

Bootstrapping Information from Corpora in a Cross-Linguistic Perspective

edited by

MASSIMO MONEGLIA
ALESSANDRO PANUNZI

RAAMMNNNOONNEEEER

REEEENNOONNMMVAU

per noi produttori di poesia
di bellezza, di arte, la poetica
è considerata cosa anormale, marginale,
scuriosa e sfacciatamente modesta. E ora si
finirà con il riconoscimento del
l'artista dopo la morte.



STRUMENTI
PER LA DIDATTICA E LA RICERCA

– 96 –

UNIVERSITÀ DEGLI STUDI DI FIRENZE
DIPARTIMENTO DI ITALIANISTICA

BIBLIOTECA DIGITALE

COMITATO SCIENTIFICO

Adele Dei
Anna Dolfi
Simone Magherini
Massimo Moneglia

Volumi pubblicati:

MODERNA [diretta da Anna Dolfi]

1. *Giuseppe Dessì. Storia e catalogo di un archivio*, a cura di Agnese Landini, 2002.
2. *Le corrispondenze familiari nell'archivio Dessì*, a cura di Chiara Andrei, 2003.
3. Nives Trentini, *Lettere dalla Spagna. Sugli epistolari a Oreste Macrì*, 2004.
4. *Lettere a Ruggero Jacobbi. Regesto di un fondo inedito con un'appendice di lettere*, a cura di Francesca Bartolini, 2006.
5. «L'Approdo». *Copioni, lettere, indici*, a cura di Michela Baldini, Teresa Spignoli e del GRAP, sotto la direzione di Anna Dolfi, 2007 (CD-Rom allegato con gli indici della rivista e la schedatura completa di copioni e lettere).
6. Anna Dolfi, *Percorsi di macritica*, 2007 (CD-Rom allegato con il *Catalogo della Biblioteca di Oreste Macrì*).
7. *Ruggero Jacobbi alla radio*, a cura di Eleonora Pancani, 2007.
8. *Ruggero Jacobbi, Prose e racconti. Inediti e rari*, a cura di Silvia Fantacci, 2007.
9. Luciano Curreri, *La consegna dei testimoni tra letteratura e critica. A partire da Nerval, Valéry, Foscolo, d'Annunzio*, 2009.
10. *Ruggero Jacobbi, Faulkner ed Hemingway. Due nobel americani*, a cura di Nicola Turi, 2009.
11. Sandro Piazzesi, *Girolamo Borsieri. Un colto poligrafo del Seicento. Con un inedito «Il Salterio Affetti Spirituali»*, 2009.
12. *A Giuseppe Dessì. Lettere di amici e lettori. Con un'appendice di lettere inedite*, a cura di Francesca Nencioni, 2009.
13. *Giuseppe Dessì, Diari 1949-1951*, a cura di Franca Linari, 2009.

LINGUISTICA [diretta da Massimo Moneglia]

1. *Bootstrapping Information from Corpora in a Cross-Linguistic Perspective*, edited by Massimo Moneglia and Alessandro Panunzi, 2010.
2. Alessandro Panunzi, *La variazione semantica del verbo essere nell'italiano parlato*, 2010.

INFORMATICA E LETTERATURA [diretta da Simone Magherini]

1. *BIL Bibliografia Informatizzata Leopardiana 1815-1999. Manuale d'uso vers. 1.0*, a cura di Simone Magherini, 2003.

Bootstrapping Information from Corpora in a Cross-Linguistic Perspective

edited by

Massimo Moneglia
Alessandro Panunzi

Firenze University Press
2010

Bootstrapping Information from Corpora in a Cross-Linguistic Perspective / edited by Massimo Moneglia and Alessandro Panunzi. – Firenze : Firenze University Press, 2010.

(Strumenti per la didattica e la ricerca ; 96)

<http://digital.casalini.it/9788884535290>

ISBN 978-88-8453-518-4 (print)

ISBN 978-88-8453-529-0 (online)

Immagine di copertina: Alessandro Geri Rustighi

Progetto grafico di Alberto Pizarro Fernández

© 2010 Firenze University Press
Università degli Studi di Firenze
Firenze University Press
Borgo Albizi, 28, 50122 Firenze, Italy
<http://www.fupress.com/>

