types, which we will refer to with the labels A, B, and C, respectively; (ii) in (1), *bene* is best analyzed as a predicate noun; (iii) finally, in both (2) and (3) *bene* fulfills adverbial functions, though a distinction between the two uses must be drawn.

1. Test 1: the omission of *bene*

Keeping out *bene* in the three clause types brings about ungrammatical outcomes in (1) and (3), but leaves (2) well-formed, as the sentences below illustrate:

(6) * Leo fece a Giulio.
Leo did to Giulio

(7) Leo fece l'esercizio.
'Leo did the exercise.'

(8) * Leo fece a tacere.
Leo did to remain silent

This test provides evidence that in type B, i.e. (2) and (7), *bene* is optional, a characteristic of prototypical adverbs, and particularly of manner adverbs (see HASER and KORTMANN, 2006: 67-68). In other terms, in (2) *bene* modifies a head (most probably the VP) and is valence-independent (BUSSMANN, 1996: 9). The ill-formed (6) and (8) show that these characteristics are not found in type A and type C.

2. Test 2: substitution

The same results concerning well-formedness obtain if *bene* is substituted for with an adverb in *–mente*: [5]

(9) * Leo fece ottimamente a Giulio.
Leo did brilliantly to Giulio.

(10) Leo fece ottimamente l'esercizio.
'Leo did the exercise brilliantly.'

(11) * Leo fece ottimamente a tacere.
Leo did brilliantly to remain silent.

3. Test 3: *bene* in final position

It is common knowledge that adverbs « often have the option of occurring in several positions in a sentence » (LANGUAGE FILES 1991: 173). Below, we carry out the test by placing *bene* in final position:

(12) ! Leo fece a Giulio bene.
'Leo did good to Giulio.'

(13) Leo fece l'esercizio bene.
'Leo did the exercise well.'

(14) * Leo fece a tacere bene.
Leo did to remain silent

The exclamation mark in (12) signals that the sentence is well-formed but needs a marked intonation pattern.[6] As expected, *bene* in final position does not modify the grammatical status of (2), as (13) demonstrates, whilst the same position generates an ungrammatical outcome in (14).

4. Test 4: quantifiers

Certain quantifiers, e.g. *molto* or *parecchio* 'much, a lot of', which can work either as adverbs or adjectives, do not reveal any differences between the three uses of *bene*:

[5] The coordination of *bene* with a manner adverb in *-mente* (see MAROTTA, 1989: 110) can be considered a variant of this test: **Leo fece bene ed egregiamente a Giulio* 'Leo did good and eminently to Giulio', *Leo fece bene e rapidamente l'esercizio* 'Leo did the exercise well and rapidly.', **Leo fece bene ed egregiamente a tacere* 'Leo did well and eminently to remain silent'.

[6] The use of the partitive article on *bene* makes sentence (12) perfectly well-formed: *Leo fece a Giulio del bene* (see section 5). If *bene* is fronted, a marked intonation pattern rescues all types: !*Bene Leo fece a Giulio*!, !*Bene Leo fece l'esercizio*, !*Bene Leo fece a tacere*.

(15) Leo fece (molto + parecchio) bene a Giulio.
'Leo did (a lot of) good to Giulio.'

(16) Leo fece (molto + parecchio) bene l'esercizio.
'Leo did the exercise (very) well.'

(17) Leo fece (molto + parecchio) bene a tacere.
'Leo did (very) well to remain silent.'

Other quantifiers, especially multi-word ones that include an article, e.g. *un gran(de)* 'real(ly)' and *un sacco di* 'a lot of', unveil a number of differences:

(18) Leo fece (un gran + un sacco di) bene a Giulio.
'Leo did (a lot of) good to Giulio.'

(19) * Leo fece (un gran + un sacco di) bene l'esercizio.
Leo did (very + a lot of) well the exercise

(20) * Leo fece (un gran + un sacco di) bene a tacere.
Leo did (very + a lot of) well to remain.silent

The sharp difference in grammaticality between type A (see (1), (15), and (18)) and the types B (see (2), (16), and (19)) and C (see (3), (17), and (20)) appears to be due to the kind of quantifier employed: *un gran* and *un sacco di* modify nouns within the domain of the noun phrase. Thus the set of sentences in (15) to (20) suggests that in type A *bene* functions as a noun, unlike in type B and type C.

5. Test 5: the partitive article

The insertion of the partitive article before *bene* gives results that are fully in line with those reached with test 4, as the examples below illustrate: [7]

(21) Leo fece del bene a Giulio.
'Leo did good to Giulio.'

(22) * Leo fece del bene l'esercizio.
Leo did of.the well the exercise.

(23) * Leo fece del bene a tacere.
Leo did of.the well to remain.silent.

The possibility of inserting the partitive article in (21) and the ill-formed outcomes in (22) and (23) show that in type A only *bene* functions as a noun (an abstract noun, uncountable, masculine, and singular).

6. Test 6: the discontinuous element *non… che*

The sentences in (24a) and (24b) are closely related:

(24) a. Mio cugino ha comprato biscotti.
'My cousin bought (some) biscuits.'

b. Mio cugino non ha comprato che biscotti.
'My cousin bought biscuits only.'

In (24b), *non* 'not' forms a single item with the following *che* 'literally: that', despite the broken linear concatenation. As the translation in (24b) shows, the value of *non* […] *che* is that of a quantifier such as *solo* (or *soltanto*) 'only' ((24b) is semantically equivalent to *Mio cugino ha comprato solo/soltanto biscotti* 'My cousin bought biscuits only'), that has scope over the post-verbal noun *biscotti*. [8]

This test can be applied to the sentences in (1)-(3), with the following effects:

(25) Leo non fece che bene a Giulio.
'Leo only did good to Giulio.'

(26) * Leo non fece che bene l'esercizio.
Leo not did that well the exercise.

(27) ?? Leo non fece che bene a tacere.
Leo not did that well to remain silent.

[7] In French, a mandatory partitive article neatly distinguishes type A from type C: *Ça fait du bien aux enfants* 'Questo fa bene ai bambini', *Il fait bien de partir* 'Fa bene a partire'.
[8] Noteworthy is the fact that, in spite of the presence of *non* 'not', the clause has positive polarity.

The ungrammaticality of (26) and the marginal acceptability of (27) originate from the lexical category to which the two occurrences of *bene* belong. The test confirms that in the structures neither of them behaves as a noun, whilst the well-formed (25) shows that *bene* does behave as a noun, as with *biscotti* 'biscuits' in (24b).

7. Test 7: diminutives, augmentatives, and superlatives

Italian allows a number of derivational suffixes on *bene*. The diminutive suffix *–ino* yields the form *benino*, whilst the augmentative *–one* brings about *benone* (e.g. *Lei non sta benone* 'She is not feeling very well'). The use of such derived forms in (1) to (3) leaves type B perfectly well-formed, as (29) displays, but generates less well-formed sentences with the other types:[9]

(28) ? Leo fece (benino + benone) a Giulio.
'Leo did good to Giulio.'

(29) Leo fece (benino + benone) l'esercizio.
'Leo did the exercise (very) well (enough).'

(30) ? Leo fece (benino + benone) a tacere.
'Leo did (very) well (enough) to remain silent.'

Another derivational suffix *bene* accepts is the superlative morpheme *–issimo* (cf. *stanco* 'tired' vs. *stanchissimo* 'very tired'), which yields *benissimo*. When we replace *bene* with *benissimo* in (1) to (3), the grammaticality outcomes are different:

(31)* Leo fece benissimo a Giulio.
'Leo did very.good to Giulio.'

(32) Leo fece benissimo l'esercizio.
'Leo did the exercise very well.'

(33) Leo fece benissimo a tacere.
'Leo did very well to remain silent.'

These results appear to be fully consistent with the analysis put forward above. The ungrammaticality in (31) can be explained on the grounds that *bene* enters the structure as a noun, and normally nouns do not take the superlative form (though a few of them do: e.g. *governissimo* 'super government', *campionissimo* 'great champion', *veglionissimo* 'New Year's Eve party', *occasionissima!* 'a real bargain!'). On the other hand, the well-formed (32) and (33) suggest that in type B and type C *bene* is not a noun.

8. Test 8: selectional restrictions

A comparison between the three types under scrutiny shows that their subjects are differently constrained. The subject of type A can also be an infinitival clause or a [−animate] noun, as (34) (cf. (4b)) and (35) (cf. (5)) respectively, illustrate:[10]

(34) Prenderti una vacanza ti farà bene.
'It'll do you good to have a holiday.'

(35) La medicina gli fece molto bene.
'The medicine did him a lot of good.'

Turning to type B, its *bene* has no bearing at all on the subject of the sentence (as well as on other elements). The subjects employed in (34) and (35) give origin to ill-formed outcomes when they replace the subject of (2) (**Prenderti una vacanza/La medicina fece bene l'esercizio*), but such results depend on the predicate of the clause (regardless of whether at clause level the predicate is *fare*, thus analyzed as an ordinary (i.e. full) verb, or the noun *esercizio*, if *fare* is analyzed as a support verb).

Finally, in type C the same kinds of subjects bring about ill-formed sentences, which shows that the subject of this clause type must be [+ Human]:

(36)* Prenderti una vacanza farà bene a tacere.
to.take.2SG a holiday will.do well to remain.silent

[9] In (30), the augmentative *benone* probably sounds more natural than the diminutive *benino*. Only the former combination occurs in the Italian *corpus* of *SketchEngine*.

[10] In type A, [−animate] subjects are more common. Indeed, in e.g. (1) the subject *Leo* can be even interpreted as 'the presence of Leo' – thus in (1) *Leo* does not mandatorily have the semantic role >Agent<.

(37)* La medicina fece bene a tacere.
the medicine did well to remain.silent

It is well-known that the constraints displayed by the head of a phrase functioning as an argument depend on its predicate. This amounts to saying that the predicate licensing the subject in type A cannot be the same that licenses the subject in type C, which in turn suggests distinct predicative roles and distinct argument frames for the two occurrences of *fare* in e.g. (1) and (3).

9. Test 9: echo questions

Compare the different outcomes in the assertion-question pairs below:

(38) Speaker 1: La passeggiata fece bene a Giulio.
Speaker 2: La passeggiata fece cosa a Giulio? / *La passeggiata fece come a Giulio?

(39) Speaker 1: Leo fece bene l'esercizio.
Speaker 2: *Leo fece l'esercizio cosa? / Leo fece l'esercizio come?

(40) Speaker 1: Leo fece bene a tacere.
Speaker 2: ??Leo fece cosa a tacere? / *Leo fece come a tacere?

The echo-questions in the replies by speaker 2 show that type A, as (38) illustrates, is the only one in which *bene* allows for the wh-word *cosa* 'what', normally used for noun phrases (e.g. *Vuole una pausa – Vuole cosa?* 'He needs a break – He needs what?'). Thus the test provides extra evidence that only in type A *bene* functions as a noun. The echo-question in (39) containing *come* 'how' confirms that in type B *bene* works as a (manner) adverb. Importantly, the lexical category to which the *bene* in type C belongs remains somewhat unclear, given that neither *cosa* 'what' nor *come* 'how' yield well-formed questions.

10. Test 10: *ne*-cliticization

The partitive clitic *ne* 'of it' targets direct objects (e.g. *Vuole una pausa – Ne vuole una* 'He needs a break – He needs one')[11] and can also be used for direct objects the head of which is uncountable (e.g. *Ha avuto* (*della*) *costanza* 'He had constancy' – *Ne ha avuta* 'He had some'). The test shows that only the *bene* of type A allows for such a clitic:

(41) La passeggiata fece (tanto) bene a Giulio – La passeggiata ne fece (tanto) a Giulio.

(42) Leo fece l'esercizio bene - *Leo ne fece l'esercizio.

(43) Leo fece bene a tacere – ??Leo ne fece a tacere.

This test too, therefore, proves that a difference exists between type A, whose *bene* functions as a noun, and the other types, in which *bene* functions as an adverb.

11. Concluding remarks

HASER and KORTMANN list the following morphosyntactic properties « as unifying characteristics of all or at least the prototypical adverbs in English » (2006: 67):

(i) adverbs are invariable;
(ii) adverbs are optional;
(iii) adverbs can be modified by items such as *very* or *quite*;
(iv) adverbs are used as modifiers of categories other than nouns.

In type B, *bene* conforms to the criteria from (ii) to (iv): it is optional, it can be modified by e.g. *molto* or *parecchio*, and does not modify a noun. As for criterion (i), *bene* (as well as other adverbs, e.g. *subito* 'soon, at once', which gives rise to the emphatic and informal *subitissimo*) differs from English prototypical adverbs in that it can be modified (see section 7), but such modifications take place by means of derivational, rather than inflectional, morphemes. Another property of type B *bene* to be added to the above list is its free position. We believe there is no reason to further pursue the examination of type B inasmuch as a large body of evidence suggests that *bene* functions as a common manner adverb.

Also the lexical category of the *bene* in type A appears to be adequately clear-cut. The tests provided above consistently show that this *bene* works as a noun (*contra* the classification found in e.g. the dictionary *Il grande italiano 2008*).

By choosing a number of pertinent criteria, shown in table 1 below, we can see how the occurrences of *bene* in type A and type B greatly differ. They take opposite values in all the criteria:

[11] See Rosen 1988.

	optional	free position	replaced by a *-mente* adverb	multi-word quantifier	superlative
Type A	−	−	−	+	−
Type B	+	+	+	−	+

Table 1: Five criteria to distinguish type A from type B

Lastly, the insertion in a lexical category of the *bene* in type C is not as straightforward as with the other types. This *bene* aligns with type A in the first three criteria: its presence is mandatory, its unmarked position is post-verbal, and no replacement with an adverb ending in *–mente* is possible. However, in type C *bene* cannot take a multi-word quantifier (cf. (20)) and accepts the superlative suffix. The latter criteria suggest treating it as an adverb. A crucial criterion in table 1 is the optional presence of *bene*. HASER and KORTMANN (2006: 67) briefly discuss cases of mandatory adverbs (*The job paid us handsomely* vs. **The job paid us*).[12] In Italian there are other cases in which *bene* is mandatory: e.g. *Comportati bene!* 'Behave yourself!', **Comportati!*; *Mario sta bene*, **Mario sta*.[13] In these examples, *bene* is free with regard to position (e.g. *Mario bene sta!*, which needs a marked intonation pattern), it can be replaced by a *–mente* adverb (*Mario sta egregiamente* 'Mario feels very well'), multi-word quantifiers are impossible (**Mario sta un gran bene*, **Mario sta un sacco di bene*, but notice the well-formed – and informal – *Mario sta un sacco bene*), and, finally, superlative forms are allowed (*Mario sta benissimo*). The different values of the criteria in table 1 that *bene* takes in these cases show that this area of Italian as well as of other languages is problematic and in need of further research.

12. References

BOCH, R. (1995, third edition), *il Boch. Dizionario francese italiano, italiano francese*. Zanichelli, Bologna.

BROWN, K. (2006), (chief ed.) *Encyclopedia of Language and Linguistics*, 14 volumes. Elsevier, Oxford.

BUSSMANN, H. (1996), *Routledge Dictionary of Language and Linguistics*. Routledge, London/New York.

DEVOTO, G. ; G. C. OLI. *Il dizionario della lingua italiana*. Edizione 2000-2001, Le Monnier, Milano.

GABRIELLI, A. (2007), *Il grande italiano 2008*. Hoepli, Milano.

HASER, V. ; B. KORTMANN (2006), *Adverbs*, in Brown 2006, vol. I, 66-69.

LANGUAGE FILES (1991, fifth edition), M. Crabtree and J. Powers (compilers), Ohio State University Press, Colombus.

LEPSCHY, A. L. ; G. LEPSCHY (1981), *La lingua italiana, storia, varietà dell'uso, grammaticale*. Bompiani, Milano.

LONZI, L. (1991), *Il sintagma avverbiale*, in Renzi, L. and G. Salvi (eds.), *Grande grammatica italiana di consultazione*, Vol. II, il Mulino, Bologna, 341-412.

MAROTTA, G. (1989), Avverbio, in *Dizionario di linguistica*, a cura di G. Beccaria, Einaudi, Torino, 109-111.

PICCHI, F. (1999), *Il grande inglese 2008*. Hoepli, Milano.

RAGAZZINI, G. (1995), *Il Ragazzini. Dizionario inglese italiano, italiano inglese*. Zanichelli, Bologna.

ROSEN, C. (1988), *The Relational Structure of Reflexive Clauses. Evidence from Italian*. Garland, New York.

SENSINI, M. (1997), *La grammatica della lingua italiana*. Arnoldo Mondadori Editore, Milano.

SERIANNI, L. (1989), *Grammatica italiana. Italiano comune e lingua letteraria*. Utet, Torino.

[12] In these examples the subject of the verb *to pay* is [− animate], as happens in *The book reads very awkwardly*.
[13] The sentence *Mario sta* is grammatical under a distinct meaning (e.g. in card games, *Sto!* 'Stick!').

Ainsi. Deux emplois complémentaires d'un adverbe type[1]

Christian Molinier

0. Introduction

La forme adverbiale *ainsi* figure dans de nombreuses constructions syntaxiques du français. On la rencontre notamment dans la fonction d'adverbe de manière auprès d'un verbe (*Ainsi vivaient nos ancêtres*), fonction dans laquelle elle est souvent considérée comme le substitut général de tout adverbial de manière[2], dans la fonction d'adverbe de phrase connecteur (*Ainsi, ces hommes vivaient dans le plus complet dénuement*), et comme marqueur de corrélation dans un système comparatif (*Comme la vue d'un portrait suggère à l'observateur l'impression d'une destinée, ainsi la carte de France révèle notre fortune* (Ch. de Gaulle))[3]. On la trouve également dans la locution conjonctive *ainsi que*, laquelle est soit locution conjonctive de subordination (*Ainsi que disent les hommes, il bat la campagne* (M. du Camp)), soit une locution conjonctive de coordination (*J'ai rencontré Luc ainsi que Paul ; Ma conversion fut le résultat de ces faits, ainsi que d'autres d'un caractère plus secret* (J. Green)), ou encore dans des structures figées (*pour ainsi dire, ainsi de suite, c'est ainsi, ainsi soit-il*, etc.).

La présente étude est consacrée à *ainsi* adverbe de manière ou adverbe de phrase. Dans chacun de ces deux cas, *ainsi* joue un rôle pronominal, *i.e*. de substitut, renvoyant ou bien au contexte discursif, ou bien à la situation de communication. Nous examinons successivement dans les pages qui suivent l'adverbe de manière et l'adverbe de phrase, tant d'un point de vue syntaxique que sémantique.

1. *Ainsi* adverbe pronominal de manière

Ainsi adverbe pronominal de manière est un modifieur du verbe, à référence contextuelle ou déictique, généralement paraphrasable par « *de cette façon / manière* ». Il apparaît soit en construction postverbale, soit en construction initiale et détermine dans ce dernier cas une organisation particulière des mots de la phrase, liée à des effets de sens. Nous examinons dans un premier temps les propriétés générales de l'adverbe (1.1.), et dans un deuxième temps la construction à adverbe initial (1.2.).

1.1. Propriétés générales

1.1.1. Propriétés syntaxiques de *ainsi* adverbe de manière verbal

Considérons les exemples suivants :

(1) *En Rhône-Alpes, A.B. a dû affronter des ignominies de la part de son propre camp. Comment des démocrates peuvent-ils se comporter ainsi ? (w.w.w.)*[4]

(2) *Sur le terrain de foot, alors que mon équipe perdait, et que je venais de rater une belle reprise de volée, j'entendis une rumeur que je crus à moi destinée. « Quand même, il peut faire cela en France, mais là, il n'est pas chez lui ». Ceux qui parlaient ainsi n'étaient pas des Nord-Irlandais, mais des continentaux (w.w.w.)*

(3) *Namer (également nommé Ménès), unifie la Haute et la Basse Egypte. Il donne ainsi naissance à la première dynastie des pharaons (w.w.w.)*

Ainsi présente dans ces exemples l'ensemble des propriétés caractéristiques d'un adverbe de manière verbal :
- Il est compatible avec la focalisation dans une phrase interrogative :

Les démocrates se comportent-ils ainsi ?

Les Nord-Irlandais ont-ils parlé ainsi ?

Donne-t-il naissance ainsi à la première dynastie des pharaons ?

dans une construction en *C'est … que* :

---

[1] Je remercie vivement Francis Cornish et Takuya Nakamura pour leurs nombreuses remarques critiques qui m'ont permis d'améliorer mon texte.

[2] Cf. M. Gross (1975 : 78).

[3] Dans cet exemple, la forme *comme* de la première proposition est corrélée à la forme *ainsi* de la seconde. Mais la corrélation peut aussi être marquée par *ainsi* dans les deux propositions (*Ainsi le rossignol n'a qu'à parler, sa voix Fait taire autour de lui tous les oiseaux des bois ; Ainsi le doux passé plein de mélancolie Fait taire le présent à l'arsouille* (Jean Richepin, *in TLF*)). Ce dernier tour est jugé archaïque.

[4] Les exemples du langage quotidien sont tirés pour la plupart du World Wide Web, source signalée par w.w.w., sans autres précisions, jugées inutiles. Le lecteur peut d'ailleurs retrouver facilement auteur et contexte, si bon lui semble.

C'est ainsi que se comportent les démocrates

C'est ainsi qu'ont parlé les Nord-Irlandais

C'est ainsi qu'il donne naissance à la première dynastie des pharaons

dans une construction négative :

Les démocrates ne se comportent pas ainsi

Les Nord-Irlandais n'ont pas parlé ainsi

Il ne donne pas naissance ainsi à la première dynastie des pharaons

- Il entre en distribution complémentaire avec un adverbe de manière dans l'emploi illustré par les exemples (1) et (2) :

Les démocrates se comportent (mal + bien + intelligemment + …)

Les Nord-Irlandais ont parlé (clairement + confusément + sans crainte + …)

et avec une construction gérondive de manière dans l'emploi illustré par l'exemple (3) :

Il donne naissance à la première dynastie des pharaons en faisant cela

- il se laisse paraphraser par *de cette (façon + manière)*, propriété qui souligne sa valeur pronominale :

Les démocrates se comportent de cette (façon + manière)

Les Nord-Irlandais ont parlé de cette (façon + manière)

Il donne naissance de cette (façon + manière) à la première dynastie des pharaons

Tout en assumant sa fonction d'adverbe de manière verbal, *ainsi* peut se rencontrer en lieu et place d'un groupe nominal attribut de l'objet, en particulier avec des verbes tels que *appeler*, *nommer*, etc. :

(4) *Théorèmes belges. On appelle ainsi les théorèmes de Dandelin et Quételet, deux illustres mathématiciens belges du 19ème siècle* (w.w.w.)

