The annotation guidelines of the
Latin Dependency Treebank and *Index Thomisticus* Treebank
The treatment of some specific syntactic constructions in Latin

David Bamman[1], Marco Passarotti[2], Roberto Busa[2], Gregory Crane[1]

[1]Tufts University
The Perseus Project, Medford, MA, USA
E-mail: {david.bamman, gregory.crane}@tufts.edu

[2]Catholic University of the Sacred Heart
Largo Gemelli 1, 20123 Milan, Italy
E-mail: {roberto.busa, marco.passarotti}@unicatt.it

Abstract

The paper describes the treatment of some specific syntactic constructions in two treebanks of Latin according to a common set of annotation guidelines. Both projects work within the theoretical framework of Dependency Grammar, which has been demonstrated to be an especially appropriate framework for the representation of languages with a moderately free word order, where the linear order of constituents is broken up with elements of other constituents. The two projects are the first of their kind for Latin, so no prior established guidelines for syntactic annotation are available to rely on. The general model for the adopted style of representation is that used by the Prague Dependency Treebank, with departures arising from the Latin grammar of Pinkster, specifically in the traditional grammatical categories of the ablative absolute, the accusative + infinitive, and gerunds/gerundives. Sharing common annotation guidelines allows us to compare the datasets of the two treebanks for tasks such as mutually checking annotation consistency, diachronically studying specific syntactic constructions, and training statistical dependency parsers.

1. The Latin Dependency Treebank and *Index Thomisticus* Treebank

Treebanks have recently emerged as a valuable resource not only for computational tasks such as grammar induction and automatic parsing, but for traditional linguistic and philological pursuits as well. This trend has been encouraged by the creation of several historical treebanks, such as that for Middle English (Kroch & Taylor, 2000), Early Modern English (Kroch et al., 2004), Old English (Taylor et al., 2003), Early New High German (Demske et al., 2004) and Medieval Portuguese (Rocio et al., 2000).

The Perseus Project (Crane et al., 2001) and the *Index Thomisticus* (IT) (Busa 1974-1980) are currently in the process of developing treebanks for Latin – the Latin Dependency Treebank (LDT) (Bamman & Crane, 2006; Bamman & Crane, 2007) on works from the Classical era, and the *Index Thomisticus* Treebank (IT-TB) (Passarotti, 2007) on the works of Thomas Aquinas[1]. In order for our separate endeavors to be most useful for the community, we have come to an agreement on a common standard for the syntactic annotation of Latin.

In this paper we present some examples from our preliminary set of annotation guidelines that illustrate how

we have adapted our general annotation model inherited from the one of the Prague Dependency Treebank (PDT) of Czech to the specific linguistic demands of Latin.

Date	Author	Words	Sentences
1st c. BCE	Cicero	5,663	295
1st c. BCE	Caesar	1,488	71
1st c. BCE	Sallust	12,311	701
1st c. BCE	Vergil	2,613	178
4th-5th c. CE	Jerome	8,382	405
	Total	**30,457**	**1,650**

Table 1: Latin Dependency Treebank composition

Date	Author	Words	Sentences
13th c. CE	Aquinas	30,145	1,352
	Total	**30,145**	**1,352**

Table 2: IT-Treebank composition

Tables 1 and 2 present the composition of both of our treebanks. Both projects work within the theoretical framework of Dependency Grammar (DG), which differs from constituent-based grammars by foregoing non-terminal phrasal categories and instead linking words themselves to their immediate head (Tesnière, 1959; Mel'cuk, 1988). This is an especially appropriate manner of representation for languages with a moderately free word order (such as Latin and Czech), where the linear order of constituents is broken up with elements of other constituents.

A DG representation of *ista meam norit gloria canitiem*,

---

[1] The IT-TB is available online at the following URL: http://gircse.marginalia.it/~passarotti; the LDT can be found at http://nlp.perseus.tufts.edu/syntax/treebank.

for instance, is provided in figure 1 (arcs are directed from heads to their dependents).
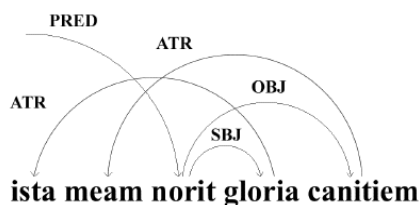

Figure 1: *ista meam norit gloria canitiem*
(Prop., *Carm*., I.8.46)
("that glory would know my old age")

DG is also appropriate for Latin since it is not too theoretically distant from Classical pedagogical grammars, where the highly inflected nature of the language leads to discussions of, for example, which adjective "modifies" which noun in a sentence. A DG simply assigns one such "modification" to every word.