Printed in Italy

INDEX

INTRODUCTION	VII
LINGUISTIC STYLES ENABLED BY THE TECHNOLOGY OF LITERACY <i>Douglas Biber</i>	1
INFORMATIONAL PATTERNING THEORY AND THE CORPUS-BASED DESCRIPTION OF SPOKEN LANGUAGE. THE COMPOSITIONALITY ISSUE IN THE TOPIC-COMMENT PATTERN <i>Emanuela Cresti, Massimo Moneglia</i>	13
LANGUAGE-TEXT INTERFACE: THE EXAMPLE OF THEMATIC PROGRESSION <i>Angela Ferrari, Anna Maria De Cesare</i>	47
INTERACTION ENTRE LA SYNTAXE DES LEXÈMES ET LE SÉMANTISME DES PARTIES DU DISCOURS: NOM VS. ADJECTIF DE COULEUR EN JAPONAIS <i>Itsuko Fujimura</i>	73
THE SEMANTIC VARIATION OF VERB <i>ESSERE</i> IN ITALIAN. THEORETICAL CONSEQUENCES OF CORPUS-BASED STUDIES <i>Alessandro Panunzi</i>	97
CHIEDE. A SPONTANEOUS CHILD LANGUAGE CORPUS OF SPANISH <i>Marta Garrote Salazar, Antonio Moreno Sandoval</i>	121
FROM TEXT TO LEXICON: THE ANNOTATION OF PRE-TARGET STRUCTURES IN AN ITALIAN LEARNER CORPUS <i>Giuseppina Turco, Miriam Voghera</i>	141
PRE-PROCESSING NORMALIZATION PROCEDURES FOR NEWSGROUP CORPORA <i>Manuel Barbera, Simona Colombo</i>	175
THE C-ORAL-BRASIL CORPUS <i>Tommaso Raso, Heliana Mello</i>	193

INTRODUCTION

Corpus Linguistics has been developing since the early 1990s as a consequence of the onset of information technologies that allow processing of huge amounts of linguistic data and it is presently one of the leading paradigms for the theoretical and practical study of language. Spoken and written corpora are now an essential source of information for many scientific domains such as, besides Linguistics and Grammar, Speech Recognition, Language Teaching, Lexicography, Artificial Intelligence, Language Pathology etc..

Corpus Linguistics methodologies have been applied to many languages, and in this frame Romance languages are now well represented by reference corpora, domain specific corpora, speech corpora, annotated corpora, language acquisition corpora, etc. The study of these resources has already produced an important literature that strongly modified the traditional background knowledge for the description and explanation of linguistic phenomena.

In this frame the Linguistics Laboratory of the Italian Department of the University of Florence (LABLITA) is specifically concerned with the constitution of Spoken Romance Corpora that are explored for the study of many fields in the domain of speech communication, such as prosody, information structure, pragmatics, semantics, lexicon and so on. However, besides our work at LABLITA, we think that the achievements of Romance languages corpus-driven studies deserve more attention from the scientific community at world level for both their amount and quality. To this end, within the Internalization program of the University of Florence, LABLITA invites yearly one of the leading scholars in the field of corpus linguistics for a workshop in which current corpus linguistics trends face data and theories derived from Romance corpora analysis. After Claire Blanche Benveniste (École Pratique des Hautes Études), John Sinclair (TWC) in 2006 and Shlomo Izre'el (Tel Aviv University) in 2007, Douglas Biber (Northern Arizona University) visited the lab in 2008.

This book hosts papers given at the 3rd International LABLITA Workshop in Corpus Linguistics, (Italian Department, University of Florence, June 4th - 5th 2008) and it aims at integrating new ideas and results derived from Romance languages corpora in the framework of the overall achievements of corpus linguistics.

The perspective of the contributions is mainly theoretical. Corpus linguistics faces the tradition of formal grammar highlighting the importance of observation adequacy and focuses on the exploration of language performance positive data. Corpora representing the actual contexts in which natural languages are used allows for the bootstrapping of properties that cannot be deduced from competence based judgments.

The Longman Grammar of Spoken and Written English¹, is still the main reference book for what regards corpus based grammars. Biber's paper, which open this volume, sketches the main achievements of the "variational approach" which has been the main method for bootstrapping the differential syntactic properties of written and spoken varieties from English corpora that have been recorded in the Longman Grammar and in subsequent works.

The following contributions, presented to Biber by notable European and extra-European linguists, range over Italian, Spanish, French, Brazilian Portuguese and report the results of long-term corpus driven research. They focus on the various linguistic levels concerned with information extraction and the methodologies applied to this end, and report new large corpus collection initiatives in the Romance language area.

Most papers deal with spoken corpora. They regard the interfaces between syntax, prosody and information structure (Cresti & Moneglia, University of Florence; Ferrari & De Cesare, University of Basel), the bootstrapping of semantic information at the lexical level (Panunzi, University of Florence; Fugimura, Nagoya University), and the comparison of speech performance in child and adult spoken corpora (Garrote & Moreno, Autonomous University of Madrid).