Ainsi tient lieu ici d'un groupe nominal attribut de l'objet du verbe *appeler* (les théorèmes de Dandelin et Quételet). Mais il partage cette fonction avec celle de complément adverbial de manière dans la mesure où il est focalisable dans une phrase interrogative, une construction en *C'est … que*, ou une phrase négative :

Appelle-ton ainsi les théorèmes de Dandelin et Quételet ?

C'est ainsi que l'on appelle les théorèmes de Dandelin et Quételet

On n'appelle pas ainsi les théorèmes de Dandelin et Quételet

et où il est en distribution complémentaire avec l'adverbe interrogatif *Comment*, le complément adverbial de manière *de cette (façon + manière)* ou encore la forme adverbiale *comme cela* :

On appelle (comment + de cette (façon + manière) + comme cela) les théorèmes de Dandelin et Quételet

De même, à la suite d'un verbe d'état, pour référer à une attitude du sujet précédemment indiquée, *ainsi* peut se trouver en position d'attribut du sujet et assumer des propriétés d'adverbe de manière :

(5) *Il s'est enfoncé dans mon canapé comme s'il n'avait plus la force de parler. Il demeura quelques minutes ainsi, immobile et muet* (w.w.w.)

1.1.2. Propriétés référentielles de *ainsi* adverbe de manière verbal

Dans chacun des exemples (1)-(5), *ainsi* est un adverbe pronominal anaphorique. Il réfère en effet à un contenu identifiable dans le contexte gauche de la phrase où il figure, soit en (1), à des ignominies de la part du propre camp de A.B., en (2), à des paroles citées, en (3) à l'unification de la Haute et de la Basse Egypte, en (4), au groupe nominal « théorèmes belges », en (5), à une attitude du sujet décrite dans la phrase précédente. Mais *ainsi* peut aussi être un adverbe pronominal cataphorique, comme dans les exemples(6)-(8) :

(6) *On procédera ainsi : les élèves liront tel épisode pour tel jour. Ils devront le lire attentivement pour être capables de restituer l'histoire* (w.w.w.)

(7) *Pourtant tourné aux Etats-Unis avec des acteurs américains parmi les plus connus, produit par UGC et censé être distribué par Warner, le film n'eut pas droit à une seule salle aux Etats-Unis. Les critiques américaines en parlaient ainsi à sa sortie : « All of it is a bizarre and badly acted bore, guaranteed to clear theatres in record time »* (Boxoffice Magazine, *in* w.w.w.)

(8) *L'enseignement devrait être ainsi : celui qui le reçoit le recueille comme un don inestimable, mais jamais comme une contrainte pénible* (w.w.w.)

Enfin, *ainsi* peut être un adverbe pronominal déictique. Son référent figure en ce cas dans la situation de communication. Il est produit au moment même de l'énonciation de *ainsi* dans un acte qui associe le geste à la parole :

(9) *Ainsi font, font, font, les petites marionnettes* (Chanson enfantine qui s'interprète avec des mouvements des mains)

(10) *Ainsi périsse quiconque franchira mes murailles* (Tite-Live, Ab Urbe Condita, I, 7. Paroles proférées par Romulus au moment où il tue son frère Rémus)[5]

(11) *Le guerrier se penche pour ramasser les armes et Clovis en profite pour lui asséner un coup de sa propre hache sur la tête : « Ainsi as-tu fait à Soissons avec le vase », lui dit-il* (w.w.w.)

1.2. *Ainsi* adverbe pronominal de manière en construction initiale

1.2.1. Contraintes syntaxiques générales propres à la construction

Le positionnement en tête de phrase de *ainsi* impose à la phrase des contraintes syntaxiques particulières que nous allons passer en revue :

- Le sujet apparaît obligatoirement à droite du verbe. On désigne traditionnellement cette construction du nom d' « inversion simple du sujet » :

(12) *Sous la Restauration, beaucoup d'orateurs préparaient leurs discours à l'avance en faisant de véritables ouvrages qui étaient imprimés sous forme de petits factums. Ainsi travaillait Royer-Collard* (w.w.w.)

Le positionnement du sujet à gauche du verbe (absence d'inversion) exclut l'interprétation adverbe de manière pour *ainsi*. Considérons la phrase suivante :

(13) *Ainsi, Pierre travaillait*

Son adverbe s'interprète comme un adverbe de phrase récapitulatif. Il introduit un jugement de synthèse à partir de faits que le locuteur vient d'exposer. Dans cette fonction, *ainsi* est compatible avec la construction dite parfois « inversion complexe » (groupe nominal sujet à gauche du verbe, pronom clitique à sa droite)[6] :

(14) *Ainsi, Pierre travaillait-il*

- Le verbe n'est pas admis à la forme négative :

(15)**Ainsi ne travaillait pas Pierre*

Comme tout adverbe de manière, *ainsi* ne peut figurer qu'à droite du verbe lorsque celui-ci est à la forme négative, et il est alors le focus de la négation :

(16) *Pierre ne travaillait pas ainsi*

Cette situation rappelle celle des adverbes de manière en général (cf. Ch. Molinier et F. Lévrier 2000 : 117) :

(17) **(Méthodiquement + Soigneusement), Pierre ne travaillait pas*

(18) *Pierre ne travaillait pas (méthodiquement + soigneusement)*

Remarquons que *ainsi* adverbe de phrase est lui, comme tout adverbe de phrase, parfaitement compatible avec la forme négative du verbe :

(19) *Ainsi, Pierre ne travaillait pas*

- Le *GN* postposé relève des principales catégories admises dans la fonction sujet. Il peut être un nom propre, comme en (12) ci-dessus ou un groupe substantival comme dans l'exemple suivant :

(20) *Jean-Pierre Benda ne sortira pas de prison en attendant son procès qui s'ouvrira le 27 avril 2010. Ainsi en ont décidé les cinq juges composant la chambre d'appel de la cour pénale internationale* (w.w.w.).

[5] Le texte original de Tite-Live est : « Sic deinde pereat quicumque transiliet moenia mea ». Notons au passage l'ordre identique des mots.

[6] Cette formule nous paraît malheureuse dans la mesure où il n'y a pas d'inversion du tout.

Il peut être aussi un pronom de première personne :

(21) *Ma haine pour Karl – ainsi avais-je baptisé le jeune officier allemand – décupla* (J. Laurent, in B. Jonare 1976 : 132),

un pronom de deuxième personne comme dans l'exemple (11) repris en (22) :

(22) *Le guerrier se penche pour ramasser les armes et Clovis en profite pour lui asséner un coup de sa propre hache sur la tête : « Ainsi as-tu fait à Soissons avec le vase », lui dit-il* (w.w.w.),

le pronom indéfini *on* :

(23) *Lorsque la pression est très faible, lorsqu'il n'y a que quelques centaines de molécules ou d'atomes par mètre cube, la condensation peut se faire directement sous forme solide. Ainsi prépare-t-on en laboratoire certains composés ultra-purs d'aluminium ou de titane* (Cl. Allègre, *in* Cl. Guimier 1997 : 88)

le pronom impersonnel *il* :

(24) *La condensation du gaz peut se faire directement sous forme solide. Ainsi en va-t-il dans le cosmos* (Cl. Allègre, in Cl. Guimier 1997 : 88)

enfin le pronom personnel *il*, sous certaines conditions[7] :

(25) *Riche, heureux, adulé à son tour, jouissant de toutes les voluptés, gourmand, débauché, ainsi vivait-il à Venise, honoré de tous et ayant Le Titien pour ami intime* (J. Richepin)

(26) *Il s'inclinait, s'agenouillait, posait son front contre l'étoffe, se redressait, se prosternait à nouveau. Quand sa prière était terminée, il pliait soigneusement son tapis et le rangeait. Ainsi procédait-il à l'aube, à midi, l'après midi, au coucher du soleil et tard dans la soirée afin que les paroles des sourates du Coran correspondent exactement au rythme de la nature* (Mohed Altrad)

(28) *« Nous avons un staff de quatre entraîneurs qui vont diriger les Lions Indomptables de football ». Ainsi parlait-il, hier, au sujet de la fédération du quartier Tsinga* (w.w.w.)

(29) *« C'est moi seul qui suis misérable à ce point. J'ai une femme, bonne, et je cours après une autre ». Ainsi pensait-il, assis dans la cabane, où, à un seul endroit, l'eau coulait* (w.w.w., page Tolstoï, *Le faux coupon et autres contes*)

L'examen des exemples fournis par le Web fait apparaître que la construction *ainsi* adverbe de manière en position initiale et le pronom personnel clitique *il* postposé au verbe se rencontre avec des verbes qui se construisent avec un complément obligatoire, comme en (25)-(26), ou bien avec des verbes intransitifs ou transitifs employés intransitivement, comme en (28)-(29). On constate à l'inverse que les phrases où figurent le pronom clitique *il* préposé au verbe et l'adverbe de manière *ainsi* postposé à lui ne permettent pas la permutation de cet ordre si leur verbe ne présente pas la structure spécifique indiquée. Soit les phrases attestées suivantes :

(30) *L'empereur Constantin convoque le tout premier concile œcuménique à Nicée dans le but d'établir l'unité de l'Eglise en Orient comme en Occident. Il espère ainsi mettre fin au conflit causé par l'Arianisme, qui nie la nature divine du Christ* (w.w.w.)

(31) *Monté sur le trône quelques années plus tôt, le jeune Ramsès II livre une bataille mémorable à Kadesh contre les envahisseurs Hittites. Il souhaitait ainsi récupérer les terres d'Afrique et d'Asie Mineure* (w.w.w.)

La postposition du pronom sujet *il* dans ces exemples, avec antéposition de *ainsi*, semble induire quasi-obligatoirement l'interprétation adverbe de phrase récapitulatif pour *ainsi* :

(30') *L'empereur Constantin convoque le tout premier concile œcuménique à Nicée dans le but d'établir l'unité de l'Eglise en Orient comme en Occident. Ainsi espère-t-il mettre fin au conflit causé par l'Arianisme*

(31') *Monté sur le trône quelques années plus tôt, le jeune Ramsès II livre une bataille mémorable à Kadesh contre les envahisseurs Hittites. Ainsi souhaitait-il récupérer les terres d'Afrique et d'Asie Mineure*[8]

[7] La structure à inversion locative exclut normalement la présence d'un pronom clitique sujet, comme le note N. Fournier (1997 : 97) : « Dans ce type de phrase, la postposition n'est possible que pour le sujet nominal, elle est impossible pour le clitique ». L'exclusion du pronom personnel clitique *il* dans la structure à inversion locative est liée à la fonction discursive très généralement reconnue à celle-ci : celle d'introduire un référent nouveau, ou de réintroduire un référent, constituant ou partie du rhème de l'énoncé. Il y a donc là un point de divergence entre *ainsi* et *ici*, *alors*, etc. cf. *infra* 1.2.3.

On en conclut que *ainsi* placé en tête de phrase, avec *il* pronom personnel sujet postposé, s'interprète prioritairement comme adverbe de phrase, *i.e.* adverbe à statut périphérique, et comme adverbe de manière seulement si la structure propre au verbe, ou la construction absolue du verbe dans un contexte particulier, exigent un complément adverbial. Le pronom personnel *il* a donc un fonctionnement à part dans le cas de l'inversion du sujet.

1.2.2. Ordre des mots à la droite du verbe

L'ordre des mots apparaissant à droite du verbe est soumis à des contraintes particulières, analogues à celles que l'on observe dans le cas du *GN* sujet pour les phrases à complément locatif initial (*Au fond de la vallée coulait une rivière*), ou pour les phrases interrogatives ou relatives (*Quand viendra votre ami* ? ; *Le jour où viendra votre ami*, cf. R.S. Kayne 1973)[9]. Ces contraintes concernent les situations où le verbe régit un ou plusieurs compléments et où le sujet est nominal.

En règle générale, le *GN* sujet postposé apparaît à la fin de la phrase, notamment quand le verbe est suivi d'un complément direct ou d'un attribut:

(32)**Ainsi analysent les sociologues ce phénomène* [10]

(33) *?ainsi analysent ce phénomène les sociologues*

L'allongement du *GN* sujet rend la phrase (33) parfaitement grammaticale :

(34) *Ainsi analysent ce phénomène les sociologues contemporains de l'école de Chicago*[11]

De la même manière, les compléments indirects étroitement dépendants du verbe semblent généralement exiger la position à droite du verbe et le rejet du sujet en fin de phrase :

(35) *?*Ainsi parlait le laboureur à ses enfants*

(36) ?*Ainsi parlait à ses enfants le laboureur*

Là encore, un sujet long (en fin de phrase) améliore la phrase :

(37) *Ainsi parlait à ses enfants le riche laboureur sentant sa mort prochaine*

Lorsque le verbe est accompagné d'un complément à statut périphérique, celui-ci peut apparaître aussi bien à la droite immédiate du verbe, laissant ainsi la position finale au sujet, qu'en fin de phrase à la suite du sujet :

(38) *Ainsi vivaient (au Paléolithique supérieur + dans la corne de l'Afrique) nos lointains ancêtres*

(39) *Ainsi vivaient nos lointains ancêtres (au Paléolithique supérieur + dans la corne de l'Afrique)*

1.2.3. Structure de l'information

Ainsi positionné en tête de phrase est nécessairement anaphorique (ou déictique). D'une part, cette fonction est observée dans l'ensemble des exemples attestés rencontrés, cf. : (12), (20), (21), etc. D'autre part, *a contrario*, la position en tête de phrase et la fonction cataphorique paraissent incompatibles[12]. Ainsi, les exemples ci-dessous construits à partir des exemples (4)-(6) –dans lesquels *ainsi* est cataphorique – en déplaçant l'adverbe en tête de phrase, paraissent déviants :

(4') *?Ainsi procédera-t-on : Les élèves liront tel épisode pour tel jour. Ils devront le lire attentivement pour être capables de reconstituer l'histoire*

(5') *?Ainsi en parlaient à sa sortie les critiques américaines : « All of it is a bizarre and badly acted bore »*

---

[8] L'interprétation adverbe de manière (« de cette manière ») exigerait un fort détachement intonatif dans la phrase.

[9] L'inversion du sujet dans de telles constructions est appelée inversion stylistique (INV-STYL). Elle s'oppose à l'inversion clitique sujet (INV-CLIT-SUJ) qui relève d'une autre analyse.

[10] Cette construction, jugée impossible aujourd'hui, était parfaitement correcte dans l'ancienne langue, et ce jusqu'au 16ème siècle, cf. : *Ainsi faisaient aucuns chirurgiens de Grèce les opérations de leur art sur des eschauffaux à la veuë des passans* (Montaigne, *in* Riegel, M., Pellat, J.-C. et Rioul, R., 2009 : 251).

[11] Les faits sont comparables à ceux que décrit R. S. Kayne (1973 : 121) :

**Quand deviendra ce comédien célèbre ?*
?Quand deviendra célèbre ce comédien ?
Quand deviendra célèbre ce comédien que nous avons vu si bien jouer hier soir à la télévision ?

[12] Remarquons que l'indéfini *tel*, qui entre lui aussi dans une construction à inversion du sujet, paraît moins contraint que *ainsi*. *Tel* dans *Telle était la situation au début du siècle* renvoie normalement à une situation évoquée, mais peut aussi renvoyer à une situation que l'on va évoquer, cf. Grevisse-Goosse § 246, 1.

(6') ?*Ainsi devrait être l'enseignement : celui qui le reçoit le recueille comme un don inestimable, mais jamais comme une contrainte pénible*

Dans cette même position, *ainsi* a statut thématique, en ce qu'il constitue le point de départ de l'énoncé, le support de l'information appelant un développement à sa suite qui constituera ou comportera le rhème, *i.e*. l'apport d'information. Une telle structure est donc comparable à la structure dite « à inversion locative » que l'on observe dans les deux exemples suivants :

(40) *Ici commence le court bonheur de ma vie* (J.-J. Rousseau)

(41) *Alors se produisit un incident étrange et véritablement surnaturel* (J. Verne)

Selon R. Le Bidois (1952 : 351), dans une telle configuration, « le terme placé en tête de phrase constitue le cadre de l'action, c'est-à-dire la circonstance de temps ou de lieu qui est le point de départ de l'énoncé. » F. Cornish (2005 : 164) caractérise précisément cette construction par les traits suivants : « the presence in clause-initial position of a locative or temporal adverbial (…); a locative or existential non-predicating verb whose inherent semantics corresponds or is reduced to this type of denotation, and which is tightly connected syntactically and semantically to the preposed constituent; and a postposed rhematic subject term occurring in predicate focus position. » Cependant, de claires différences apparaissent. Le pronom personnel *il*, exclu dans la construction à inversion locative, est possible dans la construction manière-*ainsi* :

(42) **Ici commence-t-il (le court bonheur de ma vie)*

(43) *Ainsi commence-t-il (le court bonheur de ma vie)*

La construction à inversion locative sert à introduire un référent nouveau ou à réactiver un référent (*Ici commence le court bonheur de ma vie / Ici commence ce court bonheur de ma vie*). Rien de tel avec la construction à inversion manière-*ainsi*, qui ne semble pas privilégier le sujet dans l'apport d'information comme le fait la construction à inversion locative.

2. *Ainsi* adverbe de phrase récapitulatif

2.1. Définition

Ainsi adverbe de phrase récapitulatif établit un lien discursif entre l'énoncé dans lequel il figure et le contexte gauche. Il a fondamentalement pour objet de poser la conformité de l'énoncé qu'il accompagne avec un état de choses décrit précédemment. Exemples :

(44) *A l'époque de ces sociétés primitives, les rôles de chaque sexe étaient clairement définis, et chacun savait ce qu'il avait à faire. Ainsi les hommes chassaient et les femmes gardaient la caverne et les enfants* (w.w.w.)

(45) *Pour les femmes comme pour les hommes, le niveau de formation conditionne largement le risque de chômage. Ainsi, les femmes n'ayant pas le baccalauréat sont majoritaires dans les demandes d'emploi féminines* (w.w.w.)

(46) *Ptolémée tente d'expliquer mathématiquement les mouvements des planètes, de la lune et du soleil. Ainsi il établit que les planètes du système tournent en formant un petit cercle qu'il appelle épicycle* (w.w.w.)

(47) *Le site propose assez souvent des ventes de la collection Freeman Porter avec des réductions pouvant aller jusqu'à 70%. Ainsi, il est possible d'acheter un jean de la marque à moindre frais (20 euros par exemple)* (w.w.w.)

(48) *De ce périple, il ramène un certain goût pour l'exotisme, thème prégnant de son œuvre, et une attirance pour les femmes « typées ». Ainsi, dès son retour en France, Baudelaire s'éprend de Jeanne Duval en 1842, une métisse dont il partage jusqu'à la fin la vie erratique, et qu'il érigera comme la « Vénus noire » de son œuvre, comme l'incarnation de la femme exotique et sensuelle* (w.w.w.)

(49) *Je n'ai pas proposé à M. de Narbonne de dîner ici. Ainsi, tu es maître de dire ce qui te convient* (G. de Staël, Correspondance))

Des effets de sens découlent de cette valeur première de conformité exprimée par *ainsi* : illustration ou validation d'un principe général (cf.(44)-(45)), explicitation et précisions données à une assertion (cf. (46)-(47)), conséquence d'une situation particulière (cf.(48)-(49)), conclusion ou généralisation tirée d'observations, comme c'est le cas dans les exemples (50)-(51) fabriqués en inversant les propositions constitutives de (44)-(45) :

(50) *A l'époque de ces sociétés primitives, les hommes chassaient et les femmes gardaient la caverne et les enfants. Ainsi, les rôles de chaque sexe étaient clairement définis, et chacun savait ce qu'il avait à faire*

(51) *Les femmes n'ayant pas le baccalauréat sont majoritaires dans les demandes d'emploi féminines. Ainsi, pour les femmes comme pour les hommes, le niveau de formation conditionne largement le risque de chômage.*

2.2. Positions respectives de *ainsi* et des autres constituants de la phrase

Ainsi peut figurer en position initiale comme c'est cas dans les exemples (43)-(49) ci-dessus. Il peut aussi figurer à l'intérieur de la phrase, comme dans les exemples suivants, où il est postverbal :

(52) *Le film est articulé sur la notion de triangle. Axel est ainsi partagé entre l'amour pour Grace et celui pour Elaine* (w.w.w.)

(53) *Parler créole est particulièrement répandu à La Réunion, bien plus que dans les autres départements d'Outre-Mer. La grande majorité des Créoles ne parlaient ainsi que créole durant leur enfance* (w.w.w.)

(54) *Bien évidemment, la publication sur Internet permet également de dépasser le cadre scolaire. Etymon s'adresse ainsi à toute personne curieuse de remonter aux origines de sa langue* (w.w.w.)

Notons qu'il doit figurer immédiatement à droite du verbe, la position à droite d'un complément du verbe étant interdite. Considérons l'exemple suivant, démarqué de (53) :

(55) *La grande majorité des Réunionnais parlaient le créole ainsi*

On voit que l'adverbe *ainsi* ne peut être dans ce cas qu'adverbe pronominal de manière, il signifie uniquement « de cette manière », alors qu'à la droite immédiate du verbe il peut recevoir, suivant le contexte, soit l'interprétation adverbe de manière pronominal, soit l'interprétation adverbe de phrase récapitulatif.