2. Annotation Guidelines

The development of more than one treebank for any given language has the potential to lead to balkanization, with each individual project working independently and pursuing its own research agenda. This can lead to a proliferation of annotation styles and datasets that are ultimately incompatible. The adoption of common structural standards such as XCES (Ide et al., 2000) and infrastructures mitigates this to a certain extent, but true dataset compatibility also extends to the level of the individual syntactic decisions themselves. While such compatibility is not always possible, the benefits of working together are significant.

Our two projects are the first of their kind for Latin, so we do not have prior established guidelines to rely on for syntactic annotation. As noted above, the general model for our style of representation is that used by the PDT (Hajič et al., 1999). Adopting an annotation style wholesale, however, is easier said than done. Since nearly all Latin available to us is highly stylized, we are constantly confronted with idiosyncratic constructions that could be syntactically annotated in several different ways. These constructions (such as the ablative absolute or the passive periphrastic) are common to Latin of all eras. Rather than have each project decide upon and record each decision for annotating them, we decided to pool our resources and create a single annotation manual that would govern both treebanks (Bamman et al., 2007a; Bamman et al., 2007b).

Since we deal with Latin dialects separated by 13 centuries, sharing a single annotation manual is very useful for comparison purposes, such as checking annotation consistency or diachronically studying specific syntactic constructions. In addition, the task of data annotation through these common guidelines allows us to base the decisions on a variety of examples from a wider range of texts and combine our datasets in order to train statistical dependency parsers.

2.1 Tagset

Table 3 lists all of the tags currently in use.

PRED	predicate
SBJ	subject
OBJ	object
ATR	attributive
ADV	adverbial
ATV/AtvV	complement
PNOM	predicate nominal
OCOMP	object complement
COORD	coordinator
APOS	apposing element
AuxP	preposition
AuxC	conjunction
AuxR	reflexive passive
AuxV	auxiliary verb
AuxX	commas
AuxG	bracketing punctuation
AuxK	terminal punctuation
AuxY	sentence adverbials
AuxZ	emphasizing particles
ExD	ellipsis

Table 3: Complete Latin tagset

All of the tags can also be appended with a suffix in the event that the given node is member of a coordinated construction (_Co), an apposition (_Ap) or a parenthetical statement (_Pa).

The tag PRED is given to the predicate of the main clause (or clauses, in case of coordination or apposition) of a sentence; the head verbs of the subordinate clauses are annotated according to the clause role in the sentence (for instance, a declarative clause acting as subject is annotated with the tag SBJ).

An ATR is a sentence member that further specifies a noun in some respect; typical attributives are adjectives (*bonus puer*: "good boy") and nouns in the genitive case (*domus patris*: "the father's house").

The difference between OBJ and ADV roughly corresponds to the one between arguments (inner participants) and adjuncts of verbs or adjectives, i.e., between those called 'actants' and 'circonstants' in the terms of Tesnière (1959). A special kind of OBJ is the determining complement of the object, which is tagged with OCOMP, such as *senatorem* in a sentence like *aliquem senatorem facere* ("to nominate someone senator"). The determining complement of the subject is, conversely, tagged using PNOM; this mainly occurs in case of constructions like *aliquis senator fit* ("someone becomes senator").

The tag OCOMP covers some of the functions of the ATV/AtvV tag (Verbal Attribute) as used by the PDT: departing from PDT style, we assign a different tag to object complements (OCOMP) and to complements that are not direct arguments of the verb (ATV/AtvV). These

are usually noun phrases and adjectives that agree with their head noun morphologically, but differ from typical attributes in that they also qualify the function of the verb. The use of ATV/AtvV is largely similar to the account of 'praedicativa' given in Pinkster (1990, pp. 142-162) and can be simplified to the following two examples contained therein (p. 142):

• *Galli laeti in castra pergunt*
• *Cicero consul coniurationem Catilinae detexit*