Raso and Mello (Federal University of Minas Gerais) present a new spontaneous speech corpus of Brazilian Portuguese. This new corpus, collected along the lines of the C-ORAL-ROM multilingual corpus², will properly allow the comparison of Brazilian and European Portuguese, which diverge mainly in the oral variety. The corpus will be therefore an essential source of data for linguistic studies.

In order to compute information in language corpora one crucial requirement deals with formats and annotation of raw language data. The contributions by Barbera & Colombo (University of Turin) and Turco & Voghera (Max Planck Institute, University of Salerno) fall in this domain.

Many projects in the last decade have been devoted to the collection of learner corpora. Syntactic development of learners is among the main information that can be derived from the huge amount of available data. To this end Turco & Voghera present an annotation system of L2 texts that has the same basic structure of the one

¹ Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finnegan. 1999. *Longman grammar of spoken and written English*. London: Longman.

² Cresti, E. and M. Moneglia (eds). 2005. *C-ORAL-ROM. Integrated reference corpora for spoken romance languages*, DVD + vol. Amsterdam: Benjamins.

we use to annotate native languages; this allowed a straightforward comparison between non-native and native production.

Barbera & Colombo present the procedures adopted to set up and to compute a large Multilingual Corpus of news-groups crawled from the internet. This corpus freely accessible on the net constitutes one of the most relevant textual resources now available for cross-linguistic comparison in specific semantic domains.

From a more theoretical point of view this book presents new ways to explore spoken corpora for the study of language structure. Corpus driven studies of spoken language afford a more realistic representation of lexical information and allow for the discovery of semantic properties that cannot be observed otherwise. This is the case for the semantic variation of the verb *essere* [to be] in Italian (Panunzi) and for the variation of *color adjectives* in Japanese (Fujimura).

The corpus based study of information structure bore important achievements that are now verified in Romance languages corpora. The information structure of spoken language is systematically conveyed by prosodic cues and it determines severe constraints on syntactic boundaries and semantic compositionality (Cresti & Moneglia). In the absence of prosody, information structure is also reflected in written texts and characterizes its syntax (Ferrari & De Cesare).

Massimo Moneglia
Alessandro Panunzi

LINGUISTIC STYLES ENABLED BY THE TECHNOLOGY OF LITERACY

Douglas Biber

Northern Arizona University

1. Introduction

Over the past several decades, researchers in linguistics, anthropology, and communication studies have been interested in comparisons of the spoken and written modes. However, there has been considerable disagreement on the extent to which the two modes differ linguistically. Research studies in the 1970's and early 1980's usually argued that there are fundamental linguistic differences between speech and writing. For example, researchers such as O'Donnell (1974), Olson (1977), and Chafe (1982) argued that written language generally differs from speech in being more structurally complex, elaborated, and/or explicit. This view was moderated later, when researchers such as Tannen (1982), Beaman (1984), and Chafe and Danielewicz (1986) argued that communicative task is also an important predictor of linguistic variation; therefore equivalent communicative tasks (e.g., narration) should be compared in speech versus writing to isolate the possibility of mode differences.

Multi-Dimensional (MD) studies of register variation in English (e.g., Biber 1986, 1988) went a step further by analyzing linguistic variation among the range of registers within each mode, and then comparing across speech and writing. These studies found 'dimensions' of variation that distinguish between stereotypical 'oral' versus 'literate' registers, although there are few (if any) absolute linguistic differences between speech and writing with respect to any dimension. Rather, particular spoken and written registers are more or less similar/different with respect to each underlying dimension of variation.

More recently, some scholars in the 1990's began to claim that there are essentially no linguistic correlates of literacy as a technology. Many of these researchers have taken an ethnographic perspective, studying literacy practices in communities where writing is used for specific, local functions. Having noticed that those functions do not necessarily include the stereotypical purposes of

informational exposition, these researchers have made general claims minimizing the importance of literacy as a technology; for example:

Literacy can be used (or not used) in so many different ways that the technology it offers, taken on its own, probably has no implications at all. [Bloch 1993, p. 87; see also Halverson 1991; Hornberger 1994]

Against this background, the present paper surveys a number of empirical corpus-based studies to argue that the technology of literacy does have measurable linguistic consequences. Specifically, it will be argued that the written mode enables styles of linguistic expression that are not found in spontaneous (unscripted) speech. That is, the evidence presented below shows that the spoken and written modes differ in their potential for linguistic variation: speech is highly constrained in its typical linguistic characteristics, while writing permits a wide range of linguistic expression, including linguistic styles not attested in speech. Thus, written texts can be highly similar to spoken texts, or they can be dramatically different. This difference is attributed to the differing production circumstances of the two modes: real-time production in speech versus the opportunity for careful revision and editing in writing.