On remarquera que l'adverbe de phrase *ainsi* postposé pourrait toujours aussi bien figurer en tête de phrase, l'interprétation restant identique :

(52') *Le film est articulé sur la notion de triangle. Ainsi Axel est partagé entre l'amour pour Grace et celui pour Elaine*

(53') *Parler créole est particulièrement répandu à La Réunion, bien plus que dans les autres départements d'Outre-Mer. Ainsi la grande majorité des Créoles ne parlaient que créole durant leur enfance*

(54') *Bien évidemment, la publication sur Internet permet également de dépasser le cadre scolaire. Ainsi Etymon s'adresse à toute personne curieuse de remonter aux origines de sa langue*

Comme on l'a dit plus haut (1.3.), lorsque *ainsi* est positionné en tête de phrase et que le sujet substantival figure à gauche du verbe (absence d'inversion), l'adverbe a nécessairement l'interprétation adverbe de phrase, cf. (13) repris en (56) :

(56) *Ainsi, Pierre travaillait*

C'est aussi le cas si le sujet est un pronom personnel :

(57) *Ainsi, il travaillait*

Au contraire, si le sujet est postposé au verbe, l'adverbe a nécessairement l'interprétation adverbe de manière :

(58) *Ainsi travaillait Paul*

Mais lorsque *ainsi* est positionné en tête de phrase et que le sujet est le pronom personnel *il* figurant à droite du verbe (inversion), cf. :

(59) *Ainsi travaillait-il*

l'adverbe peut recevoir l'interprétation adverbe de manière (cf. 1.2.1.), ou adverbe de phrase. Cependant, si le verbe est suivi d'un adverbe modifieur du verbe, cet adverbe bloque l'interprétation adverbe de manière pour *ainsi*, qui est alors interprété comme adverbe de phrase :

(60) *Ainsi travaillait-il jour et nuit pour nourrir sa nombreuse famille*

(61) *Il élidait les consonnes jusqu'à produire un continuum de voyelles plus ou moins expressif. Ainsi travaillait-il sans relâche à cette grande suppression des consonnes* (w.w.w.)

Il s'agit là d'un principe d'ordre que l'on peut observer ailleurs et notamment avec les compléments locatifs, comme le relève N. Fournier (1997 : 102). Ainsi dans l'exemple à inversion locative suivant, le complément initial est nécessairement le complément essentiel du verbe :

(62) *Un peu plus loin dans le virage repose une vieille carcasse de voiture rouillée*

Mais dans l'exemple suivant, muni de deux compléments locatifs, c'est le complément qui suit le verbe qui est complément essentiel, tandis que le premier joue le rôle de complément périphérique de cadre :

(63) *Un peu plus loin dans le virage repose au milieu des branchages gelés une vieille carcasse de voiture rouillée*

Enfin, *ainsi* positionné en tête de phrase peut donner lieu à l'inversion dite complexe –groupe nominal placé à gauche du verbe, pronom clitique à sa droite :

(64) *Le commissaire Juve, distrait, ne remarqua pas que Fantômas volait discrètement une épingle sur le bureau avec laquelle il ne tarderait pas à se libérer. Ainsi le plus grand criminel de l'époque put-il se libérer une nouvelle fois* (P. Souvestre et M. Allain)

Généralement, les deux constructions : absence de pronom clitique rappelant le sujet et présence du pronom clitique (inversion complexe) semblent également possibles. La phrase (64) ci-dessus amputée du pronom de rappel *il* nous paraît aussi normale que la phrase authentique et nous ne voyons aucune différence de sens entre les deux. Il est possible que l'inversion complexe ait pour effet de relier plus étroitement les deux énoncés articulés par *ainsi* et soit de ce fait plus naturelle dans le cas d'un lien de cause à conséquence, comme en (64). A l'inverse, le *ainsi* d'explicitation, de développement semblerait moins propice à l'inversion complexe :

(65) *A l'époque de ces sociétés primitives, les rôles de chaque sexe étaient clairement définis, et chacun savait ce qu'il avait à faire. Ainsi les hommes chassaient et les femmes gardaient la caverne et les enfants*

vs

(66) *?A l'époque de ces sociétés primitives, les rôles de chaque sexe étaient clairement définis, et chacun savait ce qu'il avait à faire. Ainsi les hommes chassaient-ils et les femmes gardaient-elles la caverne et les enfants*

3. Conclusion

La forme adverbiale *ainsi* présente, comme un certain nombre d'autres formes adverbiales, deux grandes classes d'emploi complémentaires : ou bien sa portée est le verbe et *ainsi* est dans ce cas un adverbe de manière de ce verbe, ou bien sa portée est la phrase et dans ce cas *ainsi* est un adverbe de phrase ayant une fonction de connecteur. Les propriétés définitionnelles générales des adverbes de manière et des adverbes de phrase sont vérifiées : focus de l'interrogation, de la négation et admission dans la structure en *C'est …que*, impossibilité de figurer en tête de phrase négative, *etc*. pour l'adverbe de manière, absence de tout lien avec le verbe, possibilité de figurer en tête de phrase négative, *etc*. pour l'adverbe de phrase connecteur.

Dans les deux cas, la signification basique de *ainsi* est celle de conformité : conformité d'une manière de faire ou d'être avec une autre manière de faire ou d'être (*Luc a toujours travaillé avec minutie et précision. Or, Jean (ne) travaille (pas) ainsi*), conformité du contenu d'un énoncé avec un état de choses présenté précédemment (*A cette époque les rôles des deux sexes étaient clairement définis. Ainsi, les hommes chassaient et les femmes gardaient la caverne et les enfants*). Dans les deux cas, *ainsi* a statut de proforme. *Ainsi* adverbe de manière se paraphrase par *de cette façon/manière* et il peut toujours se substituer à un adverbe de manière ou à un complément adverbial de manière. *Ainsi* adverbe de phrase connecteur équivaut à *conformément à cet état de choses*.

Nous avons étudié une particularité de construction de *ainsi* adverbe de manière, liée à des effets de sens : la possibilité de figurer en tête de phrase en imposant la postposition du sujet. Cette construction présente de grandes similarités, concernant l'ordre des mots, avec la construction dite « à inversion locative » (*Au fond de la forêt vivait un vieil ermite*). Au-delà de leur parenté, des différences apparaissent. Ainsi le pronom personnel *il*, interdit dans la construction locative du fait que cette construction sert principalement à introduire un nouveau référent, est parfaitement possible dans la construction à inversion de *ainsi* (**Au fond de la forêt vivait-il ; Ainsi vivait-il*).

Nous avons montré également la possibilité d'ambiguïté syntaxique entre les deux fonctions : Les Réunionnais parlaient ainsi le créole peut signifier : Les Réunionnais parlaient le créole de cette manière ou bien : Conformément à ce qui vient d'être dit, les Réunionnais parlaient le créole, alors que Les Réunionnais parlaient le créole ainsi n'a qu'une seule interprétation, la première. Ce même type d'ambiguïté se rencontre avec les adverbes acceptant les deux fonctions.

Tels sont les principaux traits qui font la singularité de *ainsi* au sein de la famille des adverbes, et le désignent comme l'un des plus représentatifs.

Références

Bonami, O. et Godard, D., 2001, Inversion du sujet, constituance et ordre des mots, *in* J.-M. Marandin (éd.), *Cahiers J.-C. Milner*, Paris : Verdier.

Borillo, A., 1990, A propos de la localisation spatiale, in *Langue française 86*, Paris : Larousse.

Cornish, F., 2005, A crosslinguistic study of ‘locative inversion’ : Evidence for the Functional Discourse Grammar model, in C. de Groot et K. Hengeveld (eds), *Morphosyntactic Expression in Functional Grammar*, Berlin et New York : Mouton de Gruyter.

Fournier, N., 1997, La place du sujet nominal dans les phrases à complément prépositionnel initial, in C. Fuchs (éd.), *La place du sujet en français contemporain*, Louvain-La-Neuve : Duculot.

Gross, M., 1975, *Méthodes en syntaxe. Régime des constructions complétives*, Paris : Hermann.

Gross, M., 1996, *Grammaire transformationnelle du français. 3 - Syntaxe de l’adverbe*, Paris : ASSTRIL.

Guimier, Cl., 1997, La place du sujet nominal dans les phrases à adverbe initial, in C. Fuchs (éd.), *La place du sujet en français contemporain* : Louvain-La-Neuve : Duculot

Jonare, B., 1976, *L’inversion dans la principale non interrogative en français contemporain*, Uppsala : Acta Universitatis Upsaliensis.

Kayne, R., 1973, L’inversion du sujet en français dans les propositions interrogatives, in *Le français moderne 4*, Paris : D’Artrey

Korzen, H., 1996, L’unité prédicative et la place du sujet dans les constructions inversées, in *Langue française 111*, Larousse : Paris.

Lambrecht, K., 1994, *Information Structure and Sentence Form : topic, focus and the mental representations of discourse referents*, Cambridge : Cambridge University Press.

Le Bidois, R., 1952, *L’inversion du sujet dans la prose contemporaine, étudiée spécialement dans l’œuvre de M. Proust*, Paris : D’Artrey.

Molinier, Ch. et Lévrier F., 2000, *Grammaire des adverbes, Description des formes en –ment*, Genève : Droz.

Muller, Cl., 2002, Inversion finale du sujet ou inversion postverbale, in *Cahiers de grammaire 27*, Toulouse : ERSS-Université Toulouse-Le Mirail.

Riegel, M., Pellat, J.-C., et Rioul, R., 2009, *Grammaire méthodique du français*, Paris : PUF.

Faire : opérateur causatif sur une phrase copulative bi-nominale

Takuya Nakamura

1. Introduction

Dans cet article, nous examinons une construction syntaxique à trois actants du verbe *faire*. Il s'agit de la structure *N faire de N N*. Nous argumentons dans ce qui suit que contrairement à l'apparence transitive, selon les cas, le complément direct n'est pas un objet direct, mais plutôt un prédicat, de telle sorte qu'on peut le considérer comme attribut du complément indirect.

Bien que son existence ait été mentionnée rapidement dans une grammaire comme *Le Bon Usage*[1], l'attribut d'un objet indirect est parfois clairement rejeté comme fonction grammaticale par des linguistes comme HERSLUND & SØRENSEN (1994)[2], et contrairement à l'attribut de l'objet direct, il a été très peu discuté dans la littérature. Il est cependant possible de considérer ce verbe *faire* comme un des opérateurs causatifs dans le sens de GROSS (1981) s'appliquant à une phrase copulative, de telle sorte qu'une case vide dans la distribution des causatifs soit remplie.

2. Ambiguïtés fonctionnelles

Voici un exemple type de notre étude :

(1) *Luc a fait de ce morceau de bois un instrument utile*

Une phrase comme celle-ci est ambiguë. Dans l'une des deux interprétations qui nous intéressent, la phrase (1) peut avoir pour paraphrase la phrase suivante :

(2) *Luc a créé à partir de ce morceau de bois un instrument utile*

Dans l'autre interprétation, la phrase peut se paraphraser par la phrase (3) :

(3) *Luc a fait que ce morceau de bois soit un instrument utile*

Dans un exemple comme (1), le complément direct du verbe semble assumer la fonction d'objet direct, mais ce n'est pas si évident. Bien qu'on puisse lui accorder cette fonction si la phrase (1) a une interprétation parallèle à celle de (2), il est difficile d'en dire autant, si elle a une interprétation parallèle à celle de (3). Dans ce dernier cas, plusieurs tests syntaxiques pour la reconnaissance de cette fonction ne se vérifient pas avec ce complément. Examinons un exemple non ambigu :

(4) *Bichat n'a pu aller jusqu'à l'élément fondamental de tous ces tissus, la cellule elle-même, mais ses conceptions et ses travaux font de lui le fondateur de l'anatomie générale des animaux.* (Histoire générale des sciences, R. Taton (dir.), Frantext)

Les tests de a) extraction par *ce que* et de b) interrogation par *que* appliqués à la phrase (4) donnent des résultats attendus d'un objet direct, mais les tests de c) pronominalisation et de d) passivation (et de formation pronominale en *se*) ne donnent pas les résultats attendus de cette fonction :

(5) a. *Il était facile à prévoir ce que ses conceptions et ses travaux auraient fait de lui*

b. *Qu'est-ce que ses conceptions et ses travaux ont fait de lui ?*

c. **Le fondateur de l'anatomie générale des animaux, ses conceptions et ses travaux l'ont fait de lui*

d-1. **Le fondateur de l'anatomie générale des animaux (a été, est) fait de lui (par ses conceptions et ses travaux)*

d-2. **Le fondateur de l'anatomie générale des animaux (s'est fait, se fait) de lui*

Selon mon hypothèse, les faits (d-1,2) sont cruciaux : si la phrase (4) n'accepte pas le changement de voix (dans les cas examinés ici, la mise en sujet de l'objet), c'est parce que dans la phrase en question *rien n'assume la fonction qui doit devenir le sujet, c'est-à-dire la fonction d'objet direct*. Tout au mieux, ces observations nous indiquent que dans la seconde interprétation, le complément direct n'est pas l'objet direct du verbe.

---

[1] LA FAUCI (1980) a mentionné la construction équivalente en italien, mais dans une optique différente de celle présentée ici. Dans un des rares travaux consacrés à l'emploi copulatif du verbe faire, LAUWERS (XXXX) parle de cette construction, mais elle est aussitôt écartée de la discussion.

[2] Ces auteurs ont proposé dans le cadre de la grammaire valencielle une fonction grammaticale appelée « Adjet » pour le troisième actant du verbe : en gros, ce sont des compléments post-verbaux faisant partie de la valence non objet direct (compl. datif, locatif, attribut, etc.). Un Adjet est par définition non répétable, d'où l'exclusion de la séquence $(Prép\ N)_{Adjet}\ (Dét\ N)_{Adjet=Attribut}$. Voir aussi HERSLUND (1994).

A propos de l'exemple (1), il faut aussi souligner que corollairement à la différence d'interprétation du complément direct, on peut également trouver une différence au niveau du sujet : dans le cas où le complément direct assume la fonction d'objet direct, le sujet doit être agentif, tandis que dans le cas contraire, le sujet peut ou non être agentif, comme le fait déjà supposer la paraphrase causative (3) :

(6) a. *(Luc, *son habileté) a fait de ce morceau de bois un instrument utile* (cf. *(Luc, *son habileté) a créé à partir de ce morceau de bois un instrument utile*)

b. *(Luc, son habileté) a fait de ce bois un instrument utile* (cf. *(Luc, son habileté) a fait que ce bois soit un instrument utile*)

3. Différents emplois

La différence fonctionnelle du complément direct et la différence sémantique du sujet observées autour du même verbe font naturellement penser à l'existence de deux emplois du verbe *faire*.

3.1. Emploi transitif

L'un des deux emplois est muni d'une grille d'arguments constituée d'Agent et de Patient, exemplifié par la paire (1) et (2). C'est un emploi des plus ordinaires de ce verbe transitif dans son sens de "création". L'objet direct réalise un argument sémantique, Patient, par rapport au prédicat *faire*. Il est un actant dans un événement décrit par le prédicat. Dans cette conception, le *SN* direct fonctionne comme une expression purement référentielle. Ajoutons l'observation que dans cet emploi, le complément prépositionnel n'est pas obligatoire : le verbe est bi-valent, sémantiquement et syntaxiquement.

3.2. Emploi non transitif

Dans le second emploi, comme la paraphrase (3) l'indique, la relation interprétative qui s'établit entre le complément indirect et le complément direct est comparable à celle de sujet-attribut du sujet. Le complément direct semble avoir la même fonction que celle de l'attribut dans la phrase suivante :

(7) *Ce morceau de bois est un instrument utile*

Le syntagme nominal *un instrument utile* dans cet exemple fonctionne comme attribut du sujet, du type prédicable "typant"[3]. La structure fonctionnelle d'une phrase copulative comme (7) à l'appui, il est possible d'analyser le *SN* direct comme attribut du complément indirect *de N*.

Le rôle de sujet que joue le complément indirect sur le plan interprétatif/sémantique explique son obligatoriété syntaxique, comme le montre l'exemple suivant :

(8) * *Ses conceptions et ses travaux ont fait le fondateur de l'anatomie générale des animaux*

Le rôle d'attribut que joue le *SN* direct sur le plan interprétatif/sémantique explique, par ailleurs, le fait qu'il n'est pas un objet direct.

Pour résumer, sur le plan descriptif, cet emploi du verbe *faire* construit une phrase avec a) un sujet de tous types (agentif ou non agentif), b) un complément indirect en *de* et c) un attribut du complément indirect. Le sens (et la forme) de la construction indique clairement qu'il s'agit d'une construction causative.

4. Quelques traits particuliers

Le SN direct qui nous intéresse présente-t-il les caractéristiques d'un attribut ?

4.1. Accord

Du point de vue morpho-syntaxique, il manifeste un accord en genre et/ou en nombre avec le nom du complément en *de* :

(9) a. *Il voulait faire de ses filles (des avocates, ?*des avocats, *un avocat, *une avocate)*

b. *Ses filles sont (des avocates, ?*des avocats, *un avocat, *une avocate)*

c. *« Plusieurs facteurs se conjuguent pour faire de 2012 une opportunité historique pour faire de la France une championne du numérique (...) »* (*lemonde.fr*)

d. ?**(...) pour faire de la France un champion du numérique (...)*

[3] RIEGEL (1985).

4.2. Article indéfini sous la négation

Un autre exemple qui différencie deux types de complément direct : sous la négation, un article indéfini déterminant un objet direct peut prendre la forme *de*, si la négation « a pour effet de rendre impossible toute véritable quantification »[4], tandis que celui qui détermine un attribut n'est pas sensible à la négation :

(10) a. *Luc n'a pas d'ami(s)*[5]

b. *Luc n'est pas un ami* (cf.**Luc n'est pas d'ami*)

Cette corrélation entre négation et article indéfini ne s'observe pas dans la construction qui nous intéresse :

(11) a. *Les 35 heures n'ont pas fait de la France « un immense parc de loisirs »... (lemonde.fr)*

b. **Les 35 heures n'ont pas fait de la France d'immense parc de loisirs...*

5. Types de l'attribut

Il existe diverses typologies formelles et surtout interprétatives en ce qui concerne l'attribut du sujet. Nous présentons celle de RIEGEL (1985) et VAN PETEGHEM (1991), et évaluons les relations entre *(de) N* et *N*.

5.1. RIEGEL (1985), VAN PETEGHEM (1991)

Quatre types d'attributs du sujet sont distingués :

(12) a. *N être ADJ* = : *Luc est gentil*

b. *N être UN N* = : *Ce meuble est un vaisselier*

c. *N être LE N* = : *Mon voisin est le champion du monde d'haltérophilie*

d. *LE N être N* = : *Le champion du monde d'haltérophilie est mon voisin*

e. *N être N* = : *Luc est professeur*

La phrase (12a) et l'une des deux interprétations de (12e) présentent un attribut « qualifiant » selon Riegel, et « prédicationnel » selon Van Peteghem. Dans les exemples (12b, c) et l'autre interprétation de (12e), l'attribut présente une interprétation « typante », pour Riegel. Selon Van Peteghem, une phrase du type (12b) est une phrase « identificationnelle » avec un sujet connu et un attribut qui représente une information nouvelle. La phrase du type (12d), toujours selon le même auteur, est une phrase « spécificationnelle », où l'ordre de surface des constituants est inversé par rapport à l'ordre canonique des fonctions : le sujet superficiel (*le champion du monde d'haltérophilie*) est un attribut profond, et l'attribut superficiel (*mon voisin*) est un sujet profond.

5.2. Les relations entre *(de) N* et *N*

5.2.1. « Prédicationnel » exclu

Dans la construction du verbe *faire* étudiée, il existe une restriction catégorielle : au complément direct ne se substitue pas un adjectif. Ce fait implique que dans la construction en question, le complément direct ne s'interprète pas comme attribut « qualifiant » (Riegel) ou « prédicationnel » (Van Peteghem). Corrélativement à ce fait, à la place du complément direct on ne peut pas avoir non plus un nom sans déterminant :

(13) *Le hasard a fait de Jean Dupont (*malin, *très professeur)*

5.2.2. Attribut « identificationnel » accepté

La construction accepte le déploiement de la prédication sujet-attribut identificationnel ou typant, avec un *SN* défini déictique comme complément indirect et un *SN* indéfini comme complément direct :

(14) a. *(Le hasard, son père) a fait de cet homme un professeur de latin*

b. *(La nouvelle classification des meubles, l'antiquaire) a fait de cet meuble un vaisselier*

5.3.3. Attribut « spécificationnel » ?

Si les deux *SN* sont définis, on peut avoir une phrase du type :

(15) *L'exercice quotidien a fait de mon voisin le champion du monde d'haltérophilie*

[4] RIEGEL *et al*. (2009 : 297).

[5] *Luc n'a pas un ami* est possible avec l'interprétation de *Luc n'a pas un ami mais il en a plusieurs* ou avec celle de *Luc n'a aucun ami*.

La relation copulative enchâssée dans la construction (15) est celle que RIEGEL (1985) appelle « typante », représentée en (12c), et c'est un cas de copulative spécificationnelle de VAN PETEGHEM (1991) mais avec l'ordre canonique des constituants. Selon cette dernière, ce type de phrase copulative a la particularité que l'ordre inversé des deux *SN* crée une phrase aussi acceptable, comme en (12d). Maintenant, si on enchâsse une copulative du type (12d) sous *faire*, le résultat est inacceptable :

(16) * *L'exercice quotidien a fait du champion du monde d'haltérophilie mon voisin*

Cette impossibilité donne une certaine confirmation à l'hypothèse poursuivie dans cet article et renforce l'analyse de VAN PETEGHEM (1991) : le complément en *de* doit être fonctionnellement équivalent du sujet d'une phrase copulative, et le complément direct de l'attribut ; le sujet et l'attribut de surface d'une phrase spécificationnelle sont, à un certain niveau de l'analyse, l'attribut et le sujet profonds.

6. *faire* opérateur causatif

A partir des observations précédentes, on peut faire l'hypothèse suivante : le verbe *faire* de cette construction est un opérateur causatif qui s'applique à une proposition de base copulative bi-nominale (17a) dont l'interprétation est identificationnelle (avec un attribut « typant »). En qualité d'opérateur causatif s'appliquant à une phrase copulative, ce verbe est en distribution complémentaire avec le verbe opérateur *rendre*, qui s'applique à une phrase copulative de base dont l'attribut est un adjectif (17b), et peut-être avec le verbe opérateur *mettre*, qui s'applique à une phrase copulative de base dont l'attribut est un syntagme prépositionnel fonctionnellement équivalent à un adjectif[6] (17c). Voici les phrases copulatives de base :

(17) a. *N être Dét N =* : *Luc est un professeur de latin*

b. *N être ADJ =* : *L'eau est rouge*

c. *N être Prép N =* : *Luc est en rage*

L'application d'opérateurs appropriés à chacune des phrases de (17) produit des phrases causatives comme suit :

(18) a. *N faire # N être Dét N =* : *(Son intelligence, le voeux de son père) a fait de Luc un professeur de latin*

b. *N rendre # N être ADJ =* : *(Le sang, Paul) rend l'eau rouge*

c. *N mettre # N être N =* : *(La désinvolture de Léa, Max) met Luc en rage*

Comme nous pouvons constater, les opérateurs causatifs énumérés ici ont l'effet d'augmenter d'un le nombre d'actants, par rapport à la proposition de base à laquelle ils s'appliquent, comme le fait le verbe opérateur causatif par exellence *faire* quand il se construit avec l'infinitif. Le sujet de ces opérateurs s'interprètent comme une « cause ».