In the first example, an attributive reading of *laeti* would lead to the translation "the happy Gauls enter the camp". As an ATV, it would be rendered "the Gauls happily enter the camp": while *laeti* agrees morphologically with the subject *Galli*, it simultaneously specifies the nature of the predicate. Since it is an inflected adjective (and not the adverb *laete*), it still bears a syntactic relationship to the noun phrase and should therefore depend on it (and not simply on the verb via ADV). This results in the following tree:
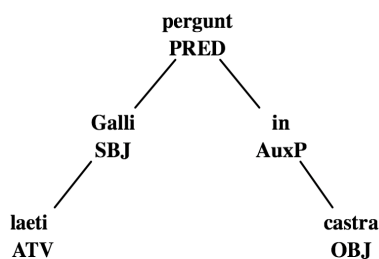
Figure 2: *Galli laeti in castra pergunt*
("the Gauls happily enter the camp")

If the head noun phrase in such constructions is implied rather than explicit, the praedicativum depends on the main verb via AtvV as in figure 3 (if *laeti* here were a SBJ depending on *pergunt*, the sentence would mean "the happy ones enter the camp").
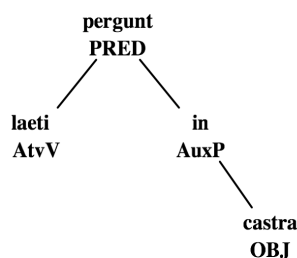

Figure 3: *laeti in castra pergunt*
("they happily enter the camp")

In Pinkster's second example from above, *consul* is not a simple attribute (or appositive) of *Cicero* since it qualifies the nature of the verb: "Cicero uncovered Catiline's conspiracy as consul (i.e., when he was consul)". Since *consul* agrees with *Cicero* morphologically while also modifying the main predicate, we annotate it as depending on the noun via ATV (figure 4).
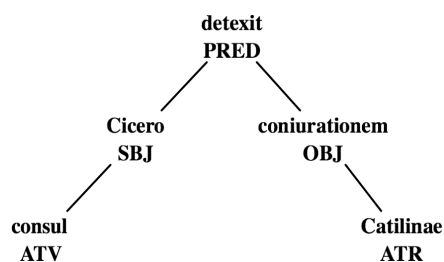

Figure 4: *Cicero consul coniurationem Catilinae detexit*
("Cicero uncovered Catiline's conspiracy as consul")

While we both adhere to these common standards in all other respects, we do differ in the annotation of a single construction: ellipsis. Since its inception, the LDT has annotated ellipsis in a manner that attempts to preserve the structure of the underlying sentence with a complex syntactic tag, while the IT-TB has followed the PDT convention of attaching an orphan to its head with the relation ExD. This difference can be seen in the differing annotations provided in figures 5 and 6.
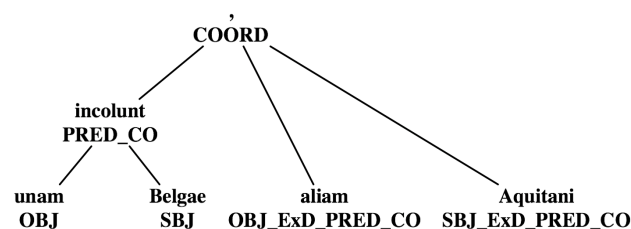

Figure 5: LDT annotation of
unam incolunt Belgae, aliam Aquitani (Caes., *B.G.*, 1.1)
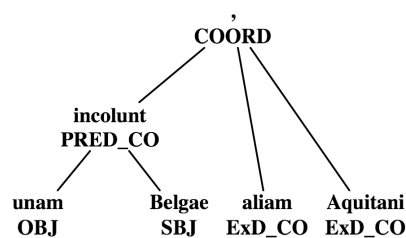("one the Belgae inhabit, another the Aquitani")

Figure 6: IT-TB annotation of
unam incolunt Belgae, aliam Aquitani (Caes., *B.G.*, 1.1)

While the edge labels we assign to these orphans are different, the structure of the tree is not, and our data is still compatible since the formalism used by the LDT can always be reduced to that used by the IT-TB.

2.2 The treatment of some specific constructions

The following sections are devoted to the description of the treatment of several specific constructions in Latin: the ablative absolute, the accusative + infinitive, and gerunds/gerundives.