I present evidence for these linguistic differences between the modes from a survey of corpus-based research studies carried out over the past 20 years. These studies consistently document the same general patterns of linguistic variation: 1) few, if any, absolute differences between speech and writing; 2) large differences in the typical linguistic characteristics of the two modes; and 3) a much larger range of linguistic variation in the written mode than the spoken mode.

2. Grammatical variation within and across the two modes

One analytical approach that has proven to be especially useful for the study of linguistic variation is corpus-based analysis (see e.g., Biber, Conrad and Reppen 1998; McEnery, Xiao and Tono 2006). Several corpus-based studies have undertaken detailed lexico-grammatical descriptions of spoken and written registers. For example, Biber et al. (1999) is a reference grammar of English that is based on empirical analysis of corpora from four registers: conversation, fiction, newspaper language, and academic prose (c. 20 million words of text overall, with c. 4-5 million words from each of these four registers). A second study that is useful for the purposes here is Biber (2006), which describes the typical linguistic characteristics of university spoken and written registers (both academic and non-academic).

I focus here on six of the registers described in these earlier studies: conversation, classroom teaching, and office hours within speech; and university textbooks, institutional texts (e.g., university catalogs or handbooks), and fiction within writing. These registers were chosen to illustrate the range of grammatical variation within the spoken and written modes.

Table 1 in Appendix summarizes many of the most important grammatical characteristics of these six registers, based on a survey of the patterns of use documented in the LGSWE (Biber et al 1999) and Biber (2006). Part A of the table lists linguistic features that are especially common in the spoken registers. These are mostly features relating to pronouns (rather than nouns), the verb phrase (verbs and adverbs), and finite clauses. Surprisingly, there are several kinds of dependent clause included here: finite adverbial clauses (e.g., *if I'm lucky, cause he can't smoke*), *that* complement clauses controlled by verbs (e.g., *I don't think [that] he does*), and WH-clauses (e.g., *I don't know what's happening*).

Two patterns are especially noteworthy in Part A of Table 1: First, there is extensive variation within writing. Thus, fiction is relatively similar to conversation in that most of these features are common in both, while none of these features are common in textbooks. In fact, other written registers are even more similar to conversation. For example, consider the dense use of verbs and pronouns in the following e-mail message:

Text Sample 1: e-mail message

Pronouns are **bold underlined**; verbs are *underlined italics*

Hey there,

How's **it** going? **It** won't *be* long now before **you're** down here (or at least close to where **I am** now). **I need** your arrival time, flight number, etc. **We** are *getting* to CC earlier than **I planned**, so **I** will *pick* up the car on the 8th and may *do* something with these folks before **you arrive**. *Let me know* if **you have** any other questions too before **you head** off.

Cheers, LC

The second noteworthy pattern in Part A of the table is that there is little variation within speech. Thus, although conversation and university classroom teaching are dramatically different in their communicative purposes (as well as the relation between speaker and audience), these two spoken registers are surprisingly similar in their linguistic styles. Text samples 2 and 3 illustrate how similar these registers are in their characteristic grammatical features:

Text Sample 2: Conversation

Pronouns are bold underlined; verbs are underlined italics; finite adverbial clauses and finite complement clauses are marked by [...].

<waiting in a car> <very long pause>

Peter: Oh brother.

Gayle: **They** might not even have *left* there yet ... the hotel.

Peter: Yeah **they** were just *getting* organized.

Gayle: Yeah.

Peter: *Were* Bob and Dorothy up already?

Gayle: Oh yeah **they** *were* up. **I** *think* [**we** better *wait*.] **You** *know* [**we** *go* out to breakfast every Sunday after church]. <laugh> And **they**'ll never, **they**'ll never *stay* there. **I** *mean* [**they** always, Bob's always gotta *go* home for some reason]. **He**'s got to *have* his bacon and egg muffin. **We** *took* **him** to breakfast on Sunday, all **he** *did* *was* *complain*. <laugh> Of course **he** *gets* mad [cause **he** can't *smoke* [cause **we** always *take* non-smoking.]]

Peter: Oh well.

Text Sample 3: Classroom teaching; English

Pronouns are bold underlined; verbs are underlined italics; finite adverbial clauses and finite complement clauses are marked by [...].