Il reste une case vide dans la distribution des causatifs sur les phrases copulatives : c'est le cas d'une phrase copulative dont l'attribut est un nom sans déterminant. Pour une certaine sous-classe d'attributs sans déterminants, le verbe *faire* peut servir d'opérateur causatif :

(19) a. *N être N =* : *F. H. est grand maître de l'ordre de la Légion d'honneur*

b. *N être N =* : *Luc est étudiant*

c. *N faire # N être N =* : *On a fait F. H. grand maître de l'ordre de la Légion d'honneur*

d. *N faire # N être N =* : **On a fait Luc étudiant/*Luc a été fait étudiant*[7]

Il existe donc une certaine restriction de sélection lexicale entre l'opérateur *faire* et l'attribut sans déterminant : ce dernier doit désigner une fonction à laquelle accède le sujet de la phrase copulative à travers un processus socialement codé de nomination, d'où l'exclusion d'un attribut comme *étudiant*. Le verbe *faire* ici est synonyme de verbes à attribut de l'objet direct comme *nommer*, *élire*. On peut en conclure qu'il ne s'agit pas d'un opérateur causatif, mais d'un verbe plein. Il va sans dire que si les mêmes attributs (qu'ils soient ceux acceptés dans une construction à attribut de l'objet direct, ou pas) se réalisent avec un déterminant, la causativation se fait normalement avec la structure *N faire de N N* :

(20) a. *Le résultat des élections a fait de F. H. un grand maître de l'ordre de la Légion d'honneur*

b. *Le concours de circonstances a fait de Luc un étudiant*

[6] GROSS (1981) mentionne les opérateurs causatifs *rendre* et *mettre*, mais pas ce *faire*.
[7] Sauf dans un contexte particulier où « *étudiant* » serait un titre.

7. Conclusions

En premier lieu, nous avons constaté qu'une même structure de surface, *N faire de N N*, peut être fonctionnellement différente : une structure transitive ordinaire avec un objet direct et un complément circonstanciel, d'un côté, et une structure à apparence transitive mais avec un complément indirect et son attribut.

GROSS (1981) à l'appui, nous avons montré la possibilité d'analyser le verbe *faire* de la construction étudiée ici comme un opérateur causatif s'appliquant à une phrase copulative bi-nominale. Dans le tableau de distribution des causatifs des phrases copulatives, la case pour un opérateur causatif d'une phrase copulative bi-nominale restait vide, à côté des opérateurs *rendre* (s'appliquant à une phrase copulative dont l'attribut est adjectival) et *mettre* (s'appliquant à une phrase copulative dont l'attribut est un syntagme prépositionnel). Ce vide ne semble pas avoir été remarqué dans la littérature, mais dès sa découverte, le voilà rempli avec notre *faire*.

Références bibliographiques

GREVISSE Maurice ; GOOSSE André (2004), *Le Bon Usage* (13ème édition), Gembloux, De Boeck-Duculot.

GROSS Maurice (1981), Les bases empiriques de la notion de prédicat sémantique, *Langages 63*, p. 8-52.

HERSLUND Michael (1994), Valence et relations grammaticales, *Linguistica XXXIV : 1*, p. 109-117.

HERSLUND Michael ; SØRENSEN Finn (1994), A Valence Based Theory of Grammatical Relations, in : FALSTER Jakobsen *et al*. (eds.), *Function and Expression in Functional Grammar*, Berlin, Mouton de Gruyter.

LA FAUCI Nunzio (1980), Aspects du mouvement de WH, verbes supports, double analyse, complétives au subjonctif, *Lingvisticæ Investigationes IV : 2*, 293-341.

RIEGEL Martin (1985), *L'adjectif attribut*, Paris, P.U.F.

RIEGEL Martin et al. (2009[4]), *Grammaire méthodique du français*, Paris, P.U.F.

VAN PETEGHEM Marleen (1991), *Les phrases copulatives dans les langues romanes*, Wilhelmsfeld, Gottfried Egert.

Declension of Czech Noun Phrases

Zuzana Nevěřilová

Abstract: This paper shows a practical application resulting from linguistic studies and language engineering. The application takes as input a Czech noun phrase (NP) or prepositional phrase and its case and outputs the same NP in the output case. The algorithm for such declension is described. Its usability is evaluated on a set of 286 (relatively short) corpus examples. The results seem to be promising as 259 of the NPs were judged to be correctly declined. NP were extracted from the corpus using a syntactic parser and the correctness of the respective NPs in different case were evaluated manually.

1 Introduction

In many natural language processing (NLP) tasks we need analysis as well as synthesis of sentences in natural language. Sometimes, the synthesised sentences use the same noun phrases (NP) as in previous input but within a different syntactic context. For example, in recognising textual entailment (RTE) a program has to be able to reformulate a sentence to another one with (almost) the same meaning, e.g. transition between active and passive sentences. In these cases NPs very often change their syntactic role (e.g. from object in active sentences to subject in passive sentences).

In languages with rich nominal inflection such as Czech (and of course most Slavic languages) NPs often have to change its case. Case changing concerns not only head nouns in the NP but also other components due to mandatory agreement of noun and its adjective modifiers. The problem looks quite straightforward but in this paper we show that in many cases it is not and sometimes it cannot even be solved without additional (semantic) information.

There are not many papers concerning noun phrases declension from the NLP point of view. On the other hand it is clear that natural languages have many features in common such as construction and attachment of NPs. According to [2] each phrase has a hierarchical phrase structure which is usually represented by a tree.

2 Czech Noun Phrases and Sentence Generation

Syntactic analysis is one of the most important parts of NLP analysis. From a broader perspective, syntactic analysis consists of three main steps: sentence boundaries detection, sentence constituents detection and (dependency/ phrasal/ hybrid) parse tree construction.

This work is closely related to the second step – sentence constituents detection. Thanks to syntactic parsers for Czech language such as `synt` [3] or SET [4] we are able to determine sentence constituents such as:

- noun phrases (NP)
- prepositional phrases (PP)
- verb phrases (VP)
- adverbial phrases (AP)

For the purpose of sentence generation we have to determine sentence constituents as well as generation rules that determine syntactic roles of these sentence constituents. Sentences are seen as tuples

$$(VP, NP_1, \ldots, NP_m, PP_1, \ldots, PP_n, AP_1, \ldots, AP_p)$$

and generation rules are tuples of functions $(f_1, \ldots, f_r)$. Each of these functions f_i describes how does the *i*th sentence constituent change in the newly generated sentence.

For example, the sentence "Petr včera snědl bramborovou polévku."[1] is a tuple (sníst[2]/ 3PERSON, SG, PAST; Petr[3]/ NOM; bramborová polévka[4]/ ACC; včera[5]/ ADV, TIME). A corresponding generation rule for passive sentence construction is a tuple:

(f1 (sníst/ 3PERSON, SG, PAST)□ sníst/ PASSIVE, SG, PAST;

f2 (Petr/ NOM)□ ε;

f3 (bramborová polévka/ ACC)□ bramborová polévka/ NOM;
f4 (včera/ ADV, TIME)□ včera/ ADV, TIME).

[1] Yesterday, Peter ate the potato soup.
[2] to eat
[3] Peter
[4] potato soup
[5] yesterday

In case of passive sentences the mandatory agreement concerns gender and number of the subject ("bramborová polévka"[6] and the verb phrase "být snězen"[7].

The resulting sentence is "Byla snězena bramborová polévka včera."[8], generated by a tuple (sníst/ PASSIVE, SG, PAST, ε, bramborová polévka/ NOM, včera/ ADV, TIME).

In the sentence generation described above the order of sentence constituents does not matter. Since Czech is a nearly free word order these newly generated sentences may sound unusual but are (syntactically) correct sentences. We extended the generation algorithm to generate more correct sentences by permutation of the sentence constituents but this step goes beyond the scope of this paper.

Moreover, in this paper we treat prepositional phrases as preposition + noun phrase. We therefore work with both prepositional (PP) and noun phrases (NP) in nearly the same matter.

3 Morphological Analyser

For declension of Czech noun phrases we are using automatic morphological analyser/ generator `majka` [9]. Authors of the software [5] keep in mind linguistic approaches such as distributed morphology analysis [10] but finally the application uses a dictionary lookup – it lists "all combinations of recognized input and corresponding outputs" [9]. This dictionary lookup is much faster than a "real" analysis and therefore `majka` outputs are usable in analysis and in generation as well. The software works with *words*, *lemmata* (base forms) and *tags* (syntactic categories described by abbreviations) and it operates in two modes:

- analyser: for a given word all lemmata and possible tags are returned, e.g. for the word "vysokou", possible pairs (lemma, tag) are (vysoká/ NOUN, ACC, SG, FEM) and (vysoký/ ADJ, ACC, SG, FEM)[9]
- generator: for given lemma and tag it returns the word, e.g. for the lemma "pes[10]" and the tag NOUN, DAT, PL it returns "psům"

It is clear that for the generation purpose a smaller set of tags is needed. For example in case of "vysoká" the generator does not need any information about gender. In syntactic parsing a larger set of tags is needed because of agreement checking between nouns, adjective modifiers and verbs.

N.B. that `majka` uses a rich set of tags described in [8]. This system is better in automatic processing but more difficult for human comprehension. In this paper we use a more explicit set of tags.

4 Declension Algorithm

In general, in each NP the syntactic analyser finds the *head noun*. Afterwards, it marks the whole NP with the head noun tag, e.g. the accusative NP such as "bramborovou polévku" in "Petr snědl bramborovou polévku." has the following tags: ACC, SG, FEM.

The algorithm takes the original NP and its tag as input and outputs the same NP in a different case. In case of prepositional phrases the preposition is omited if the input case differs from the output case.

After determining the head noun *h* at *j*th position in the NP, the NP is processed subsequently and the following rules (based on [1, p. 175–209]) are applied to each word *wi* in the NP:

1. determine all possible lemmata and tags for *wi*
2. if *wi* is not recognized and $i<j$ output *wi* and do not stop declension for the following words
3. if there is no agreement with *h* stop declension for the following words
4. if $i>j$ and *wi* is not adjective stop declension for the following words
5. if *wi* is a coordination conjunction (such as and, or, neither, respective) and $i<j$ allow declension for the following words
6. if *wi* is a numeral proceed in a different mode (decline the following genitive if the numeral means more than 4 and the output case is neither nominative not accusative)
7. if *wi* has agreement in gender, case and number, change the case according to the output case

Changing the case is quite straightforward for NPs containing:

[6] "Potato soup" is SG, FEM.
[7] "to be eaten" has to be changed to "was eaten"/ SG, FEM
[8] The potato soup was eaten.
[9] The word "vysoká" has at least two meanings – high (adjective) and university (informal).
[10] dog

1. a single noun or pronoun (which is at the same time the head noun) changes its case: “Petr snědl polévku (accusative).” “Polévka (nominative) byla snězena.”[11]

2. a single noun with preceding adjective modifiers changes its case, its modifiers change their cases as well: “Petr snědl bramborovou polévku (accusative).” “Bramborová polévka (nominative) byla snězena.”[12]

3. a coordination of nouns: “Petr snědl polévku a rýži (accusative).” “Polévka a rýže (nominative) byly snězeny.”[13]. In this case the last member of the coordination is considered to be the head noun. All preceding nouns, pronouns and adjectives change their case.

Declension becomes more complicated when NPs are more structured. These cases concern:

1. modifiers following the head noun. In case of agreement with the head noun these adjectives are treated in the same manner as preceding adjectives, e.g. “Petr snědl polévku plnou lesních hub.”[14]. The declension stops after the adjective modifier “plnou” resulting in “polévka plná lesních hub”.

2. several subsequent nouns starting with capital letters. This case happens in case of proper names, e.g. “pan Pavel Novák”[15]. In this case the last noun is considered to be the head noun and all the preceding nouns have to change their case. Since in locative and dative there are variants of most (human) names a rythmical rule is used. The forms “panu Pavlovi”[16] and “panu Pavlu Novákovi”[17] are correct NPs.

3. several subsequent words starting with capital letters. This case is similar to the previous one but concerns naming nominatives such as “v hotelu Praha”[18]. The proper name is not the subject of declension, it always stays in (naming or citation) nominative. We can distinguish both cases using a database of persons’ proper names.

4. genitive groups. Declension stops after the head noun, e.g. “fakulta” in “fakulta sociálních studií (nominative)”[19]. If the whole NP is in genitive the following adjective is ambiguous because it could be treated as the case 1, e.g. “fakulty sociálních studií (genitive)”.

5. numerals. In Czech (as well as many other Slavic languages) the declension of NPs with numerals depends of the number expressed by the numeral. If it means one, two, three or four, declension proceeds like in case 2, e.g. “dva psi (nominative)”[20] and “dvou psů (genitive)”. If the numeral expresses a number bigger than four, the NP is treated as a genitive group (e.g. “pět psů (nominative)”[21]) but only in nominative and accusative. Declension of numerals becomes even more complicated with numbers ending with one to four. Here both forms are valid e.g. “dvacet jeden pes” and “dvacet jedna psů“[22].

6. ambiguous (noun or adjective) words occuring in the NP. There is a specific class of Czech nouns derived from adjectives. Originally these nouns were parts of a noun phrase progressively reduced to the adjective part, e.g. “vysoká škola”[23] is transformed to “vysoká”[24] (at present used informally). After this transformation the nouns have the same forms as the adjective. When `ajka` determines a word both as adjective and noun a significant ambiguity can occur if the whole NP is in genitive, e.g. “vysoké stromy”[25]. Since this ambiguity can be recognized by language users we assume that language users tend to eliminate it. Therefore if there are two subsequent nouns and in the NP and can also be an adjective, the algorithm prefers the adjective.

[11] Petr ate the soup. The soup was eaten.
[12] Petr ate the potato soup. The potato soup was eaten.
[13] Petr ate the soup and the rice. The soup and the rice were eaten.
[14] Petr ate the soup full of wild mushrooms.
[15] Mr. Pavel Novak
[16] Mr. Pavel (dative)
[17] Mr. Pavel Novak (dative)
[18] in the Praha hotel
[19] faculty of social studies
[20] two dogs
[21] five dogs
[22] twenty one dogs
[23] university but literally vysoká škola means high school
[24] high
[25] high trees but also trees of the university

7. coordinations of genitive groups. In cases of NPs in genitive such as “Byli jsme tam i bez předsedy oblastní rady a prezidenta asociace malých výrobců.”[26] it is not feasible to decide correctly the head noun without additional (semantic) information and therefore the algorithm has lower precision.

5 Evaluation

We have evaluated the precision of the algorithm on 286 NPs from the corpus. We did not measure recall since the NPs detection depends on the syntactic analyser. Table 1 shows the number of correct declensions depending on the output case.

input NP length	# of input NPs	# of correct output NPs
6	1	0
5	1	1
4	5	4
3	25	14
2	37	31
1	212	209
total	286	259

Table 1: Number of correct declension depending on the length of the input NP (expressed by number of words).

The preliminary evaluation gives promising results. From 286 NPs 259 were judged correct.

6 Conclusion

This paper presents a practical application that takes a noun phrase (or a prepositional phrase) and its case (called input case) and outputs the noun phrase in the new case (called output case). The application is used when sentence generation in Czech language is used, namely in the game X-plain [6] and the inferencing system using verb valency frames [7].

A larger discussion is needed on whether the algorithm is transferable to other languages. In case of most Slavic languages NP/PP generation is a complex issue. In future we should test the algorithm on near languages such as Slovak or Polish. The success rate depends heavily on appropriate morphological analysers/generators.

We have evaluated the precision of the results on a small corpus and proved that the application is usable. It can be run from http://nlp.fi.muni.cz/projects/declension. In future we plan most detailed evaluation on different corpora.

Acknowledgments

This work has been partly supported by the Czech Science Foundation under the project P401/10/0792.

References

[1] Grepl, M.; Karlík, P.: *Skladba spisovné češtiny*. Edice Učebnice pro vysoké školy, Státní naklad., 1986.

[2] Hawkins, J. A.: The Typology of Noun Phrase Structure from a Processing Perspective. Říjen 2008. http://linguistics.ucdavis.edu/People/jhawkins/recent-papers/NPStructureLingTypologyDouble.pdf

[3] Horák, A.; Kadlec, V.: New Meta-grammar Constructs in Czech Language Parser synt. In *Proceedings of the 8th International Conference on Text, Speech and Dialogue (TSD 2005)*, Springer-Verlag, 2005, s. 85–92.

[4] Kovář, V.; Horák, A.; Jakubíček, M.: Syntactic Analysis Using Finite Patterns: A New Parsing System for Czech. In Human Language Technology. Challenges for Computer Science and Linguistics: 4th Language and Technology Conference, LTC 2009, Roznań, Poland, November 6-8, 2009, Revised Selected Papers, Springer, 2011, str. 161.

[5] Šmerk, P.: *K počítačové morfologické analýze češtiny [online]*. Disertační práce, Masarykova univerzita, Fakulta informatiky, 2010 [cit. 2012-05-21]. http://is.muni.cz/th/3880/fi\s\do5(d)/

[6] Nevěřilová, Z.: X-plain – A Game That Collects Common Sense Propositions. In *Proceedings of NLPCS*, Funchal, Portugal: SciTePress, 2010, ISBN 978-989-8425-13-3, str. 47–52.

[7] Nevěřilová, Z.: Common Sense Inference using Verb Valency Frames. In *Proceedings of the 15th International Conference on Text, Speech and Dialog TSD 2012*, Brno: Springer-Verlag, 2012, submitted.

[26] This sentence is ambiguous and most humans will translate is as ”We were there without the local council chair and small manufacturers association president.“ The other interpretation can be such as ”We were there without the chair of the local council and small manufacturers association president.“

[8] Sedláček, R.: ajka tagset. online; accessed 2012-05-21 from nlp.fi.muni.cz/projekty/ajka/tags.pdf, 2006.
[9] Šmerk, P.: Fast Morphological Analysis of Czech. In *Proceedings of the Raslan Workshop 2009*, Masarykova univerzita, 2009.
[10] Ziková, M.; Caha, P.: Princip synkretismu aneb Augiášův chlév české deklinace. *Linguistica ONLINE*, ročník 1, 2006, ISSN 1801-5336. http://www.phil.muni.cz/linguistica/art/zikcah/zic-001.pdf

La préposition *de* dans la construction N_0 *V de* N_1 du verbe *changer*

Kozue Ogata

1. Introduction

Le verbe *changer* se distingue par la diversité des constructions dans lesquelles il entre et par l'hétérogénéité des sens qui y sont dénotés. On remarque par ailleurs que sa fréquence (96ème dans tous les verbes (Baudot 1992)) est relativement élevée parmi les verbes dénotant un changement d'état. *Changer* est de loin plus fréquent que les autres verbes exprimant un changement d'état : *remplacer, déplacer, modifier, transformer, échanger*, etc. Nous nous intéressons à l'ensemble des constructions de ce verbe et aux rapports qu'elles entretiennent entre elles. Dans un premier temps, nous nous limitons aux constructions N_0 *V* N_1 et N_0 *V de* N_1.

En nous appuyant sur l'analyse d'exemples dans les corpus du *Monde* (1994) (noté *LM*) et de *Frantext*, nous essaierons de déterminer les propriétés syntaxiques de ces constructions. Il s'agit notamment de savoir :

1. Quelle est l'opposition entre les constructions N_0 *V* N_1 et N_0 *V de* N_1 ?
2. Ces constructions-ci sont-elles chacune homogènes ou faut-il distinguer plusieurs sous-groupes ?
3. Le complément *de* N_1 de N_0 *V de* N_1 est-il le même que *de dét* N_1 de N_0 (*parler + traiter + servir*) *de dét* N_1 ?

2. Les constructions de *changer* et leurs fréquences

Nous avons relevé, à l'aide d'*Unitex*, dans une année *du Monde*, 4509 occurrences du verbe *changer* ainsi que des noms *change* et *changement*. Nous nous sommes limitée aux 1000 premières occurrences verbales. Elles se répartissent ainsi, du point de vue des formes syntaxiques, dans les tables du *Lexique-Grammaire* indiquées entre parenthèses. Afin de situer les constructions N_0 *V* N_1 et N_0 *V de* N_1 (en caractères gras) dans l'ensemble des constructions de *changer*, nous donnons ci-dessous l'effectif de toutes les constructions, mises à part les constructions figées (*N+h* : substantif humain, *N-h :* substantif non humain) :

N_0 *V* (31 R) *La situation a changé* [N_0*+h V* ..... 70 ; N_0*-h V* ..... 304]
N_0*-h V* en N_1 1

N_0 V N_1 (32R1) *J'ai un peu changé mon programme*
[N_0*+h V* N_1..... 165 ; N_0*-h V* N_1 .....86]
N_0*-h V* N_1 à N_2*-h(y)*..2 ; N_1 est *V*é..... 16 ; N_0 *V* (N_1) 8
N_0 *V* N_1 (*pour + contre*) N_2 (38R) *J'ai changé ma voiture contre une jeep*
[N_0*+h V* N_1 (*pour + contre*) N_2 ...1]
N_0 *V* N_1 à N_2 [N_0*+h V* N_1 à N_2*-h*..... 8]
N_0*+h V* N_1 de N_2 1
N_0 *V* N_1 en N_2 (38R) *Les dernières pluies ont changé les rivières en torrents.*
[N_0*+h V* N_1 en N_2 ..... 5 ; N_0*-h V* N_1 en N_2 ..... 3]
***N_0 V* de N_1** (35 R) *Il a changé cinq fois de nationalité.*
[N_0*+h V* de N_1 ....230 ; N_0*-h V* de N_1 .....48]
N_0*+h V* de N_1 avec N_2..0 (35 RR) *Veux-tu changer de place avec moi ?*
N_0*+h V de* N_1 *à* N_2.....1

N_0 se *V* (31H) *Je n'ai pas le temps de me changer* [N_0*+h* se *V* .....3]
N_0*+h* se *V* en N_2.....2 ; N_0*-h* se *V* en N_2 .....8 ; N_0 se *V* de N_1 1

3. Les constructions N_0 *V* N_1 et N_0 *V de* N_1

3. 1. N_0 *V* N_1

Dans cette construction, le N_1 est le siège du changement. Le N_1 n'est pas contraint par rapport au N_0, contrairement à ce qui se passe pour le N_1 dans la construction N_0 *V de* N_1 (3.2. ci-dessous). Le N_0 est soit l'agent, soit la cause de ce changement. Le changement peut être du type « modification » ou « remplacement ».