2.2.1 The ablative absolute

The ablative absolute is a grammatical construction similar to the English nominative absolute, where a noun and (typically) a participle form a phrase that is disjoint from the grammar of the rest of the sentence; in Latin both the noun and participle are inflected in the ablative case.

Following Pinkster (1990), we treat ablative absolutes as an embedded predication that functions as an adjunct. In common absolutes (with a noun + participle), the noun is annotated as the subject of the participle, with the participle (as the head of the ablative absolute phrase) dependent on the main verb as an adverbial. Figure 7 provides one such example.
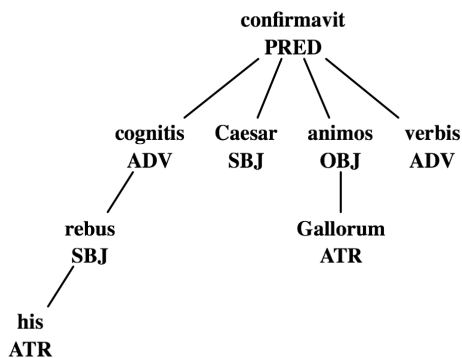
Figure 7: *his rebus cognitis Caesar Gallorum animos verbis confirmavit*
(Caes., *B.G.*, 1.33)
("these things known, Caesar calmed the minds of the Gauls with words")

In absolutes involving no participle (as in figure 8), the head noun is dependent on the main verb via ADV, with its child (the element the head is "functioning as") dependent on it via ATV.
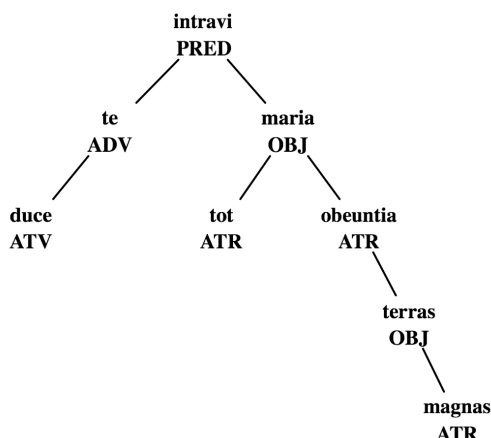
Figure 8: *magnas obeuntia terras tot maria intravi duce te*
(Verg., *Aen.*, 6.58)
("I have entered so many seas breaking upon great lands with you as my guide")

2.2.2 Accusative + infinitive

In accusative + infinitive constructions (most commonly found in indirect discourse), the infinitive verb is the head of its phrase. This verb represents the entire clause and depends, usually via OBJ, on the word that introduces the discourse. Within the phrase, standard annotation applies (so that the subject, while accusative, still depends on the indirect infinitive via SBJ).

Thus, a sentence such as *dicit deum apparuisse in corporalibus formis* is annotated in the following way:
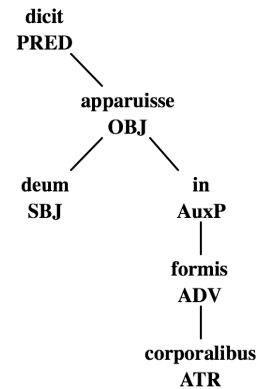
Figure 9: *dicit deum apparuisse in corporalibus formis*
(Thomas, *Super Sententiis Petri Lombardi*, II, Dist. 8, Qu. 1, Prologus, 14-1, 14-6)
("it says that God had appeared in bodily forms")

2.2.3 Gerunds and gerundives

As a verbal noun, gerunds are relatively straightforward to annotate: they are simply treated as nouns and annotated according to their syntactic function in the sentence.
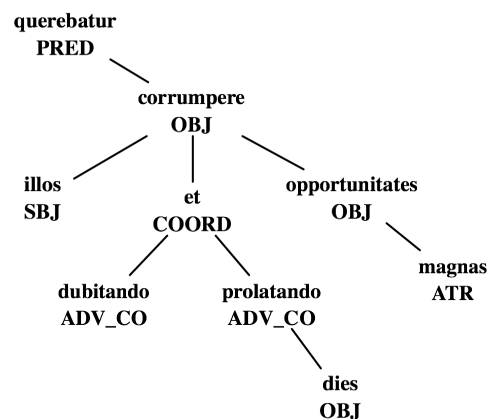
Figure 10: *querebatur ... illos dubitando et dies prolatando magnas opportunitates corrumpere*
(Sal., *Cat.*, 43)
("he complained that they wasted great opportunities by doubting and delaying")