Instructor: [What I want you to do in your free writes] is kind of reflect on [what do you think [he means here]]. Maybe - and [what you could answer] is would you want to live in that kind of place. Would you want to live there? And [if you do], Why? and do not, Why? And how does Rymmer give you clues? I think [Rymmer, especially in a poem like this, he talks about this hollowness at his core, sort of the absence of the bona fide, legitimate purpose to the whole thing]. I think [clues like this are embedded throughout that suggest [that Rymmer's pretty negative, or skeptical about this whole project]], right? And [what I wanna know] is, [if you do want to live there], why is that, and [if you don't], what is it about Rymmer's writing, or Rymmer's ideas that lead you to believe [that you wouldn't want to live there].

Part B of Table 1 lists features that are especially common in written registers. Many of these features are nouns or features that can be used for noun phrase modification, such as adjectives, prepositional phrases, or relative clauses. Similar to Part A of the table, Part B shows that there is extensive linguistic variation among written registers. Thus, these features are very common in informational written registers, but many of them are not common in fiction. Text sample 4 illustrates the dense use of nouns, attributive adjectives, and post-nominal modifiers typical of informational writing:

Text Sample 4: Medical textbook

There was no significant difference between the two groups regarding blood pressure, family history of ischaemic heart disease, obesity or alcohol consumption. There was, however, a high incidence of heavy alcohol consumption amongst patients who subsequently required coronary artery surgery.

Classroom teaching is especially noteworthy in Part B of the table, in that it does *not* make dense use of nouns and complex noun phrase structures, even though it has similar informational communicative purposes to textbooks. Text Sample 3 is dramatically different from Text Sample 4 in this regard. Text Sample 3 is typical of most classroom teaching in that the instructor uses comparatively few nouns overall, and few complex noun phrases (e.g., with embedded prepositional phrases or relative clauses). Rather, the linguistic complexity of this passage is expressed primarily through the dense use of finite dependent clauses – adverbial clauses with *if* or *because*, and complement clauses (*WH* or *that*).

Thus, communicative purpose has a surprisingly small effect on the typical linguistic characteristics of spoken discourse. That is, whether a speech event is interactive and interpersonal (as in normal conversation), or primarily monologic and informational (as in classroom teaching), it is characterized by the same set of typical linguistic features: verbs, pronouns, finite adverbial and complement clauses, etc. And all of these speech events are characterized by the relative absence of nouns and complex noun phrase structures.

Considering the overall patterns of linguistic variation, Table 1 shows three general distributional patterns: 1) linguistic features that are common in informational writing tend to be rare in the spoken registers, and vice versa; 2) spoken registers are surprisingly similar to one another in their typical linguistic characteristics, regardless of differences in communicative purpose, interactiveness, and pre-planning; and 3) in contrast, written registers have a much wider range of linguistic diversity.

The linguistic uniformity among spoken registers can be attributed to their shared production circumstances. Spoken texts are normally produced in real time. As a result, spoken registers share a heavy reliance on finite clausal syntax. And conversely, it seems that the dense use of complex noun phrase structures – typical of some kinds of written prose – is simply not normally feasible given the production constraints of the spoken mode. At the same time, we find large linguistic differences among written registers, corresponding to differences in purpose, interactiveness, author involvement, etc. This variation is attributed to the production circumstances of writing, which give the author maximum flexibility to choose styles of linguistic expression very similar to those typical of speech, or to produce of expression that are apparently not feasible in speech.

Multi-dimensional studies of spoken and written registers in English (e.g., Biber 1988, 1995; Conrad and Biber 2001) show similar patterns. Several general patterns and conclusions about spoken and written language have emerged from multi-dimensional studies:

- Some dimensions are strongly associated with spoken and written differences; other dimensions have little or no relation to speech and writing;

- There are few, if any, absolute linguistic differences between spoken and written registers;
- However, there are strong and systematic linguistic differences between stereotypical speech and stereotypical writing, that is, between conversation and written informational prose;
- The spoken and written modes differ in their linguistic potential: they are not equally adept at accommodating a wide range of linguistic variation. In particular, there is an extremely wide range of linguistic variation among written registers, because writers can choose to employ linguistic features associated with stereotypical speech. In contrast, there is a more restricted range of linguistic variation among spoken registers. As in the discussion above, I attribute this last pattern to the real-time production circumstances of the spoken mode, making it difficult to employ many of the linguistic features associated with stereotypical informational writing.

3. Historical Evidence

Historical corpus-based studies help us to better understand the nature of these differences between the spoken and written modes. For example, Biber and Clark (2002) investigate historical change in the noun phrase structures commonly used in drama, fiction, and medical prose (from the ARCHER Corpus). The study shows how non-clausal 'compressed' types of noun modification – attributive adjectives, nouns as pre-modifiers, and prepositional phrases as post-modifiers -- have increased dramatically in use over the past 3 centuries. (Clausal noun modifiers, like relative clauses, have remained relatively constant in frequency across these periods.) For example:

Attributive adjectives:

gradually expanding cumulative effect

Nouns as pre-modifiers:

baggage inspection procedures

Prepositional phrases as post-modifiers:

a high incidence **of** heavy alcohol consumption **amongst** patients ...