3.1.1. « modification »

Quand *changer* indique une modification du N_1, le N_0 n'est pas limité au *N+humain* et peut être *-humain*.

(1) L'alternance n'a rien changé. (*LM*)

3.1.2. « remplacement »

Quand le changement implique un remplacement, c'est-à-dire qu'un exemplaire du N_1 est remplacé par un autre exemplaire, le N_0 est limité au *N+h*. Ainsi restreint, le sens « remplacement » est moins fréquent que le sens « modification » et est marqué.

> (2) Les ravitaillements ont permis de démontrer la virtuosité des mécaniciens qui ont mis 7 secondes 4/10 chez Benetton (…) pour changer les quatre roues (…). (*LM*)

Contrairement au verbe *remplacer*, le verbe *changer* n'entre pas dans la construction avec *par* : N_0 *V* N_1 *par* N_2 (par exemple, *Léa remplace son mobilier ancien par du neuf.*) ; *N_0 *changer* N_1 *par* N_2. Le « remplacement de N_1 par N_2 » est dénoté avec le verbe *changer* par la construction : N_0 *V* N_1 (*contre* + *pour*) N_2 : *J'ai changé ma voiture contre une jeep.*

Remarquons que dans la construction transitive N_0 *V* N_1, quand le N_0 est *+humain*, il peut y avoir une ambiguïté entre l'interprétation de modification et celle de remplacement. Par exemple, *J'ai changé la nappe* peut vouloir dire la modification ou le remplacement. Bien que l'interprétation de remplacement soit limitée au N_0*+h*, et qu'on puisse parfois relever des indices dans le contexte comme apportés par des compléments circonstanciels, la construction en soi admet les deux interprétations : modification et remplacement. Ce qui est commun dans la construction transitive N_0 *V* N_1, c'est l'existence du changement apporté par le N_0 au N_1.

3. 2. N_0 V de N_1

Dans cette construction, le N_1 n'admet pas de déterminant : N_0 *V de* (*E* +**dét*) N_1. Le N_1 indique une catégorie et ne réfère pas à une entité spécifique. Le changement est du type « remplacement », mais le N_1 ne réfère ni à une entité ancienne ni à une entité nouvelle. On ne peut pas exprimer, avec cette construction, le procès de « remplacement du N_1 par un autre exemplaire spécifique N_2 » :*N_0 *V de* N_1 (*pour* + *contre*) N_2, contrairement à ce que nous avons vu pour la construction transitive (3.1.2.). Après le changement, un exemplaire de la catégorie N_1 est remplacé par un autre exemplaire de la même catégorie. À travers ce remplacement, c'est le N_0 qui est affecté.

3.2.1. N_0

Dans la construction N_0 *V de* N_1, le N_0 n'est pas limité au *N+h*, contrairement à ce que nous avons remarqué pour la construction transitive directe N_0 *V* N_1 dénotant le « remplacement » pour laquelle le N_0 est limité au *N+h*. Cela signifie que le procès de remplacement du N_1 dans la construction N_0 *V de* N_1 n'exige pas la volonté du N_0. Le N_0 dans (3), l'OPAC, société, assimilable au *N+h*, peut ou non prendre l'initiative du remplacement du directeur :

> (3) L'OPAC qui, (…), a changé de directeur, s'est engagé à corriger ces regrettables pratiques. (*LM*)

Dans (4) avec N_0-*h*, le changement du N_1 a lieu sans la volonté du N_0 :

> (4) Passé dans le rayon des musiques sérieuses, (…), le jazz a changé d'image publique. (*LM*)

3.2.2. N_1

Le N_1, indiquant des catégories apparemment diverses, a cependant un point commun par rapport au N_0 : il faut que le N_1 ait un certain rapport avec le N_0[1]. Mel'cuk caractérise ce que dénote la construction N_0 *changer de* N_1, en fonction du type de N_1 (1992 :181, cité aussi dans Ulland (2007)) :

1. commencer à se trouver dans un exemplaire différent de *N*
2. commencer à avoir une valeur différente du *N* (une caractéristique, un attribut ou une propriété)
3. commencer à posséder ou utiliser un exemplaire différent du *N*.

Pottier (1987) définit, pour sa part, le N_1 par rapport au N_0 ainsi : localisation, caractéristique inhérente, objet externe. Nous constatons que les N_1 dans notre corpus peuvent en effet être caractérisés ainsi, dans l'ordre de fréquence : (a) propriété du N_0, (b) lieu où se trouve le N_0, (c) objet que le N_0 utilise : **(a) propriété du N_0** : Pierre a changé de (avis + coiffure + secrétaire + attitude), par exemple :

> (5) Paris avait changé *d'humeur.* (*LM*)

(b) lieu où se trouve le N_0 : Pierre a changé de (train + place), par exemple :

> (6) Ils ont été priés (…) de changer *de place, de pièce, d'étage*, voire *d'immeuble*. (*LM*)

(c) objet que le N_0 utilise : Pierre a changé de (gants + voiture), par exemple :

> (7) Après cinq jours de défilé, on change *de bobine*. (*LM*)

[1] Picoche (1986) remarque que tous les *N* qui entrent dans cette construction « sont tenus pour faire, d'une certaine manière, partie du » N_0. « Pour que *X change de Z*, il faut que préalablement, *il ait Z* ». p. 52-53.

Remarquons cependant que, selon le contexte, l'interprétation du même nom peut diverger quant à la caractérisation. Par exemple :

(8) *Pierre a changé de train*

peut signifier, qu'il est descendu d'un train pour en prendre un autre comme (b). Mais il peut également dénoter qu'il a changé de billet pour prendre un train différent de ce qu'il avait prévu, c'est-à-dire (c).

« de N_1 »

Le complément *de* N_1 sert, dans cette construction, donc à apporter la précision au changement en N_0. De ce point de vue, le complément *de* N_1 ici est comparable au complément *de* N_1, complément non obligatoire, employé, par exemple, dans la construction de *ressembler*, comme suit :

(9) Léa ressemble à sa mère *de visage.*

Le complément *de* N_1 : *de visage* précise de quelle partie Léa ressemble à sa mère.

Le changement dénoté dans cette construction concerne toujours le N_0, étant donné que le N_1 possède un certain lien avec le N_0. Il y a une continuité de rapport entre le N_0 et le N_1 : le N_0 a un exemplaire du N_1 avant le changement et un autre exemplaire du N_1 après le changement. Contrairement à ce qui se passe dans N_0 *V* N_1, où le changement porte sur le N_1, dans N_0 *V de* N_1, le changement affecte le N_0, le N_1 étant lié au N_0. En ce sens, on peut paraphraser certaines de ces constructions par une construction intransitive « N_0 *change quant au* N_1 » ou « N_0 *change en ce qui concerne* (*son* + *le*) N_1 » : Elle *a changé de coiffure* peut être paraphrasé : *Elle a changé quant à la coiffure* ou *Léa a changé. Cela concerne* (*sa* + *la*) *coiffure*. C'est Léa qui a changé en partie.

La construction pronominale correspondant à N_0 *V de* N_1 donne une phrase peu naturelle ou inacceptable, car elle est redondante en ce qui concerne la valeur réflexive :

(10) ?? Léa s'est changé de (coiffure + train + attitude + mari).

« Adverbes de degré »

Avec certains N_1, la possibilité de faire intervenir les degrés s'observe, bien qu'il s'agisse de remplacement : Max a *un peu* changé de (coiffure + place + attitude). Il s'agit bien du remplacement, par exemple, d'une coiffure par une autre, mais s'agissant de propriété de N_0, on peut considérer le changement comme continuum dont on peut mesurer le degré dans l'écart : *un peu*, *beaucoup*, etc.

4. N_0 *V* N_1 *vs* N_0 *V de* N_1

Dans l'opposition entre les constructions N_0 *V* N_1 et N_0 *V de* N_1, on peut remarquer ceci :

— Dans la construction N_0 *V* N_1, le N_1 n'est pas restreint. C'est à ce N_1 que s'applique le procès du changement, soit la « modification » soit le « remplacement ». Au sens de « remplacement », le N_0 est limité au *N+h*. Dans cette construction, le N_1 a une référence spécifique. Le remplacement du N_1 par une autre entité (=N_2) peut être dénoté par la construction N_0 *V* N_1 *(contre+ pour)* N_2.

— Dans la construction N_0 *V de* N_1, le N_1 n'a pas de référence spécifique. On ne peut pas exprimer, avec cette construction, le procès de « remplacement du N_1 par un autre exemplaire spécifique N_2 ». Le N_0 n'est pas limité au *N+h* et ce changement ne requiert pas la volonté. Le changement dénoté dans cette construction concerne toujours le N_0, étant donné que le N_1 possède un certain lien avec le N_0. Le N_0 a un N_1 avant le changement et un autre N_1 après le changement. Il y a une continuité de rapport entre le N_0 et le N_1.

L'occurrence suivante est intéressante en ce sens qu'elle met en évidence la différence d'implication des constructions N_0 *V* N_1 et N_0 *V de* N_1 :

> (11) Il faut aujourd'hui non pas *changer le socialisme*, qui reste un socle majeur de représentation d'un monde qui ne se confond pas avec la libre loi de l'argent, mais *changer de socialisme*. (*LM* 2009) (C'est nous qui soulignons)

L'auteur affirme que « changer le socialisme », c'est-à-dire « modifier le socialisme actuel », n'est pas suffisant et qu'il faut « changer de socialisme », c'est-à-dire « remplacer le socialisme actuel par un nouveau socialisme ».

5. N_0 *V de* N_1 d'autres verbes

Quant à la construction du verbe *changer* : N_0 *V de* N_1, comment se caractérise-t-elle par rapport aux autres constructions en *de* N_1 ? Dans les constructions des verbes qui entrent à la fois dans la construction transitive directe en N_0 *V* N_1 et dans la construction transitive indirecte en N_0 *V de* N_1, le N_1 dans N_0 *V de* N_1 peut être accompagné d'un déterminant (N_0 *V de* (*E* + *dét*) N_1), excepté pour les verbes *changer* et *redoubler*, selon nos enquêtes. Dans la construction N_0 *V de dét* N_1 dans laquelle le N_1 apparaît accompagné d'un déterminant, par exemple avec les verbes *parler, traiter, servir*, etc., le N_1 peut référer à une entité spécifique et le procès que dénote le verbe s'applique au N_1. Il répond à la question : « De quoi V N_0 ? » :

(12) De quoi Pierre a-t-il parlé ? — Pierre a parlé de son nouveau livre.

Avec *N_0 changer de N_1*, nous avons vu que le *N_1* ne réfère pas à une entité spécifique. Il dénote une catégorie qui a un lien avec le *N_0* et le procès de changement s'applique ainsi à cette partie du *N_0*. La construction *N_0 changer de N_1* ne répond pas à la question : « De quoi V-N_0 ? » [2]:

(13) Pierre a changé de (avis + coiffure + secrétaire + attitude)
*? De quoi Pierre a-t-il changé ? (plutôt inacceptable)

Avec le verbe *redoubler*, qui n'admet pas non plus de déterminant au *N_1 : N_0 V de* (*E* + ***dét***) *N_1*, la question « De quoi V-N_0 ? »n'est pas possible :

(14) La tempête a redoublé de (intensité + violence).
*De quoi la tempête a-t-elle redoublé ?

Le complément *de N_1* dans la construction *N_0 changer de N_1* (et de *redoubler* également) est à distinguer ainsi du complément indirect *de dét N_1* dans la construction *N_0 V de dét N_1* que l'on trouve avec d'autres verbes.

6. Conclusion

A propos de l'opposition entre les constructions *N_0 V N_1* et *N_0 V de N_1*, nous avons remarqué que, dans *N_0 V N_1*, c'est le *N_1* qui change, qu'il soit modifié ou remplacé, le *N_0* n'étant pas affecté. Par contre, dans *N_0 V de N_1*, le changement affecte le *N_0*, le *N_1* étant lié au *N_0*. Le *N_0* a un exemplaire de la catégorie *N_1* avant le changement et un autre exemplaire après le changement. Il y a une continuité de rapport entre le *N_0* et le *N_1*. Le complément *de N_1* est à distinguer ainsi du complément indirect d'autres verbes, qui admet le déterminant : *N_0 V de dét N_1*.

Il sera nécessaire désormais d'approfondir nos recherches sur les autres constructions du verbe *changer* et les constructions *N_0 V de N_1* d'autres verbes comparables (*redoubler*, et encore, *augmenter, se tromper*, par exemple).

Références

Baudot, J., (1992), Fréquences d'utilisation des mots en français écrit contemporain, Montréal : Les Presses de l'Université de Montréal.

Boons, J.-P., Guillet, A., Leclère, Ch. (1976), La structure des phrases simples en français : classe des constructions transitives, Rapport de Recherches 6. LADL. Paris.

Gross, M. (1975), Méthodes en syntaxe, Paris : Hermann.

Mel'cuk, I. (1992), « Changer et changement en français contemporain (étude sémantico- lexicographique) », Bulletin de la Société Linguistique de Paris, 87/1 : 161-223.

Picoche, J. (1986), Structures sémantiques du lexique français, Paris : Nathan.

Pottier, B. (1987), Théorie et analyse en linguistique, Paris : Hachette.

Ulland, H. (2007), « Un verbe changeant : Etude sur les diverses structures argumentales de changer », 26th conference on Lexis and Grammar, Bonifacio, 2-6 October 2007.

[2] L'inacceptabilité de la question semble légèrement varier, selon nos informateurs, suivant le type du *N_1* : *Pierre a changé de train.— *De quoi Pierre a-t-il changé ?* (inacceptable)
Pierre a changé de voiture. — ? *De quoi Pierre a-t-il changé ?* (douteux)

From Treebanks to Lexical Entries. Clustering the *Index Thomisticus*

Marco Passarotti

Abstract: Language resources (LRs) such as corpora, lexica, grammars and ontologies are strictly related to each other at both development and exploitation stage. In particular, a strong relation holds between lexical resources and annotated corpora.
Recent years have seen a large growth of projects aimed at building LRs for Classical languages. Among these new LRs are syntactically annotated corpora (*treebanks*), which can be exploited to provide empirical evidence to test and refine lexical resources developed over the centuries by Ancient Greek and Latin lexicography.
This paper describes the application of clustering techniques to the *Index Thomisticus* Treebank corpus to organise the meanings of lemma *forma* in Thomas Aquinas' works, according to its textual and syntactic behaviour.
Clustering is an unsupervised learning method dealing with finding a structure in a collection of data. Applying clustering techniques to textual data grounds on the theoretical assumption that words that are used in similar contexts tend to have the same or related meanings (Distributional Hypothesis by HARRIS (1954)).
Our results show that syntactic metadata are indeed helpful for clustering purposes.

1. Introduction

Language resources (LRs) such as corpora, lexica, grammars and ontologies are strictly related to each other at both development and exploitation stage. In particular, a strong relation holds between lexical resources and annotated corpora. This relation is twofold: linguistic annotation of textual data is indeed supported and improved by the use of lexica and dictionaries, while these latter can be induced from annotated data in corpus-driven fashion.

Recent years have seen a large growth of projects aimed at building LRs for Classical languages. Among these new LRs are syntactically annotated corpora (*treebanks*), which can be exploited to provide empirical evidence to test and refine lexical resources developed over the centuries by Ancient Greek and Latin lexicography.

This paper describes the application of clustering techniques to a Medieval Latin treebank to organise the meanings of a specific lemma, according to its textual and syntactic behaviour.

2. The *Index Thomisticus* Treebank

Started in 1949 by father Roberto Busa, the *Index Thomisticus* (IT; BUSA, 1974-1980) is the first digital corpus of Latin. The IT contains the opera omnia of Thomas Aquinas (118 texts) as well as 61 texts by other authors related to Thomas, for a total of around 11 million tokens. The corpus is morphologically tagged and lemmatised.

The *Index Thomisticus* Treebank project (IT-TB; http://itreebank.marginalia.it) aims at performing the syntactic annotation of the whole IT. Presently, the IT-TB corpus consists of around 160,000 annotated tokens for a total of approximately 9,000 sentences excerpted from *Scriptum super Sententiis Magistri Petri Lombardi*, *Summa contra Gentiles* and *Summa Theologiae*.

Treebanks are textual corpora annotated at syntactic level. Data can be annotated according to two main grammar frameworks, both representing syntactic structures with tree-graphs. These are phrase structure grammar (PSG), mostly used in generative-based approaches, and dependency grammar (DG), started by TESNIÈRE (1959). While in DG trees all nodes are labeled with words or empty strings, in PSG trees all, and only, the leaf nodes are labeled with words or empty strings, and internal nodes are labeled with non-terminal symbols.

Latin is a richly inflected language, featuring discontinuous constituents and a moderately free word-order. As PSG has proved to be a framework more suitable for representing poorly inflected languages, these features of Latin have influenced the choice of DG as the most appropriate grammar framework for building Latin treebanks. Moreover, DG has recently gained wide interest because it is simple, yet providing useful information for many NLP tasks and for representing predicate-argument structure.

The first two Latin treebanks have started in 2006: they are the Latin Dependency Treebank (LDT) by Tufts University in Boston, based on texts of the Classical era, and the *Index Thomisticus* Treebank (IT-TB) by Catholic University in Milan, based on the IT.

IT-TB and LDT share the same annotation guidelines, which resemble those of the Prague Dependency Treebank of Czech (PDT). They are theoretically grounded on Functional Generative Description (SGALL ET AL., 1986), a dependency-based grammar framework developed in Prague and intensively applied and tested while building the PDT.

3. The 'Lessico Tomistico Biculturale'

The IT-TB is part of a bigger project named 'Lessico Tomistico Biculturale' (LTB). LTB aims at building a new lexicon of Thomas Aquinas by empirical confrontation with the evidence provided by the IT. Indeed, the entries of the available lexica of Thomas are systematically biased by the criteria for the selection of the examples adopted to describe the different meanings of lemmas. This limitation can now be overcome by exploiting the IT and, particularly, the IT-TB.

The first lemma we want to analyse for LTB is *forma*. This lemma has 18,357 occurrences in the IT, 16,525 of which in Thomas' works and 1,832 in the texts of other authors. We devoted the first years of the project to annotate those sentences that feature at least one occurrence of *forma*. So far, 5,191 occurrences of *forma* have been annotated in the IT-TB.

Forma is a technical word in Thomas' writings, showing high polysemy. In the lexicon of Thomas Aquinas by DEFERRARI & BARRY (1948-1949: 433-438), *forma* has 5 meanings:

a) 'form, shape', synonym of *figura*;
b) 'form', the configuration of an artificial thing as distinct from 'figure' (which is the configuration of natural things);
c) 'form', the actualizing principle that makes a thing be what it is;
d) 'mode, manner';
e) 'formula'.

In Latin Wordnet (http://multiwordnet.fbk.eu), *forma* has 21 senses, which do not include all those present in Thomas.

4. Clustering[1]

Clustering is an unsupervised learning method dealing with finding a structure in a collection of (un)labeled data. Objects are organised into groups (*clusters*) whose members are similar in some way and are dissimilar to the members of other clusters.

Word clustering techniques usually follow a two-step procedure:

1. classification: each word occurrence is represented as an observation in a matrix and the (dis)similarity of two observations is computed;
2. clustering: some clustering algorithm is applied, such that similar occurrences are grouped together.

Hierarchical clustering is a specific method of cluster analysis which seeks to build a hierarchy of clusters. It makes use of two main strategies:

- agglomerative (bottom-up): each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy;
- divisive (top-down): all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

Applying clustering techniques to textual data in order to organise the different meanings of words grounds on the theoretical assumption that words that are used in similar contexts tend to have the same or related meanings (Distributional Hypothesis (DH) by HARRIS (1954)). Similar to DH is the notion of context of situation by Firth, who points out the context-dependent nature of meaning, as reported in his famous quotation: "You shall know a word by the company it keeps" (FIRTH, 1957: 11).

Considering the contextual behaviour of a word to organise its meaning is closely related to word sense disambiguation. Word sense disambiguation must be kept separate from word sense discrimination. The former assigns a word a range of meanings excerpted from a preexisting sense inventory, usually provided by a dictionary or some other handcrafted resource. The latter is a method that does not use any sense inventory: clusters do not have a definition or sense label associated, but they are (semi-)automatically labeled with a gloss that describes the underlying meaning of the target word in those contexts that share distributional characteristics.

The contexts used for clustering can be of two different types:

- headed contexts have a target word (*head*) that serves as the focus of the context;
- headless contexts do not contain any specific target word and are used to make determinations about the overall text analysed and not a specific word.

The similarity itself between contexts can be considered in two different fashions. First-order similarity is the most intuitive approach: two contexts that share a large percentage of words are likely to be similar and scores are based on the number of matching words. Second-order similarity replaces the contexts with something else that still represents it and yet provides a richer basis for measuring similarity. For instance, a word can be replaced with its definition provided by a lexical resource like Wordnet.

5. Clustering *forma* in the *Index Thomisticus*

We apply hierarchical clustering techniques to organise the occurrences of *forma* in both IT and IT-TB into clusters, so that occurrences showing similar behaviour fall in the same cluster (or in clusters that are close to each other). Our method features the following properties:

- we perform word sense discrimination instead of disambiguation: our method is language-independent and does not use any lexical resource at clustering stage. Hence, we consider first-order similarity and do not associate any sense label to clusters;
- we use headed contexts: our input data are the concordances of a specific target word (i.e. *forma*);

[1] The contents of this section are mostly taken from PEDERSEN (2006).

- we carry out a DIvisive hierarchical clustering ANAlysis (KAUFMAN & ROUSSEEUW, 1990) by using the function DIANA (MAECHLER ET AL., 2011), available in the library *cluster* of the R free statistical software (http://www.r-project.org).