Gerundives, on the other hand, behave more like participles in that they can function either as an attribute or in a dominating construction. When attributive, gerundives are labelled ATR; when dominating, they are annotated according to their specific role in the sentence. A test for which tag is appropriate is whether or not the gerundive

can be omitted: if it can be left out of the sentence without changing the lexical meaning of the predicate, it is annotated via ATR; if not, then it is dominating. In the example provided in figure 11, *effeminandos* cannot be left out of the sentence since *quae ad animos pertinent* ("which pertain to the minds") isn't able to stand on its own.
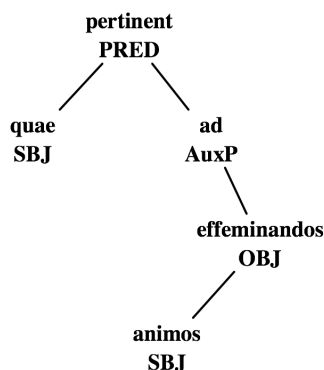

Figure 11: *quae ad effeminandos animos pertinent*
(Caes., *B.G.*, 1.1.3)
("which pertain to the mind being effeminated")

Our intuition here may be to treat the noun *animos* as the direct object of the gerundive (since we idiomatically translate the phrase with such a sense: "which pertain to effeminating the mind"), but we should keep in the mind that a gerundive is a passive form, which then makes *animos* a subject.

An attributive use of a gerundive can be seen in the fragment *privatio formae inducendae* ("the privation of the form to be inserted") in figure 12. Here *inducendae* is omissible and is therefore labelled with ATR.
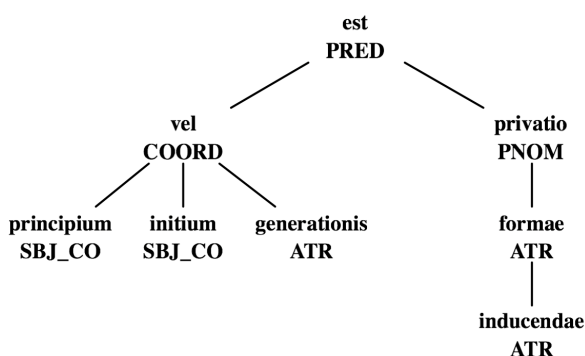

Figure 12: *principium vel initium generationis est privatio formae inducendae*
(Thomas, *Super Sententiis Petri Lombardi*, I, Dist. 5, Qu. 3, Art. 1, Solutio, 7-4, 8-4)
("the beginning or origin of generation is the privation of the form to be inserted")

When a gerundive appears in a passive periphrastic construction, it is treated as a predicate nominal, as shown in figure 13.
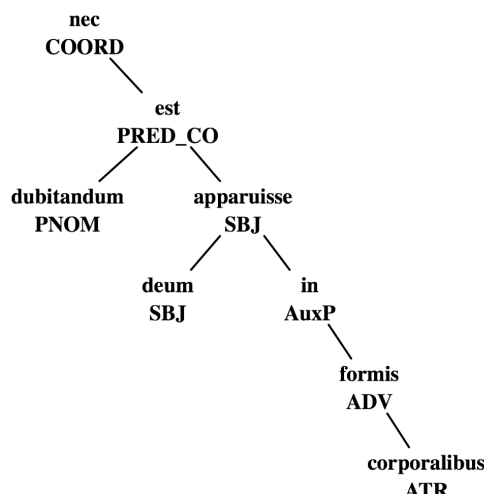

Figure 13: *nec dubitandum est deum in corporalibus formis apparuisse*
(Thomas, *Super Sententiis Petri Lombardi*, II, Dist. 8, Qu. 1, Prologus, 12-2, 13-2)
("and that God has appeared in bodily forms should not be doubted")

3.   Conclusion

The examples above all reflect a mutual effort by our two independent projects at adopting a common set of annotation guidelines. While our overall annotation style is based on that used by the PDT, each of these examples illustrates a way in which those general guidelines (developed for Czech) have been extended and refined for use in Latin.