However, as Figures 1-3 show, the increase in non-clausal modifiers has occurred only in informational written prose (medical prose in this study; see also Biber 2003). In contrast, drama and fiction have shown only slight increases, if at all.

These historical developments can be attributed to two influences: 1) an increasing need for written prose with dense informational content, associated with

the 'informational explosion' of recent centuries, and 2) an increasing awareness among writers of the production possibilities of the written mode, permitting extreme manipulation of the text. Specifically, it seems that the extremely dense use of complex noun phrase constructions (described in Section 2 above) is not normally feasible in speech, regardless of the communicative purpose. As a result, we did not have models for this style of linguistic expression in earlier centuries. As research specialists developed a communicative need for such styles, authors became increasingly aware that the written mode provided extended production possibilities, allowing dense phrasal embedding in ways that are not attested in the spoken mode. Here again, it is important to note that the written mode does not necessitate these distinctive linguistic styles; thus written fiction has not changed historically to employ these complex noun phrase structures. Rather, it seems that the written mode provides possibilities for styles of linguistic expression not normally possible in speech, and that authors have only gradually come to exploit those possibilities over the past four centuries.

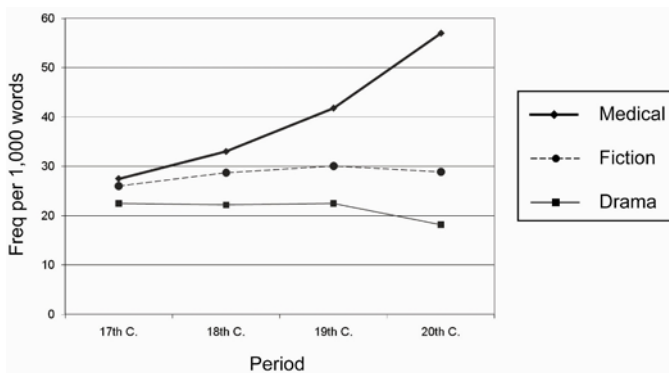


Figure 1. Attributive adjectives across periods

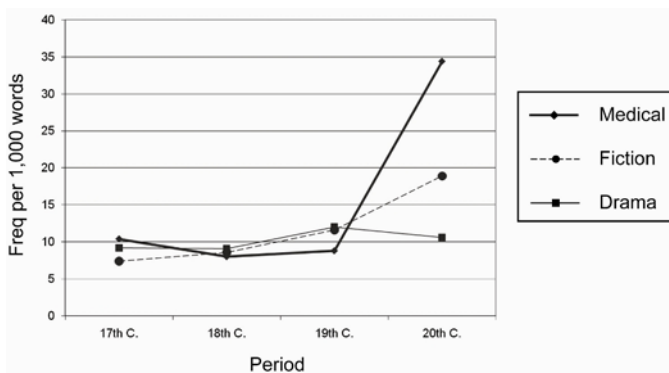


Figure 2. Noun-noun sequences across periods

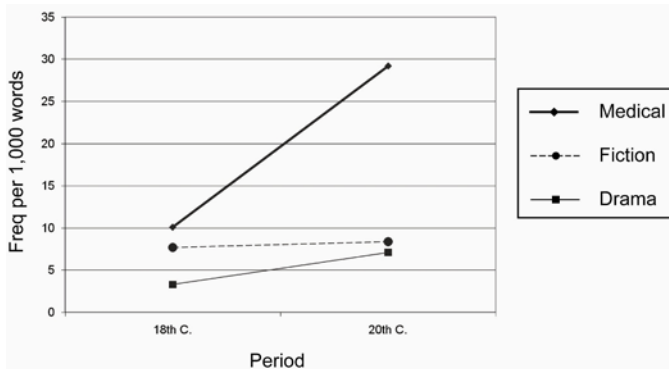


Figure 3. Prepositional phrases as postnominal modifiers, across periods

4. Conclusion

Some individual speakers are especially gifted and can produce notably elaborate and fluent styles of discourse. Oral poets in traditional societies are examples of such gifted speakers. For example, Andrzejewski and Lewis (1964) describe the elaborated linguistic styles typical of oral poetry in Somali. A Somali poet can spend days working on a single poem, to produce a text governed by strict rules of alliteration, with a strong preference to avoid repetitions of words.