Our input data are organised into two matrices, one built from the IT-TB and one from the IT:

- matrix A (5,191 observations): for each occurrence of *forma* in the IT-TB, one observation reports the lemmas of[2]:
  a) its parent and grandparent in the tree;
  b) its attributives (dependent nodes with syntactic label 'Atr');
  c) its coordinated nodes in the tree;
  d) up to 2 words preceding and 2 words following the occurrence of *forma* concerned in the observation.

  While (a), (b) and (c) report information taken from the IT-TB (i.e. syntactic information), (d) features information concerning the linear word order in the text (taken from the IT);
- matrix B (18,357 observations): for each occurrence of *forma* in the IT, one observation reports the lemmas of up to 3 words preceding and 3 words following the occurrence of *forma* concerned.

For each of these matrices, we produce a dissimilarity matrix generated by considering a modification of the *simple matching distance* (SOKAL & MICHENER, 1958). The *simple matching distance* between two observations, *r* and *s*, with categories (x_{r1}, x_{r2}, …, x_{rk}) and (x_{s1}, x_{s2}, …, x_{sk}) over *k* variables is:

$$dist(r,s) = \frac{k - \sum_{j=1}^{k} sim(x_{rj}, x_{sj})}{k}, \qquad sim(x_{rj}, x_{sj}) = \begin{cases} 0 \ if \ x_{rj} \neq x_{sj} \\ 1 \ if \ x_{rj} = x_{sj} \end{cases}.$$

For every pair of observations in each matrix, we calculate their (dis)similarity by grouping together more variables (i.e. the values of the columns in observations), hence not considering each variable on its own, as specified by *simple matching*. Thus, we define:

$$dist(r,s) = \frac{sim_{\max} - \min\left(\sum_{g=1}^{w} sim(x_{rg}, x_{sg}), \sum_{g=1}^{w} sim(x_{sg}, x_{rg})\right)}{sim_{\max}}$$

where *w* is the number of groupings of variables, $sim(x_{rg}, x_{sg})$ is an asymmetric measure for the number of values in the *s* observation matching with the elements in the *r* observation for group *g*[3], and

$$sim_{\max} = \max_{r,s}\left(\min\left(\sum_{g=1}^{w} sim(x_{rg}, x_{sg}), \sum_{g=1}^{w} sim(x_{sg}, x_{rg})\right)\right)$$

is the overall observed maximum number of matches.

We performed several experiments, using different settings. For instance, while computing the (dis)similarity between observations, in some experiments we excluded specific kinds of words (like function words and pronouns). In other experiments, we chose different settings for grouping the variables. For instance, we grouped together the values of the first lemma preceding and the first following *forma*, and used this as one variable separated from another which features the second lemma preceding and the second following.

6. Evaluation and Results

In order to evaluate the results of our experiments, we built two gold standards:

- Gold standard A (GsA): we manually annotated the meaning of 672 randomly chosen occurrences of *forma*;
- Gold standard B (GsB): among the occurrences of GsA, we selected a subset of 356 featuring a easy-to-detect meaning of *forma*.

We used a tagset consisting of 10 labels defined according to both DEFERRARI & BARRY (1948-1949) and Latin Wordnet[4].

For each experiment, we evaluated the results by using several evaluation metrics, among which are precision and recall (VAN RIJSBERGEN, 1979: 112-113). We compute precision and recall according to both labels and clusters: thus, we talk about precision/recall of label *n* in cluster *x*. For instance, if a cluster *x* contains 100 observations (i.e. 100 occurrences of *forma*), 80 out of which are labelled with the same label *n* in the Gold

[2] The values in the columns of the observations are lemmas, which represent different word types.

[3] As multiple occurrences of one value may appear in a group, this measure is introduced in order to skip repetitions and to compute the value only once. For instance, this happens when the same lemma appears both as the first following *forma* in the text and as one of its attributives in the dependency tree, and, thus, it is reported twice in the observation.

[4] 1: *forma inhaerens, substantialis*; 2: *forma corporis*; 3: *forma artificiati*; 4: *forma naturalis*; 5: *forma praedicati*; 6: *forma materialis*; 7: *forma* as 'shape' (e.g. *forma domus*); 8: *forma accidentalis*; 9: *forma* as 'formula' (e.g. *forma baptismi*); 10: *forma participata*.

Standard, this means that the precision rate of label *n* in cluster *x* is 0.80. Conversely, if the total number of observations labelled with *n* in the Gold Standard is 80, the recall rate of label *n* in cluster *x* is 1.

The best performing setting on GsA (matrix A) is the following:

- grouping features: two separate groups, namely syntactic information (lemmas of parent, grandparent, attributive dependents and coordinated nodes) and textual information (lemmas of 2 words preceding and 2 following *forma* in the text);
- function words, pronouns and verb *sum* ('to be') excluded while computing (dis)similarity.

The best performing setting on GsA (matrix B) and on GsB (both matrices A and B) presents the following features:

- grouping features: the same two groups reported above;
- function words, pronouns and all verbs (not only *sum*) excluded while computing (dis)similarity.

Table 1 reports the best results achieved with these settings[5]:

	Precision	**Recall**
GsA (matrix A)	0.9545 (2)	0.8235 (5)
GsA (matrix B)	0.9444 (6)	0.6522 (8)
GsB (matrix A)	0.9687 (6)	0.8933 (6)
GsB (matrix B)	0.9375 (8)	0.8267 (6)

Table 1. Best results

While plotting the results of hierarchical clustering, we highlight in red some specific observations from GsA or GsB. This allows us to visually check if those observations annotated with the same label in the gold standard do indeed appear close to each other in the plot. Figure 1 reports the plot showing the results on GsB (matrix A): the observations tagged with label 6 are highlighted in red and do appear close to each other.

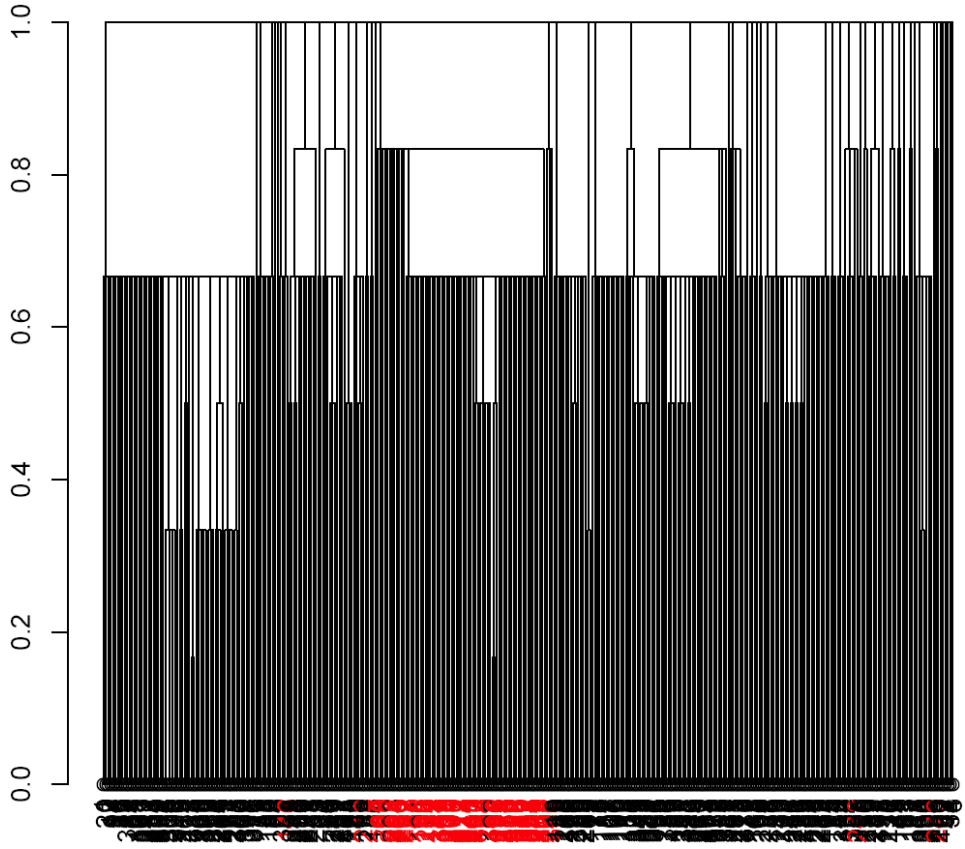


Figure 1. Plot of the results on GsB (matrix A)

[5] In the columns featuring precision and recall, the numbers in round parentheses correspond to labels in the evaluation tagset. For instance, the best precision rate on GsA (matrix A) concerns those observations that are tagged with label 2. Precision and recall are computed by cutting the dendrogram at the height of 0.8 (see Figure 1).

7. Conclusion and Future Work

Looking at results, we can conclude the following:

- while calculating the (dis)similarity between observations, excluding function words and pronouns leads to higher results;
- excluding the verb *sum* leads to higher results on GsA, while excluding all verbs leads to higher results on GsB;
- recall rates on matrix B are much lower than on matrix A; precision rates on matrix B are fairly lower than on matrix A;
- results on matrix A are always higher than on matrix B.

Our general conclusion is that syntactic metadata (available in matrix A, but not in matrix B) are indeed helpful for clustering purposes, particularly if recall is concerned. We must consider this aspect in our future work, when we will have to decide which metadata from annotation we should use while applying other clustering techniques, as for instance Latent Semantic Analysis.

Considering the above mentioned relation between LRs, we would like to exploit Latin Wordnet to perform word sense disambiguation, associate sense labels to clusters, and use second-order similarity.

We would also like to allow partial matches between variables. When full matches are allowed, lemmas are considered equal to themselves only (for instance, *baptisma* is equal to the value *baptisma* only). Instead, when partial matches are allowed, lemmas are considered equal to a set of values (for instance, *baptisma* is equal not only to *baptisma*, but also to *baptismalis*, *baptizo* and other values, i.e. lemmas sharing the same morphological root).

References

BUSA Roberto (1974-1980), *Index Thomisticus*, Stuttgart-Bad Cannstatt, Frommann-Holzboog.

DEFERRARI Roy J.; BARRY M. Inviolata (1948-1949), *A Lexicon of St. Thomas Aquinas: based on the* Summa Theologica *and selected passages of his other works*, Washington, Catholic University of America Press.

FIRTH John R. (1957), *Papers in Linguistics 1934-1951*, London, London University Press.

HARRIS Zelig (1954), Distributional Structure, *Word 10*, p. 146-162.

KAUFMAN Leonard; ROUSSEEUW Peter J. (1990), *Finding Groups in Data: An Introduction to Cluster Analysis*, New York, Wiley.

MAECHLER Martin; ROUSSEEUW Peter J.; STRUYF, Anja; HUBERT Mia; HORNIK Kurt (2011), *cluster: Cluster Analysis Basics and Extensions*, R package version 1.14.1. http://CRAN.R-project.org/package=cluster.

PEDERSEN Ted (2006), Unsupervised corpus-based methods for WSD, in: AGIRRE Eneko; EDMONDS Philip (eds.), *Word Sense Disambiguation: Algorithms and Applications*, New York, Springer, p. 133-166.

SGALL Petr; HAJIČOVÁ Eva; PANEVOVÁ Jarmila (1986), *The Meaning of the Sentence in its Semantic and Pragmatic Aspects*, Dordrecht NL, Reidel.

SOKAL Robert R.; MICHENER Charles D. (1958), A statistical method for evaluating systematic relationships, *Univ. Kansas Sci. Bull. 38*, p. 1409-1438.

TESNIERE Lucien (1959), *Éléments de syntaxe struc-turale*, Paris, Editions Klincksieck.

VAN RIJSBERGEN 'Keith' C.J. (1979), *Information Retrieval*, London, Butterworths.

False diminutives in Brazilian Portuguese

Roana Rodrigues, Oto Araújo Vale

Abstract. This study analyses the suffix -inho and -inha on nouns in order to establish a typology of false diminutives in Brazilian Portuguese. This phenomenon is often mentioned in literature, but it was never described in detail. In addition to the descriptive interest itself, we intend to contribute to the creation of resources for Natural Language Processing (NLP).

1. Introduction

It is common to find in Brazilian Portuguese the same morphemic constructions – marked by derivation – with different meanings. It is exactly in these constructions that one is able to notice the existence of false augmentatives and false diminutives. That is, the suffix *-ão* in *portão* (entrance/gate) is not the augmentative of *porto* (harbor) or *porta* (door), as *-inha* in *calcinha* (panties) is not the diminutive of *calça* (trousers). We do not disregard the fact that a considerable amount of lexemes present the ending *-zinho* and *-zinha*, but actually we assume, along with BISOL (2010), that there is only one morpheme, *-inho/a*, which is realized with an epenthetic consonant [z], in order to satisfy structural needs, in lexemes as *flor* (flower) > *flor-z-inha*.

This paper aims at describing false diminutives on nouns in Brazilian Portuguese found in the two major dictionaries currently available for Brazilian Portuguese and also in NILC corpus. In doing so, we attempt to enhance computational resources which deal with this kind of phenomena, especially the ones that use electronic dictionaries.

2. Related works

There are few studies about false augmentatives and false diminutives in Brazilian Portuguese among which it is possible to mention (BECHARA, 2002; CAMARA JR. 1970, 1971; LAROCA, 2005; SANDMAN, 1992), who described morphological phenomena in Brazilian Portuguese. Nevertheless, in their analysis, the authors only show that there is an important distinction between the processes of inflection and derivation, in which the diminutive morphemes we intend to describe play a relevant role, in the sense that they contribute to the creation of new words, a process that is not obligatory nor regular. Apart from that, there is nothing else about false diminutives in Brazilian Portuguese. As these lexemes may pose some problems concerning, for example, disambiguation and stemming, it is worth classifying them, for a better understanding of the language.

3. Diminutives and false diminutives in the dictionaries

Initially, for the systematization of false diminutives, we have consulted the Dictionaries Aurélio (FERREIRA, 2009) and HOUAISS (2009), and a list of lexical units ending with *-inho* and *-inha* was compiled. This list consists of 948 entries (479 ending in *-inho* and 469 ending in *-inha*) found in both dictionaries. Following the list compiled in the dictionaries, we have proposed a classification in which some features where taken into account, as the diminutive form, the original form, the false diminutive and the synonym. Besides, due to the considerable amount of occurrences, we have decided to include the semantic fields “plants”, “birds”, “fish”, “aguardente” (Brazilian spirit made of sugar cane), that are recorded in the dictionaries, as it can be seen in table 1:

Table 2. classification of lexical units ending in *–inho* e *-inha* in the Dictionaries Aurélio and Houaiss

Aurélio	Houaiss	Diminutive form	Original Form	Original Diminutive	False Diminutive	Sinonym	Plant	Bird	Fish	Aguardente	POS-DimForm	POS-OrigForm	Intensity	Youth	Afectivity
+	+	adivinha	adivinha	+	-	+	-	-	-	-	-	N	-	-	-	-
+	+	amarelinha	amarelo	-	-	+		+	-	-	-	N	Adj	-	-	-
+	-	azulinha	azul	-	-	+	-	-	-	-	+	N	Adj	-	-	-
-	+	devagarinho	devagar	-	+	-	-	-	-	-	-	Adv	-	+	-	-
-	+	mocinha	moça	-	+	+	-	-	-	+	-	N		-	+	-
-	+	povinho	povo	-	-	+	-	-	-	-	-	N	-	-	-	-

4. Diminutives and false diminutives in NILC corpus

We have also analysed the occurrences of lexemes ending with *-inho* and *-inha* found in NILC Corpus, in order to observe how these morphemes were used. This corpus is composed by texts in different genres, as journalistic and didactic, among others. We have found 80 thousand occurrences of lexemes ending in *-inho*, *-inha*, *-inhos*, *-inhas* in the corpus, that refers to 1392 different entries (693 ending with *–inho*, *-inhos* and 699 ending *–inha*, *-inhas*). Some of these uses are shown in Table 2:

Table 3. classification of lexical units ending in *–inho* e *-inha* in NILC corpus

Diminutive Form	Original Form	Original	Diminutive	False Diminutive	Synonym	POS FD	POS OF	Occurrence DIM	Occurrence FD1	Occurrence FD2	Intensity	Youth	Affection
amarelinha	amarela	-	-	+	-	N	Adj		a ingenuidade de uma criança pulando **amarelinha**	Há jogadores que tremem com a **amarelinha**	-	-	-
cestinha	cesta	-	+	+	-	N	N	acomodaram **as cestinhas** e embrulhos de lembranças que traziam de Vassouras	No jogo, foi o **cestinha**, fazendo 29 dos 127 pontos do time do Leste.		-	-	-
devagarinho	devagar	-	+	-	-	Adv	Adv	Vou **devagarinho**, como der, mas vou, garantiu			+	-	-
paradinha	parada	-	+	+	-	N	N	não dispensa aquela **paradinha** na rodovia dos Tamoios	Romário deu a famosa **paradinha** a ntes de marcar o gol		+	-	-

We noticed that the diminutive lexical units refer not only to something small, but are also understood as something that may denote intensity (*fino* x *fininho* – slim x very slim), youth (*mocinha* x *jovem moça* – girl x young girl), and affection (*amorzinho* – little love – showing caress/*povinho* – little people – pejorative).

Even though the suffixes *–inho* and *-inha* are used in order to convey a specific meaning (intensity, youth, affection) in a certain context, we have observed that there is a substantial amount of entries that refer to false diminutives: 192 entries of *-inho* and 254 of *–inha*, as for example *carinho* (lit: little darling = care) and *fominha* (lit: little hunger = someone who is mean, as is soccer, a player who does not pass the ball).

5. Some values of false diminutives

Some definitions found in the dictionaries allowed us to classify the false diminutives, as they were synonymous to their 'original forms'. In NILC corpus, however, the number of entries of this kind was drastically reduced, once these derivational suffixes are rarely used as their 'original forms'. They are generally used when one wants to convey some sort of evaluation, judgements, politeness strategies, despise etc. In this sense, we may agree with ALVES (2006), who considers that a discursive act itself is not only realized in linguistic units as the text and the utterance, but also at the word level, since the use of the suffixes we described brings about new meanings and shows intentions of the text in a whole.

Moreover, we have observed in NILC corpus that some frozen expressions in Brazilian Portuguese are mostly, when not obligatorily, composed by lexemes ended in *-inho* and *-inha* as in:

> *cavalinho*: as in *Pode ir logo tirando o cavalinho da chuva* (lit: little horse – take you little horse out of the rain = "You may forget it");
>
> *figurinha*: as in "*as duas bandas trocavam muita figurinha no começo* (lit. little card – the two bands use to share little cards in the beginning = "The two bands had used to cooperate in the beginning");
>
> "*lasquinha*": as in *muitos patos se aproximam, e tiram uma lasquinha* (lit: little chips – many ducks approach and take a little chip = "Many ducks take advantage");
>
> Regular lexemes ending in *-inho* and *-inha* also observed especially in NILC corpus refer to sports, as in:
>
> *bandeirinha* (assistant referee in soccer);
>
> *carrinho* (sliding tackle, in soccer);
>
> *cestinha* (scoring leader in a basketball game).

It is possible to consider, thus, that the results were satisfactory, as from the lists we produced one is able to notice how false diminutives are used in different situations and contexts in Brazilian Portuguese. This preliminary typology allowed us to attest which cases are actual uses of false diminutives, as the lexemes calcinha, camisinha and sardinha. We have also shown that the use of false diminutives in Brazilian Portuguese poses a problem for

their description, and that the scope of their use is far more important than the works that deals with them seem to consider.

6. Future work

Taking into account the results obtained in the list, our preliminary typology of flase diminutives is been improved. In this new typology, we intend to compare the units with the ones found in electronic dictionaries that are part of the Unitex Software (PAUMIER,2002; MUNIZ et al 2005). Our goal is to enhance these electronic dictionaries, in order to turn the searcher more précised, avoiding thus unnecessary ambiguities.

We hope that this investigation may contribute to sentiment analizes systems, once the units we have listed could be considered as affection markers. Another possible application is the enhancement of analyses based on stemming. The use of our results could avoid that entries as *camisinha* (lit: little T-shirt = condom) were reduced to *shirt* or *carrinho* (lit: little car = sliding tackle, in soccer), as in example 5 above, to *car*, a reduction that could imply an obvious lack of information. It would be also interesting to observe whether systems fed with this kind of information are more precise in their searches.

7. References

ALVES Elisabeth (2006) , O diminutivo no português do Brasil: funcionalidade e tipologia. *Estudos linguísticos* XXXV. Campinas, GEL, p. 694-701.

BECHARA, Evanildo (2002) Moderna Gramática Portuguesa. Rio de Janeiro, Lucerna.

BISOL Leda (2010), O diminutivo e suas demandas. *DELTA*, São Paulo, v. 26, n. 1, p. 59-85

CAMARA Jr., Joaquim Mattoso (1970), *Estrutura da língua portuguesa*. Petrópolis: Vozes.

CAMARA Jr., Joaquim Mattoso (1971), *Problemas de linguística descritiva*. Petrópolis: Vozes.

FERREIRA, Aurélio Buarque de Holanda (2009), *Novo dicionário Aurélio da língua Portuguesa*. Rio de Janeiro: Positivo.

HOUAISS, Antonio (2009) *Dicionário Houaiss da língua portuguesa*. São Paulo: Objetiva.

LAROCA Maria Nazaré de Carvalho (2005), *Manual da morfologia do português*. Campinas: Pontes.

MUNIZ Marcelo ; NUNES Maria das Graças Volpe ; LAPORTE, Eric, (2005) UNITEX-PB, a set of flexible language resources for Brazilian Portuguese. In: *Workshop em Tecnologia da Informação e da Linguagem Humana - TIL, 2005 anales*, São Leopoldo, RS v. 1. p. 2059-2068.

PAUMIER, Sesbastien (2002), *Unitex Manual*. Université Marne-la-Vallée.

SANDMAN, Antonio José (1992), *Morfologia Lexical*. São Paulo: Contexto.