4.   Acknowledgements

5.   References

Bamman, D. & Crane, G. (2006). The design and use of a Latin dependency treebank. In *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT 2006)*. Prague: ÚFAL MFF UK, pp. 67-78.

Bamman, D. & Crane, G. (2007). The Latin Dependency Treebank in a cultural heritage digital library. In *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007)*. Prague: Association for Computational Linguistics, pp. 33-40.

Bamman, D., Crane, G., Passarotti, M. & Raynaud, S. (2007a). *Guidelines for the Syntactic Annotation of Latin Treebanks*. Technical report. Tufts Digital Library: http://nlp.perseus.tufts.edu/syntax/treebank/1.3/docs/guidelines.pdf.

Bamman, D., Crane, G., Passarotti, M. & Raynaud, S. (2007b). A Collaborative Model of Treebank Development. In *Proceedings of the Sixth Workshop on Treebanks and Linguistic Theories (TLT 2007)*. Bergen:

Northern European Association for Language Technology (NEALT) Proceedings Series, Vol. 1, pp. 1-6.

Busa, R. (1974-1980). *Index Thomisticus: sancti Thomae Aquinatis operum omnium indices et concordantiae, in quibus verborum omnium et singulorum formae et lemmata cum suis frequentiis et contextibus variis modis referuntur quaeque / consociata plurium opera atque electronico IBM automato usus digessit Robertus Busa SJ*. Stuttgart - Bad Cannstatt: Frommann - Holzboog.

Crane, G., Chavez, R.F., Mahoney, A., Milbank, T.L., Rydberg-Cox, J.A., Smith, D.A. & Wulfman, C.E. (2001). Drudgery and deep thought: Designing digital libraries for the humanities. *Communications of the ACM*, 44(5), pp. 34-40.

Demske, U., Frank, N., Laufer, S. & Stiemer, H. (2004). Syntactic interpretation of an Early New High German corpus. In *Proceedings of the Third Workshop on Treebanks and Linguistic Theories (TLT 2004)*. Tübingen, pp. 175-182.

Hajič, J., Panevová, J., Buráňová, E., Urešová, Z. & Bémová, A. (1999). *Annotations at analytical level: Instructions for annotators* (English translation by Z. Kirschner). Technical report. Prague: UFAL MFF UK.

Ide, N., Bonhomme, P. & Romary, L. (2000). XCES: An XML-based encoding standard for linguistic corpora. In *Proceedings of the second international Conference on Language Resources and Evaluation (LREC 2000)*. Athens, pp. 825-830.

Kroch, A., Santorini, B. & Delfs, L. (2004). *The Penn-Helsinki Parsed Corpus of Early Modern English*. University of Pennsylvania, Philadelphia: Department of Linguistics.

Kroch, A. & Taylor, A. (2000). *The Penn-Helsinki Parsed Corpus of Middle English, second edition*. University of Pennsylvania, Philadelphia: Department of Linguistics.

Mel'cuk, I. (1988). *Dependency Syntax: Theory and Practice*. New York: State University of New York Press.

Passarotti, M. (2007). Verso il Lessico Tomistico Biculturale. La *treebank* dell'*Index Thomisticus*. In R. Petrilli, D. Femia (Eds.), *Il filo del discorso. Intrecci testuali, articolazioni linguistiche, composizioni logiche. Atti del XIII Congresso Nazionale del la Società di Filosofia del Linguaggio, Viterbo, 14-16 Settembre 2006*. Roma: Aracne Editrice, Pubblicazioni della Società di Filosofia del Linguaggio 04, pp. 187-205.

Pinkster, H. (1990). *Latin Syntax and Semantics*. London: Routledge.

Rocio, V., Alves, M.A., Lopes, J.G., Xavier, M.F. & Vicente, G. (2000). Automated Creation of a Medieval Portuguese Partial Treebank. In A. Abeillé (Ed.), *Treebanks: Building and Using Parsed Corpora*. Dordrecht: Kluwer Academic Publishers, pp. 211-227.

Taylor, A., Warner, A., Pintzuk, S. & Beths, F. (2003). *York-Toronto-Helsinki Parsed Corpus of Old English Prose*. University of York: Department of Language and Linguistic Science.

Tesnière, L. (1959). *Éléments de syntaxe structurale*. Paris: Editions Klincksieck.