The existence of such speakers shows that the production difference between the spoken and written modes is not an absolute one. That is, a few speakers are able to mentally compose dense, lexically elaborated texts, relying on memory without the aid of writing. Such texts go through multiple rounds of planning, revision, and editing, similar to the process of careful production described above for written registers. In this case, the process of careful production and revision relies heavily on an exceptional memory – the entire text is planned, revised, and edited over a period of weeks, relying on the powers of memory. The case of Somali oral poets show that such feats are humanly possible.

However, these are truly exceptional spoken registers. The vast majority of speech, in any language, is not memorized and has not been mentally revised and edited. Rather, speech is normally produced spontaneously in real-time (even if it has been pre-planned, as in the case of university lectures). And corpus-based studies of spontaneous spoken registers have shown consistently that they differ from written registers in that they do not provide the possibility of extreme lexical diversity, or the dense use of complex noun phrase constructions. Rather, such linguistic styles require extensive planning, revision, and editing – processes that are normally possible only in writing.

This does not represent an absolute or necessary difference between speech and writing. Rather, authors can exploit the written mode to produce texts that are very similar to the typical linguistic styles of speech. However, the converse is not true: that is, speakers are not normally able to revise and edit their texts because they are constrained by real-time production circumstances. As a result, some written registers have evolved to employ linguistic styles – with extreme lexical diversity and a dense use of complex noun phrase structures – that are not normally feasible in the spoken mode.

In sum, there are genuine linguistic consequences of literacy; but these consequences have to do with the linguistic potential of the two modes rather than the necessary linguistic characteristics of the two modes. In particular, the present paper has shown that language production in the written mode enables styles of linguistic expression not normally attested in speech, even though writers often choose not to exploit that linguistic potential.

References

- Andrzejewski, B.W. and I.M. Lewis. 1964. *Somali poetry: An introduction*. Oxford: Oxford University Press.
- Beaman, K. 1984. Coordination and subordination revisited: Syntactic complexity in spoken and written narrative discourse. In D. Tannen (ed.), *Spoken and written language: Exploring orality and literacy*. Norwood, NJ: Ablex Publishing, 45-80.
- Biber, D. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.
- Biber, D. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.
- Biber, D. 2003. Compressed noun phrase structures in newspaper discourse: The competing demands of popularization vs. economy. In J. Aitchison and D. Lewis (eds), *New media language*. London and New York: Routledge, 169-181.
- Biber, D. 2006. *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.
- Biber, D. and V. Clark. 2002. Historical shifts in modification patterns with complex noun phrase structures: How long can you go without a verb? In T. Fanego, M. J. López-Couso and J. Pérez-Guerra (eds), *English historical syntax and morphology*. Amsterdam: John Benjamins Publishing, 43-66.
- Biber, D., S. Conrad and R. Reppen. 1998. *Corpus linguistics: Exploring language structure and use*. Cambridge: Cambridge University Press.
- Biber, D. and E. Finegan. 1989. Drift and the evolution of English style: A history of three genres. *Language* 65: 487-517.
- Biber, D. and E. Finegan. 1997. Diachronic relations among speech-based and written registers in English. In T. Nevalainen and L. Kahlas-Tarkka (eds), *To explain the present: Studies in changing English in honor of Matti Rissanen*. Helsinki: Société Néophilologique, 253-276 [reprinted in Conrad and Biber 2001, 66-83].

- Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finnegan. 1999. *Longman grammar of spoken and written English*. London: Longman.
- Bloch, M. 1993. The uses of schooling and literacy in a Zafimaniry village. In B.V. Street (ed.), *Cross-cultural approaches to literacy*. Cambridge: Cambridge University Press, 87-109.
- Chafe, W. 1982. Integration and involvement in speaking, writing, and oral literature. In D. Tannen (ed.), *Spoken and written language: Exploring orality and literacy*. Norwood, NJ: Ablex Publishing, 35-54.
- Chafe, W.L. and J. Danielewicz. 1986. Properties of spoken and written language. In R. Horowitz and S. J. Samuels (eds), *Comprehending oral and written language*. New York: Academic Press, 82-113.
- Conrad, S. and D. Biber (eds). 2001. *Variation in English: Multi-Dimensional studies*. London: Longman.
- Halverson, J. 1991. Olson on literacy. *Language in Society* 20: 619-640.
- Hornberger, N.H. 1994. Continua of biliteracy. In B.M. Ferdman, R.M. Weber and A.G. Ramirez (eds), *Literacy across languages and cultures*. Albany: State University of New York Press, 103-139.
- McEnery, T., R. Xiao and Y. Tono. 2006. *Corpus-based language studies*. London: Routledge.
- O'Donnell, R.C. 1974. Syntactic differences between speech and writing. *American Speech* 49: 102-110.
- Olson, D. 1977. From utterance to text: The bias of language in speech and writing. *Harvard Educational Review* 47, 3: 257-281.
- Tannen, D. 1982. The oral/literate continuum in discourse. In D. Tannen (ed.), *Spoken and written language: Exploring orality and literacy*. Norwood, NJ: Ablex Publishing, 1-16.