L'attribut dans les constructions impersonnelles

Yoichiro Tsuruga

Abstract: The complement (French « attribut ») is a syntactical function and presupposes the subject in personal constructions. But in impersonal constructions, it can function without a syntactical personal subject nor a so-called semantic « real » subject : *Il est tard*. In *Il semble que Luc chante*, it is *que Luc chante* that constitutes a complement. The next derivation is not acceptable : **Que Luc chante semble→+Il→Il semble que Luc chante*. More natural is the next one (the complement is inside the parenthesis [ ]): *Luc chante→+sembler→Luc semble* [*chanter*]*→+Il→*Il semble* [*chanter*] *Luc→+que→Il semble* [*que Luc chante*]. In *Il semble* [*que Luc chante*], *chante* is the core of the complement. *Paraître* has an intransitive absolute construction: *Un livre paraît→Il paraît un livre*. But in *Il paraît* [*que Luc chante*], it is *que Luc chante* that is a complement. The impersonal construction permits in this way the complement to function without a syntactical nor semantic subject. It is furthermore not rare to find utterances like *Il est* [*temps*], *Il est* [*9 heures*], etc.

1. Introduction

Le sujet qui est à gauche du verbe intransitif dans la construction personnelle se place à sa droite dans la construction impersonnelle: *Un accident arrive→Il arrive un accident*. Ce déplacement est difficile pour les verbes transitifs à deux arguments: *Des gens mangent du fromage→*Il mange du fromage des gens*. C'est que les verbes intransitifs ont une place vide à leur droite, tandis que les transitifs n'en ont pas. Les verbes attributifs ont deux arguments apparents, sujet et attribut: *Luc est* [*content*]. (L'attribut [1)] sera entre [ ], ci-dessous.) Mais certaines constructions attributives permettent le déplacement en question: (*De*) *chanter est* [*intéressant*]*→Il est* [*intéressant*] *de chanter*. Cette construction personnelle a en fait un seul argument: (*De*) *chanter* [*est* [*intéressant*]] [2)]. L'attribut peut être assumé par différentes formes: noms (*Luc est* [*Français*]), infinitifs (*Vouloir, c'est* [*pouvoir*]), adjectifs (*Luc est* [*gai*]), adverbes (*C'est* [*loin*]), syntagmes variés (*Luc est* [*d'esprit simple*], *Cela est* [*à la mode*], *Le fait est* [*qu'il t'aime*]). Il est ainsi quelquefois délicat d'identifier ce qu'est l'attribut dans les constructions impersonnelles.

2. Les constructions de *sembler*

L'adjectif peut, sans problème, être identifié comme attribut. L'adjectif en tant que tel ne peut assumer la fonction sujet syntaxique ni le sujet sémantique dit « réel ». Il n'assume pas la fonction d'objet direct ni celle d'objet indirect.

(1) Que Luc chante semble [*curieux*]→Il semble [*curieux*] que Luc chante.

Ci-dessus, quand il y a un syntagme nominal et un adjectif autour de *sembler*, verbe ayant besoin d'un attribut, c'est l'adjectif qui est attribut. Dans la construction impersonnelle, la position du sujet syntaxique est occupée par *Il*, et à droite du verbe il y a un adjectif et une proposition nominale en *que* (*que*V, ci-dessous). Formellement, l'attribut peut être assumé par un adjectif ou une proposition *que*V, mais c'est *curieux* qui est prioritairement attribut, car l'adjectif ne peut y assumer d'autres fonctions.

(2) *Que Luc chante semble—Il semble [*que Luc chante*].

Ci-dessus, le verbe *sembler* avec le sujet *que*V ne peut constituer une phrase intransitive absolue, car *sembler* a besoin d'un attribut. Dans la construction impersonnelle, c'est la fonction attribut qu'assume *que*V. *Que*V pourrait formellement assumer le sujet sémantique, mais ce n'est pas le cas ici. Et il est hors de question de considérer ce *que*V comme objet, car même dans les constructions personnelles de *sembler*, il n'y a aucune position possible pour la fonction d'objet direct (cf. la passivation impossible; voir aussi RUWET, 1987 : 67). La dérivation suivante est plus naturelle: *Luc chante→+sembler→Luc semble* [*chanter*]*→+Il→*Il semble* [*chanter*] *Luc→+que→Il semble* [*que Luc chante*] (cf. *chante* est le noyau de l'attribut [*que Luc chante*]).

(3) Et dans les rêves, à *ce* [*qu'*]*il me semble*, ces choses-là ne comptent pas. (*Le Monde* 1994)

Dans *ce* [*qu'*]*il me semble*, ci-dessus, le pronom relatif *qu'* assume la même fonction que celle de *que*V de (2). Dans la construction impersonnelle de *sembler*, l'attribut nominal, seul suivi de rien, est difficile à accepter: **Il semble une découverte*. Mais dans le cas de « pronom démonstratif+pronom relatif », l'attribut pronominal seul est acceptable.

(4) [*Que*] te semble *de cette affaire* ?

[1)] L'attribut peut, en principe, être défini comme une fonction prédicative dont a besoin un verbe « copule » dans la construction personnelle. L'attribut de l'objet direct ne fait pas l'objet principal de cet article.
[2)] [[ ]]: partie prédicative.

La forme *que* ne peut assumer le sujet. Par conséquent, même sans *il*, l'énoncé ne peut être qu'impersonnel. On peut suppléer *il*: [*Que*] *te semble-t-il de cette affaire?* Dans cet énoncé-ci, c'est le pronom interrogatif *que* seul qui peut être attribut. *De cette affaire* représente une fonction de « propos ». Sémantiquement, cet énoncé peut avoir un lien avec la construction personnelle: *Cette affaire*, [*que*] *te semble-t-elle?* Dans (4), il s'agirait alors de l'attribut d'un élément ayant la fonction indirecte de « propos » (et non de celui ayant la fonction sujet ou objet direct). Rappelons que traditionnellement quand on parle d'attribut, il est question ou bien d'un sujet ou bien d'un objet direct [3)].

(5) Prenez ce que [*bon*] vous semble.

Dans (5), il y a plus d'éléments à suppléer que dans (4). On peut rétablir: *Prenez ce qu'il vous semble* [*bon*] *de prendre*. *Ce qu'il vous semble* est acceptable avec *à* (cf. (3)), mais ne l'est pas dans ce contexte: **Prenez ce qu'il vous semble* (cf. *Prenez ce qu'il vous semble* [*bon*]).

3. Les constructions de *paraître*

Le verbe *paraître* a un emploi intransitif absolu, ce qui est différent du cas de *sembler*.

(6) Un nouveau livre paraît demain→Il paraît un nouveau livre demain.

Ci-dessus, avec *demain*, élément temporel, *paraître* fonctionne comme verbe de « parution » qui n'a pas besoin d'un attribut. L'élément temporel n'est même pas nécessaire.

(7) *Que l'euro monte paraît—Il paraît [*que l'euro monte*].

*Que*V, pourtant, assume difficilement le sujet de l'intransitif absolu. Dans la construction impersonnelle de (7), *que*V ne peut donc pas être sujet sémantique. *Que*V ne peut assumer que la fonction attributive exactement comme dans *Il semble* [*que l'euro monte*].

(8) Que l'euro monte paraît [*curieux*]→Il paraît [*curieux*] que l'euro monte.

*Que*V de (8) assume le sujet syntaxique ou sémantique, respectivement dans les constructions personnelle et impersonnelle. Ces constructions sont les mêmes que celles de *sembler*.

(9) ?Que l'euro monte paraît dans le journal→Il paraît dans le journal que l'euro monte.

Avec un locatif *dans le journal*, l'énoncé (9) personnel n'est pas inacceptable (cf. (7)). Et la construction impersonnelle est acceptable avec *paraître* de « parution ». Remarquons toutefois que l'on trouve difficilement des énoncés comme (9), personnel ou impersonnel. (Cf. Aucun exemple dans nos corpus de 1.496 occurrences du *Monde* 1994 et de *Frantext* 1986-97.)

(10) […] les diplomates russes multipliaient des déclarations ambiguës, *d'où il paraissait* [*ressortir*] *que* si la Russie désirait vivement bloquer l'action des Occidentaux *elle n'en avait pas vraiment les moyens*. (*Le Monde*, 1994)

Dans *il paraissait* [*ressortir*] *que* […], c'est l'infinitif *ressortir* qui assume la fonction attributive (cf. GREVISSE, 2001[13] : §245).

En comparant *Luc paraît* [*content*] et *Luc paraît* [*chanter*], par exemple, il est difficile de ne pas reconnaître que *content* et *chanter* appartiennent exactement au même paradigme attributif. Traditionnellement, on considérerait *content* comme attribut, tandis que *paraît* est un auxiliaire modal de *chanter*. En fait, ces analyses ne s'opposent pas. Dans le premier énoncé, *paraît* joue le rôle de la copule *est* de *Luc est content*. La copule est un élément actualisateur de ce qui suit. Elle aide son successeur à fonctionner comme attribut qui est centre de l'énoncé, soit prédicat. Dans le second énoncé, si *paraît* est auxiliaire, c'est qu'il aide son successeur à fonctionner comme prédicat. Dans les deux cas, il faut reconnaître que le centre formel de l'énoncé est le verbe fini *paraît*. Si, donc, on observe les énoncés du point de vue de *paraître* même, l'adjectif *content* et l'infinitif *chanter* assument la même fonction prédicative.

On peut avoir comme contruction personnelle *Que*V-[*paraît* [*ressortir*]] *de cela*. Si *paraît ressortir* est conçu comme prédicatif, il y a à gauche un syntagme nominal *que*V et à droite un syntagme indirect avec *de*. La position d'un syntagme nominal à droite du verbe n'est pas encore remplie. Et si l'on introduit *il* et que le complétif *que*V se déplace à droite de la partie prédicative, on aura *Il paraît* [*ressortir*]-*que*V-*de cela*. Et avec *de cela* antéposé, on aura *De cela il paraît* [*ressortir*]-*que*V. Fonctionnellement, *ressortir* ne peut être ni sujet sémantique ni objet direct. Il ne peut être qu'attribut. C'est *que*V qui est le sujet sémantique.

3) Cf. Un exemple de l'attribut de l'objet indirect: « *Peut-être certains* (...) *jugeront du tout* [*comme peu novateur*]*; d'autres* (...) *le jugeront* [*visionnaire*]. » (*Frantext*, 1969). Voir aussi les exemples (19)-(22), ci-dessous.

4. Les constructions de *être*

On peut reconnaître deux *être* dans *Luc est* [*content*] (→*Luc* [*l'*]*est*) et *Luc est à Paris* (→*Luc y est*). Mais l'emploi de *être* sans attribut ni aucun autre élément support comme dans *Même si cela ne vous plaît pas, cela est* est exceptionnel (cf. deux exemples seuls dans nos corpus de 2.122 occurrences du *Monde* 1994 et de *Frantext* 1980-82).

Mais cet *être* accompagné d'un élément locatif (*Luc est là*) ou temporel (*Le rendez-vous est à trois heures*) est si fréquent qu'on est tenté de considérer *être* comme inacceptable s'il n'est suivi de rien. Ce qui accompagne *être* est ainsi très proche de l'attribut dont la caractéristique essentielle est celle d'être indispensable. En plus, l'emploi de *être* comme copule a un sens affaibli et le sens de *être* suivi d'un élément locatif n'est pas si manifeste que celui de l'emploi absolu. Et ce sens faible de *être* s'affaiblit encore plus quand le sujet est impersonnel, puisqu'il n'y a, à sa gauche, aucune fonction avec le paradigme ouvert qui le supporte. Et il y a souvent des expressions figées. Tout cela rend quelquefois difficile d'identifier ce qu'est l'attribut dans la construction impersonnelle de *être*.

(11) [...] *c'est* [*dans cette partie du monde*] que l'on parle le plus de prolifération nucléaire. (*Le Monde* 1994)

Ce peut être considéré comme impersonnel dans la mesure où son paradigme est fermé. Et *dans cette partie du monde* est locatif mais aussi attribut. La relation entre *est* et *dans cette partie du monde* est la même que celle entre *est* et *Luc* dans *C'est* [*Luc*] *qui chante* ou *est* et *ainsi* dans *C'est ainsi que Luc chante*. *Est* ne signifie pas « exister ».

(12) Il est [*tard*].

(13) Il est [*temps*], Il est [*6 heures*].

Dans (12), ci-dessus, *tard*, adverbe, ne peut être qu'attribut, et *temps* et *6 heures* de (12) appartiennent au même paradigme.

(14) *Il n'est bon champagne* que de France.

Il est, ci-dessus, a le sens proche de *il y a*. Si, donc, *bon champagne...* n'est pas attribut, cette partie qui suit *être* doit-elle être considérée comme sujet sémantique ayant une caractéristique d'un objet direct? Peut-on supposer la dérivation suivante?: *?*Bon champagne de France est*→**Bon champagne de France n'est que*→**N'est que bon champagne de France*→**Il n'est que bon champagne de France*→*Il n'est bon champagne que de France*.

(15) (...) ou ce n'est pas amusant, et on ne rit pas, ou c'est amusant, et on rit, sans qu'*il soit besoin* de nous le faire savoir. (*Ibid.*)

Il soit de *il soit besoin...*, ci-dessus, est aussi interprété comme *il y a*.

(16) Il est [*question*] de "délivrer le pays d'un chancre (...)". (*Ibid.*)

Question, ci-dessus, est attribut et *de délivrer...* est une fonction de « propos » (cf. *Il en est* [*question*]). La construction de (16) est formellement identique à celle de (15). La différence vient seulement de celle existant entre *besoin* et *question*. La construction *si besoin est* est acceptable, tandis que ce n'est pas le cas de *question*.

(17) Toujours est-il [*certain*] que Luc chante.

(18) Toujours est-il [*que Luc chante*].

Dans (17), c'est *certain* qui est attribut et *que Luc chante*, sujet sémantique. Et si l'on identifie que *Toujours* et *est* restent les mêmes dans (17) et (18), il n'y a, dans (18) que *que Luc chante* qui peut être attributif. C'est comparable à : *Il semble* [*curieux*] *que Luc chante* et *Il semble* [*que Luc chante*]).

(19) Voilà *ce* [*qu'*]il *en* est.

(20) *Quoi* [*qu'*]il *en* soit *des affirmations* des uns et des autres, le gouvernement iranien avait formulé des menaces à peine voilées (...). (*Le Monde*, 1994)

Dans (19), *qu'* est attribut. *En* représente une fonction de « propos » assez vague.

Il en va de même pour (20). *Qu'* est attribut et il y a deux fonctions de « propos », l'un (*en*) vague et l'autre (*des affirmations...*) explicitée. Deux fonctions identiques sont compatibles si elles ont des extensions différentes.

(21) Il en est [autrement/ ainsi/ de même] de lui.

(22) Il en est de lui [comme de nous].

Dans (21) *autrement*, par exemple, est attribut. Et dans (22), *autrement* est remplacé par *comme de nous*. *En* et *de lui* assument la même fonction que celle de *en* et de *des affirmations* de (20). On identifie bien sûr le même *est* dans (19)-(22).

Dans les constructions personnelles, normalement l'attribut suppose comme sujet un élément nominal direct. Dans les constructons impersonnelles, c'est *il* qui le remplace. Mais *il* est sémantiquement vide, et il peut là y avoir

un sujet nominal « réel » qui est aussi l'équivalent d'un direct (le cas de (17)). L'énoncé attributif impersonnel de *être* peut en fait être indépendant avec *il* formel, *être* et attribut, mais sans sujet « réel » (cf. le cas de (12) *Il est tard*). Mais un élément indirect correspondant sémantiquement au sujet nominal direct peut accompagner cet énoncé attributif impersonnel. Ce sont les cas de *en* de: (19)-(22), de *de cette affaire* de (4), de *des affirmations* de (20), de *de lui* de (21) et de (22). Dans les constructions attributives impersonnelles, il y a un sujet formel *il*, mais un sujet « réel » n'est pas indispensable. C'est d'ailleurs là l'essentiel de l'impersonnel. Mais un sujet « réel », en principe de forme directe, peut apparaître sous une forme indirecte. Un élément sémantique peut ainsi s'insérer sous différentes formes.

(23) Il n'est pas [*que vous ne sachiez son nom*]. (vieux)

C'est une construction vieillie signifiant à peu près « il n'est pas [*vrai*] que... ». C'est la même construction que celle de (18). Ici aussi, l'attribut seul est présenté avec l'élément formel *il*.

5. Conclusion

Dans les constructions attributives personnelles : "Sujet-Verbe« copule »-Attribut" (*Luc semble* [*gai*]), l'attribut présuppose la fonction sujet. Mais dans les constructions attributives impersonnelles, le sujet qui accompagne l'attribut est *Il*, impersonnel, qui est sémantiquement vide : *Il est* [*tard*], *Il est* [*temps*], *Il semble* [*que Luc chante*]. "Sujet$_{personnel}$-Partie prédciave" constitue un cadre binaire, tandis que "*Il*-Partie prédicative", un cadre qui est en fait unaire. L'attribut (= prédicat) est indispensable, mais ce n'est pas le cas du sujet personnel, syntaxique ou sémantique. Rappelons aussi que l'attribut peut aussi être accompagné d'un sujet sémantique de forme indirecte.

Références

DE GAULMYN, M.-M. et REMI-GIRAUD, S. (dir.) (1991), *A la Recherche de l'attribut*, Lyon, PUL.

DEULOFEU, J. (2011), « Deux notions toxiques en linguistique française: *prédicat* et *prédication* », G. CORMINBOEF et M.-J. BÉGUELIN (dir.) *Du système linguistique aux actions langagières, Mélanges en l'honneur d'*Alain BERRENDONER, Bruxelles, De Boeck et Duculot, pp. 125-146.

ERIKSSON, O. (1980), *L'Attribut de location et les nexus locatifs en français moderne*, Götegorg, Kompendiet, Londome.

GREVISSE, M. (1939, 2001[13]), *Le Bon usage, Grammaire française*, refondue par A. GOOSSE, Gembloux, Duculot.

GROSS, M. (1975), *Méthodes en syntaxe*, Paris, Hermann.

KURODA, S.-Y. (1973), « Le jugement catégorique et le jugement thétique : exemples tirés de la syntaxe japonaise », *Langages* 30, p. 81-110.

LAMIROY, B. (2012, à paraître), « L'Impersonnel », A. ABEILLÉ, D. GODARD (éd.) *Grammaire de référence du français*, 2 vols, Paris, L'Harmattan, 12 p.

MARTINET, A. (1979), *Grammaire fonctionnelle du français*, Paris, Crédif, Didier.

RIEGEL, M. (1988), *L'Adjectif attribut*, Paris, PUF.

RIVIÈRE, N. (1981), *La Construction impersonnelle en français contemporain*, Paris, Éd. Jean Favard.

RUWET, N. (1982), « Montée du sujet et extraposition », *Grammaire des insultes*, Paris, Éd. du Seuil, p. 27-75.

RUWET, N. (1991a), « Raising and Control revisited », *Syntax and Human Experience*, Chicago, The Univ. of Chicago Press, p. 56-81.

VAN PETEGHEM, M. (1991), *Les phrases copulatives dans les langues romanes*, Rainweg, Gottfried Egert Verlag.

Les corpus

Le Monde, 1994.

Frantext, 1980-1997.

A Lexicon of Verb and *–mente* Adverb Collocations in Portuguese Extraction from Corpora and Classification

Lucas Nunes Vieira, Cláudio Diniz, Nuno Mamede, Jorge Baptista

1. Introduction

Collocations started to be a target of research in the twentieth century after FIRTH (1957) coined the term and called attention to the fact that the way we combine words in natural language is far from being unconstrained. In the sense of Firth, a pair or group of words can be considered a collocation if the probability for their co-occurrence exceeds chance levels. For a long time this concept has prevailed in the literature as the rationale behind the task of collocation extraction. However, the more recent formulation of MEL'ČUK (2003) provides a more semantic-based view on the phenomenon that does not necessarily coincide with an attested high frequency of the word combinations. According to Mel'čuk, the meaning of certain words would dictate the adjacent use of others, forming groups or pairs of *base* words and *collocates*.

Concerning the linguistic pattern investigated in this study, namely pairs of verb and *–mente* ('-ly') ending adverbs, the verb would be the base of the combination, while the adverb would be the collocate. The strategy here adopted for the extraction of this pattern profits both from Firth's and Mel'čuk's formulations, since, at different stages, it relies both on frequency of distribution and on meaning-oriented human annotations.

The corpus used for the extraction of verb-adverb bigrams was the CETEMPúblico[1] (SANTOS & ROCHA, 2001) corpus of European Portuguese, consisting of 192M words of journalistic texts. This is, to the best of our knowledge, the largest freely distributed corpus of Portuguese. Albeit constituting just over 10% of all simple adverb occurrences in the corpus, adverbs ending in *–mente*, henceforth *Adv-mente*, represent in fact the majority of the simple-word lemmas of this grammatical class, based on data from the CETEMPúblico.

While a number of initiatives at collocation extraction have relied substantially on a search for adjacent words, as CHOUEKA (1988), newer studies suggest that methods involving some level of syntactical parsing might prove more precise for certain linguistic patterns (SERETAN, 2011). In view of this, we experiment with a syntax-based approach for the extraction of verb-adverb pairs from the corpus. This pattern could be deemed a challenging one in respect to the extraction task, since adverbs can occupy different positions in the sentence, being commonly associated with a rather loose mobility in the speech (BECHARA, 2003).

In the remainder of this paper, we describe the syntax-based approach adopted to extract collocations from the corpus (Section 2), explain the linguistically-motivated classification of collocation candidates (Section 3), and present an empiric evaluation of statistical association measures in identifying {*V, Adv-mente*} collocations (Section 4). We conclude by discussing the appropriateness of the methods experimented with in view of {*V, Adv-mente*} pairs and by proposing future work (Section 5).

2. Collocation extraction

2.1 Syntactically parsing the corpus

In order to illustrate some of the pitfalls that the linguistic pattern investigated may pose to the task of collocation extraction, a potentially problematic context is presented below:

(1) *O aluno leu o livro atentamente e resumiu-o*

« The student read the book attentively and summarised it »

Even though the *Adv-mente* in this example co-occur with two verbs, *ler* « to read » and *resumir* « to summarise », it only holds a syntactical dependency with the verb *ler* « to read », forming the pair *ler atentamente* « to read attentively », which can be considered to have collocational status in Portuguese. However, there is a high probability that the pair *resumir atentamente* « to summarise attentively » would erroneously come up as a collocation candidate in a search for adjacent words, since the adverb co-occurs with the two verbs in considerable proximity in the same sentence.

In order to overcome similar problems, the strategy described in this paper included the syntactical parsing of the corpus envisaging a more precise extraction of verb-adverb pairs that actually hold a syntactical dependency between them. The STRING computer-based text processing chain (MAMEDE ET AL., 2012) has been used for that purpose. In broad terms, the chain comprises three main stages: pre-processing, disambiguation, and syntactic analysis, respectively. The pre-processing stage is responsible for text segmentation, part-of-speech (POS) tagging, and for the splitting of the input into sentences. In the last stage, the syntactic parsing of the text is performed by XIP (Xerox Incremental Parser) (AÏT-MOKHTAR et al., 2002), a rule-based finite-state parser that establishes syntactic dependencies between words. In this framework, the output provided by the system includes a

[1] http://www.linguateca.pt/CETEMPublico/ [Accessed 5 May 2012]

dependency relation (called MOD[fier]), with the correct verb-adverb pair. An example of the output yielded for sentence (1) is provided below:

```
VDOMAIN(leu,leu)
VDOMAIN(resumiu-o,resumiu-o)
MOD_POST(leu,atentamente)
SUBJ_PRE(leu,aluno)
SUBJ_PRE_ANAPH0(resumiu-o,aluno)
CDIR_POST(resumiu-o,o)
CDIR_POST(leu,livro)
TOP{NP{O aluno} VF{leu} NP{o livro} ADVP{atentamente} e VF{resumiu-
o} NP{o}}
```

After processing the entire corpus, MOD dependency relations of this kind between verbs and *Adv-mente* were extracted from the output. The number of verb-adverb bigrams obtained was of 65,535, whose frequency in the corpus exceeds 290,000 occurrences altogether.

2.2 Filtering the extraction output

In order to narrow down the universe of potential cases of collocations, a number of filtering strategies were applied to the total set of bigrams extracted from the corpus so as to eliminate from the outset cases that did not present any potential of forming collocations. A frequency threshold of five ($f \geq 5$) was established for the consideration of pairs as collocation candidates. This is a threshold that can be deemed considerably low given the fact the total number of words in the corpus is 192M.

With regard to *Adv-mente*, we have augmented the adverbial classification carried out by FERNANDES (2011) for *Adv-mente* in Portuguese, which initially covered approximately 520 adverbs. This number has now been increased to nearly 1,000.

Having knowledge of the class or classes a given *Adv-mente* belongs to played an important role in filtering out adverb categories that do not hold a straight connection with the verb, which consequently impedes the formation of verb-adverb collocations. That would be the case of adverbs that play the single role of modifying a sentence, i.e. sentence-modifying *Adv-mente*, namely conjunctive adverbs (PC), disjunctive adverbs of style (PS), and disjunctive adverbs of attitude (PA) (MOLINIER & LEVRIER, 2000). Focus adverbs (MF), albeit being commonly integrated in the clause, were also filtered out due to their low potential of receiving collocation status, since their sole purpose in an utterance is to emphasise a sentence constituent.

Certain verbs with little semantic content were also filtered out at this stage. So-called *support verbs* (GROSS, 1981), such as *fazer* « to do », *dar* « to give », *ter* « to have », and *haver* « to exist »/« there is/are » were amongst the verbs to be discarded, as well as copula verbs such as *ser* « to be », *estar* « to be », *permanecer* « to remain », *ficar* « to stay », and *parecer* « to seem ». However, if these verbs were part of a verb chain, i.e. a verb phrase forming a compound tense with a past participle form, the modifying relation would be established between the participle, as head of the phrase, and the adverb. Participles used as adjectives were also ignored.

Due to the high frequency with which the verbs just mentioned are used in the language, they were present in the vast majority of verb-adverb combinations extracted from the corpus. After the filtering process, the remaining number of verb-adverb different bigrams was 5,973, which was then considered the set of collocation candidates on which manual annotations would be made.

3. Manual classification of collocation candidates

3.1 Establishing empiric criteria

The criteria empirically devised to establish the collocational status of candidate pairs take into account the relationship of a given term with its possible collocates as well as the meaning of the words involved in the combination. An explanation of these criteria, along with illustrating examples, is presented bellow.

1. The adverb has a hyperbolic meaning in the combination, e.g.:
   (2) *Ele **esperou eternamente** pelo telefonema*

   « He waited eternally for the phone call »

2. The adverb holds a non-literal meaning in the combination, e.g.:

   (3) *O time **venceu confortavelmente** a partida* « The team **won** the match **comfortably** »
   ≠ *O time estava confortável* « The team was comfortable »

   (4) *Ele **deitou-se confortavelmente** na cama* « He **lay comfortably** in bed »
   = *Ele estava confortável* « He was comfortable »

While in (4) the adverb *confortavelmente* « comfortably » holds its literal meaning, connected to the idea of physical comfort, in (3) it assumes a figurative meaning adopted to express the idea that the match was won *effortlessly* or by a large scoring difference. The non-literal meaning in this case attributes a unique character to this combination that could be potentially indicative of its collocational value in Portuguese. In (4), the adverb in the combination, a manner adverb with scope on the action itself and on the subject of the verb, can be paraphrased by its equivalent base adjective operating on the same subject. In (3) this transformation is not possible, which reinforces the non-literal meaning of the adverb in the context of this sentence.

In another example, the adverb modifies the verb by according a quantifying/intensive value to it, such as *perdidamente* « lost-ly », below:

(5) *Ele **apaixonou-se perdidamente** por ela* 'He **fell lost(ly)**[2] in love for her'
≠ Ele estava perdido 'He was lost'

In (5), the adverb *perdidamente* « lost-ly » is derived from the adjective *perdido* « lost », but its intensifying meaning in the combination cannot be (regularly) derived from the meaning of the base adjective (« one who does not know or is unable to find his/her whereabouts »).

3. The combination belongs to the specific vocabulary of a scientific or technical area of expertise, e.g.:

(6) *Ele **respondeu civilmente** pelo crime que cometeu*
« He **responded civically** for the crime he committed »

In (6), the {*V, Adv-mente*} pair is part of the vocabulary commonly used in the domain of law, which accounts for the fixedness of the expression in Portuguese.

4. Synonymic relations between adverbs are broken in the collocational context, e.g.:

(7) Ela chorava **copiosamente** « She cried **copiously** »

(8) ?Ela chorava **abundantemente** « She cried **abundantly** »

Even though the adverbs *copiosamente* « copiously » and *abundantemente* « abundantly », in (7) and (8) respectively, could be considered synonymous, only the adverb *copiosamente* holds a collocational value in this context, since the use of *abundantemente* renders the construction unnatural in Portuguese. We thus say that the synonymic relation between these adverbs is broken. In (8), the adverb would be substituted by *hysterically* or *uncontrollably* in equivalent collocations in English.

5. In a collocation context, the adverb holding collocational status cannot be combined with the antonymous of the verb in question, e.g.:

(9) *O time **venceu** a partida confortavelmente* « The team **won** the match comfortably »

(10) **O time **perdeu** a partida confortavelmente* « The team comfortably **lost** the match »

While the {*V, Adv-mente*} combination in (9) can be considered a collocation in Portuguese, the antonymous of the verb seems to impede a coherent construction in (10), which can be used as an index of the collocational value of the pair in (9). Naturally, this criterion only holds true if an antonymous form of the verb exists in the language. Equally noteworthy is the fact that the simple use of negation does not function as a deciding parameter, as both the collocation status and coherence of the combination would be maintained in this case:

(11) *O time **não** venceu a partida confortavelmente*
« The team **did not** win the match comfortably »

6. The adverb can be combined with often only one subset of the possible meanings of the verb, e.g.:

(12)*A secretária **reproduziu fielmente** os documentos*

« The secretary **reproduced** the documents **faithfully** »

(13)**Coelhos **reproduzem-se fielmente*** « Rabbits **reproduce faithfully** »

[2] This is a case where equivalent collocations in English and Portuguese have adverbs that differ altogether, both morphologically and etymologically. *Perdidamente* « lost-ly » is an adverb for which there is no literal translation in English. There is also no equivalent verb for *apaixonar-se*, so that instead the support verb construction *to be in love*, in the inchoative variant *to fall in love*, was considered here (the inchoative variant was used for consistency with the aspectual value of the Portuguese verb). For this English expression, an adverbial collocation could be *madly*. Interestingly, the base adjective *perdido* also appears in a compound adjectival construction *estar perdido de amores por* « to be lost of/from love-pl for ». A similar construction also exists with the adjectives *doido* and *louco* « mad »: *estar louco/doido de amores por* « to be mad of/from love_pl for ».

While the adverb *fielmente* « faithfully » can be combined with the verb *reproduzir* « to reproduce » in (12), the combination is not possible in (13), as the verb in this sentence, albeit being homonymous to the one in (12), consists in fact of a different lexical item with a different syntactic construction.

Based on these criteria, the set of candidate bigrams were classified as to their collocation status, receiving either the tag of collocation or the tag of non-collocation. Table 3.1 shows the number of candidate bigrams per frequency range in the corpus, and the number of cases among them that have been classified as collocations.

Freq. Range	# Candidates	# Collocations
> 100	65	39
100 - 10	2700	334
5 - 10	3208	128

Table 3.1 Number of candidates and of collocations per frequency

3.2 Assessing native speakers' intuitions

In order to test the intuition of native speakers of Portuguese with regard to the collocational status of the linguistic pattern investigated, a sample classification task was carried out with 21 subjects, of which 13 were native speakers of European Portuguese and 8 of Brazilian Portuguese. The dataset to be classified was composed of 30 collocation candidates randomly selected, 15 having been previously classified as collocations, and 15 as non-collocations. The candidate pairs were presented to the subjects in the contexts where they actually occurred in the corpus, with the {*V, Adv-mente*} pairs highlighted in each sentence. Prior to making a decision on the status of the pairs, the subjects were asked to attentively consider the set of guiding criteria presented in Section 3.1.

Cohen's κ statistic chance-corrected inter-annotator agreement (COHEN, 1960) was calculated for classifications made on the entire set of 30 pairs randomly selected for the experiment. Results are presented in Table 3.2. Results based solely on the 15 pairs that had been previously classified as collocations are presented in Table 3.3.

κ for 30 randomly selected candidates	
Percent of overall agreement	0.57
Fixed-marginal kappa	0.06

Table 3.2 κ for 30 randomly selected pairs of collocation candidates

κ for 15 cases of collocation in the sample	
Percent of overall agreement	0.62
Fixed-marginal kappa	0.10

Table 3.3 κ for 15 pairs among random selection previously classified as collocations

Cohen's κ values can vary from -1.0 to 1.0, where 0 would represent chance agreement. The κ results for the entire set of randomly selected collocation candidates and just for the cases previously classified as collocations were of 0.06 and 0.10 respectively, which can be considered to stand in the range of slight agreement according to the interpretation scale proposed by LANDIS & KOCH (1977).

Even though these results are above what could be considered agreement by chance, they can arguably be deemed low. The most likely reason for this lies in the fact that the sample used in the experiment was too small, requiring an extremely high raw agreement percentage in order for the κ value to reach higher levels of significance. Because of this, κ values achieved in the experiment do not allow for definitive conclusions to be taken with respect to the agreement of the recruited subjects on the collocational status of word pairs. The limited size of the sample was due to the foreseen resistance that a larger set would most likely find among potential voluntary annotators, and to the risk of losing consistency if a larger list of examples had been presented to them.

The poor κ values could also be considered indicative of the elusive nature of the concept of collocation. In this respect, it can be seen in Tables 3.2 and 3.3 that the agreement achieved among cases that had been previously classified as collocations was higher than the overall agreement. This denotes that identifying negative cases poses more difficulty than identifying positive ones, which only confirms that the limit between both is far from being clear-cut. Considering just the positive cases, it can be noticed that a raw agreement of 62% has been reached, which, despite the low κ value, could be considered indicative in some degree of the collocational phenomenon addressed in this paper. One of the factors that denote this phenomenon is the premise that certain lexical combinations, in certain language producing contexts, tend to be given preference over other combinations that can be deemed nearly semantically equivalent. That would be the case of example (7) above, where to express the idea of crying in excess, the Portuguese speaker seems to give preference to the adverb *copiosamente* « copiously » or even *convulsivamente* « convulsively », as opposed to other adverbs of very similar semantic content, such as

abundantemente « abundantly » and *excessivamente* « excessively ». This is one of the reasons to consider the former pairs as collocations, which does not seem to be the case in regard to the latter ones.

4. Correlation of association measures with human classification

A number of statistical association measures have already been tested for capturing the linguistic phenomenon of collocations. PECINA (2010) provides an extensive account in this respect, remarking the particularly good performance of Unigram Subtuples (UnigSub) (PECINA, 2010) and Mutual Information (MI) (FANO, 1961) for large-sized corpora. SERETAN (2011), in turn, mentions the appropriateness of Log-likelihood Ratio (LLR) (DUNNING, 1993) for capturing low-frequency word combinations. In this Section, the manual classification of the collocation candidates will be contrasted with association measures that have received significant attention in previous research. The aim of this comparison is to unveil the measures that are most sensitive to the collocational pattern investigated.

The following measures were chosen for the experiment: *t*-test, Pearson's chi-square (χ^2), Mutual Information (MI), Log-likelihood Ratio (LLR), Dice Coefficient (Dice), and Unigram Subtuples (UnigSub).

The set composed of 5,973 collocation candidates, already classified as to their collocational status, was stratified into three subsets according to the frequency of the bigrams in the corpus. The subsets correspond exactly to the three frequency ranges presented in Table 3.1. The groups were denominated S1, S2, S3, in decreasing order of frequency, respectively.

The *t*-test and χ^2 are both measures that have pre-established statistical significance thresholds for the analysis of results. The performance of these two measures was analysed in terms of precision, recall, and *F*-measure, taking into account a threshold value of 2.576 for the *t*-test, and 3.841 for χ^2, values that correspond to a confidence level of $\alpha = 0.005$ and $\alpha = 0.05$, respectively, which have been previously adopted in similar contexts aimed at identifying collocations (MANNING & SCHÜTZE, 1999 : 153 ; 159). Results of these two measures for S1, S2, and S3 separately, as well as for the set altogether, are shown in Table 4.1

	***t*-test**			**χ^2**		
	Precision	**Recall**	***F*-measure**	**Precision**	**Recall**	***F*-measure**
S1	0.603	0.974	0.745	0.609	1	0.757
S2	0.129	0.937	0.227	0.123	0.964	0.218
S3	0.082	0.460	0.140	0.041	1	0.079
All	0.128	0.818	0.222	0.084	0.976	0.156

Table 4.1. *t*-test and χ^2 results on collocation candidates

Figures in Table 4.1 clearly denote that the *t*-test and χ^2 significance threshold values fell far short of identifying the collocation pattern investigated. The reason behind the poor performance of these measures is most likely connected to the low frequency of the linguistic phenomenon dealt with, a fact that has already been reported in the literature with regard to the *t*-test (DUNNING, 1993). It can also be observed in Table 4.1 that the higher the frequency of the collocation candidates in the corpus, the more satisfactory the performance of the *t*-test and χ^2 is in identifying the phenomenon. The *F*-measure of both tests increases from S3 to S1.

Even though a decision can always be made with regard to a threshold value to be applied to the results of statistical association tests, the other measures calculated for the collocation candidates extracted from the corpus – namely MI, LLR, Dice, and UnigSub – do not have a pre-established threshold for filtering results. The correlation of these measures with the manual classification of collocation candidates was assessed based on the Pearson product-moment correlation coefficient (*r*) (PEARSON, 1896), which measures the linear relationship between two variables – in this case, the referred measures and the classification of bigrams as (non-)collocations. Pearson's *r* values for the aforementioned measures, considering S1, S2, and S3 and the set altogether, are presented in Table 4.2.

	Pearson Correlation Coefficient (*r*)						
	t-test	χ^2	MI	LLR	Dice	UnigSub	# Instances
S1	0.0321	0.2358	*0.4562*	*0.3610*	*0.3831*	*0.3469*	65
S2	0.0759	0.0633	0.2876	0.2403	0.1711	*0.3816*	2700
S3	0.1126	0.0447	*0.3137*	*0.3312*	0.1144	0.1707	3208
All	0.1519	0.0528	*0.3093*	*0.3109*	0.2287	*0.3453*	5,973

Table 4.2. Pearson results for *t*-test, χ^2, MI, LLR, Dice, and UnigSub for considering the classification of collocation candidates

Values for *r* can range from -1.0 to 1.0. According to COHEN (1988), an *r* of .10 could be considered to have a small *effect size* (ES), while an *r* of ± .30 would have a medium ES, and an *r* equal to or above .50 ($r \geq .50$), a large ES. In other words, the furthest the *r* value is from zero, the stronger the relationship between the two variables analysed should be.

In Table 4.2, it can be observed that the four association measures presented a medium ES for S1, the subset with frequent collocation candidates in the corpus. Concerning S3, the *r* value of Dice and UnigSub presented a considerably small ES, which stood at approximately 0.1 for both measures. The small ES of *r* for Dice and UnigSub seems to suggest that these two measures are not appropriate to capture the collocation pattern investigated when it occurs infrequently. LLR, on the other hand, has maintained *r* values from 0.24 to 0.36 across the three subsets. This corroborates findings of previous research that affirm this measure could be deemed reliable for the task of collocation extraction in general (SERETAN, 2011), since it would be sensitive to both high and low-frequency phenomena (DUNNING, 1993 : 62). MI showed a similar trend in this respect, with *r* values ranging from 0.28 to 0.45, where the lowest value corresponds to S2, the subset including pairs of medium frequency in the corpus.

Considering the entire set of collocation candidates, UnigSub, LLR, and MI were, in descending order, the measures that presented the highest correlation with the human classification on the collocation status of the pairs. The *t* and χ^2 tests presented a notably low correlation with the classification, which seems to confirm the poor Precision, Recall and *F*-measure results of these two measures, as shown in Table 4.1.

5. Conclusion and future work

In this paper, we described a syntax-based approach to collocation extraction from corpora, and assessed the performance of traditional association measures in identifying the collocation pattern {*V, Adv-mente*} in Portuguese, based on a reference manual classification carried out in accordance with a set of empirically devised linguistic criteria.

Having information on collocational patterns is of high importance to fields such as NLP and Second Language Learning, since conforming to these patterns is of interest both to humans and to computer systems aimed at processing natural language. Pairs such as *apaixonar-se perdidamente* « fall in love madly » for example, due to their morphological/etymological difference between English and Portuguese, might pose a challenge to Machine Translation (MT) engines for instance, which at a number of times fail to provide correct translations for the linguistic pattern here addressed, as it has been shown in VIEIRA (2012).

The identification of collocations conforming to this pattern has proven to be challenging both for statistical association measures and for native speakers of the language. The relatively low frequency of the phenomenon seems to contribute to that difficulty, especially in regard to the statistical classification. Pre-established threshold values proposed for the *t* and χ^2 tests have shown to be particularly unsatisfactory in that respect. The correlation of other measures with the reference classification was considerably more promising.

As future work, we intend to use results of these measures as training data to build an automatic collocation classifier for the {*V, Adv-mente*} pattern in Portuguese using machine learning techniques. This is a strategy that would arguably profit more significantly from the results of these measures, as it disregards any decision based on critical values, but rather takes all results into account as being potentially useful for the classification task. As further assessments to be carried with native speakers, we intend to repeat the experiment previously described making use of a different questioning strategy, asking respondents to rate candidate pairs of similar meaning as opposed to deciding on the collocation status of a single pair.

References

AÏT-MOKHTAR Salah ; CHANOD Jean-Pierre ; ROUX Claude (2002), Robustness Beyond Shallowness: Incremental Deep Parsing. *Natural Language Engineering*, 8. New York: Cambridge University Press, p. 121–144.

BECHARA Evanildo (2003), *Moderna gramática portuguesa*. 37 ed. Rio de Janeiro: Lucerna.

CHOUEKA Yaacov (1988), Looking for Needles in a Haystack or Locating Interesting Collocational Expressions in a Large Corpus. In: *Proceedings of RIAO '88*, p. 609–623.

COHEN Jacob (1960), A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20, p. 37-46.

COHEN Jacob (1988), *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

DUNNING Ted (1993), Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics* 19(1), p. 61-74.

FANO Robert (1961), Transmission of Information: A Statistical Theory of Communications. MIT Press, Cambridge, MA.

FERNANDES Gaia (2011), *Classification and Word Sense Disambiguation: The case of -mente Ending Adverbs in Brazilian Portuguese.* M.A. Thesis, Universidade do Algarve/ Universitat Autònoma de Barcelona.

FIRTH John Rupert (1957), A Synopsis of Linguistic Theory 1930-55. In: Firth, J. R. et al. *Studies in Linguistic Analysis*. Special volume of the Philological Society. Oxford: Blackwell.

GROSS Maurice (1981), Les bases empiriques de la notion de prédicat sémantique. *Langages* 63, p. 7–52.

LANDIS J Richard ; KOCH Gary G (1977), The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, p. 159-174.

MAMEDE Nuno ; BAPTISTA Jorge ; DINIZ Cláudio ; CABARRÃO Vera. (2012), STRING: A Hybrid Statistical and Rule-Based Natural Language Processing Chain for Portuguese. Demo PROPOR 2012. Available at <http://www.propor2012.org/demos/DemoSTRING.pdf> [Accessed 14 May 2012].

MANNING Christopher ; SCHÜTZE Hinrich (1999), *Foundations of Statistical Natural Language Processing.* Cambridge, MA: MIT Press.

MEĽČUK Igor (2003), Collocations, definition, rôle et utilité. In: GROSSMANN Francis ; TUTIN Agnès (eds.), *Les collocations, analyse et traitement,* Amsterdam: De Werelt, p. 23-31.

MOLINIER Christian ; LEVRIER Françoise (2000), *Grammaire des adverbes. Description des formes en* -ment. Genève: Droz.

PEARSON Karl (1896), Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity and Panmixia. *Philosophical Transactions of the Royal Society of London* 187, p. 253-318.

PECINA Pavel (2010), Lexical Association Measures and Collocation Extraction. *Language Resources and Evaluation* 44(1), p. 137-158.

SANTOS Diana ; ROCHA Paulo (2001), Evaluating CETEMPúblico, a free resource for Portuguese. In: *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, ACL'2001*, Toulouse, 9-11 July 2001, p. 442-449.

SERETAN Violeta (2011), *Syntax-Based Collocation Extraction. Text, Speech and Language Technology*. Dordrecht: Springer.

VIEIRA Lucas Nunes (2012), PT-EN Collocation Equivalents and Machine Translation Evaluation. *BULAG Natural Language Processing and Human Language Technology*, Forthcoming.