Appendix

Distinctive linguistic characteristics of spoken registers compared to written registers [based on Biber et al. 1999 and Biber 2006]

Table 1. Part A: Features generally more common in spoken registers (and fiction)

Linguistic Feature	SPOKEN UNIVERSITY REGISTER		OTHER SPEECH	WRITTEN REGISTER		
	Classroom Teaching	Office Hours	Conversation	Textbooks	Institutional Writing	Fiction
verbs (e.g., <i>get, go, see</i>)	**	**	**			**
progressive aspect (e.g., <i>bringing, making</i>)	*	*	*			*
adverbs (e.g., <i>here, now, again</i>)	*	*	*			*
stance adverbs:						
certainty adverbs (e.g., <i>really, actually</i>)	**	**	**			
likelihood adverbs (e.g., <i>probably, maybe</i>)	*	**	*			
pronouns (e.g., <i>I, you, it</i>)	*	**	**			*
discourse markers (e.g., <i>ok, well</i>)	*	**	**			
adverbial clauses:						
conditional clauses (e.g., <i>if you read...</i>)	*	**	*		*	
causative clauses (e.g., <i>because ...</i>)	*	**	*			
temporal clauses (e.g., <i>when/while ...</i>)			*			*
that-clauses:						
certainty verb + <i>that</i> -clause (e.g., <i>I know [that]...</i>)	**	**	*			*
likelihood verb + <i>that</i> -clause (e.g., <i>I think/guess [that]...</i>)	*	**	**			*
communication V + <i>that</i> -cls (e.g., <i>he said [that]...</i>)	*	*	*			*
WH-clauses (e.g., <i>you know how to...</i>)	*	**	**			*

** = extremely common; much more frequent than in other registers

* = very common; more frequent than in other registers

Table 1. Part B: Features generally more common in written registers

Linguistic Feature	SPOKEN UNIVERSITY REGISTER		OTHER SPEECH	WRITTEN REGISTER		
	Classroom Teaching	Office Hours	Conversation	Textbooks	Institutional Writing	Fiction
diversified vocabulary (e.g., <i>sanctimonious</i>)				*		
nouns/nominalizations (e.g., <i>term, assumption</i>)				**	**	*
rare nouns (e.g., <i>abscission, ambivalence</i>)				*		
common abstract / process nouns (e.g., <i>system, factor, problem</i>)				*	*	
style adverbs (e.g., <i>generally, typically</i>)				*		
adjectives (e.g., <i>important, likely</i>)				*	*	*
linking adverbials (e.g., <i>however, for example</i>)				**	*	
passive voice (e.g., <i>was determined</i>)				*	*	
relative clauses (e.g., <i>the sequence which determines ...</i>)				*	*	*
prepositional phrases (e.g., <i>patterning of behavior by households</i>)				**	**	*
noun + <i>that</i> -clause (e.g., <i>the fact/assumption that... </i>)				*		
mental verb + <i>to</i> -clause (e.g., <i>remember to...</i>)				*	*	
adjective + <i>to</i> -clause (e.g., <i>unlikely/difficult to...</i>)				*		
noun + <i>to</i> -clause (e.g., <i>the opportunity to learn</i>)					*	

** = extremely common; much more frequent than in other registers

* = very common; more frequent than in other registers

INFORMATIONAL PATTERNING THEORY AND THE CORPUS-BASED DESCRIPTION OF SPOKEN LANGUAGE

THE COMPOSITIONALITY ISSUE IN THE TOPIC-COMMENT PATTERN

Emanuela Cresti, Massimo Moneglia¹
LABLITA – University of Florence

1. Introduction

1.1 *Speech Act Theory* and *Teoria della lingua in atto*

The *Language into Act Theory* (*Teoria della lingua in atto*, see Cresti 1987-2000) and the *Informational Patterning Theory* (*Teoria dell'Articolazione dell'informazione*, see Cresti 1994; Moneglia 1994; Scarano 2003; Scarano 2009; Moneglia and Cresti 2006; Cresti and Moneglia, in press), derive from *Speech act Theory* (Austin 1962), which assumes that *speech acts* are the reference unit for linguistic behavior. Austin claims that one speech act is composed by the simultaneous performance of three different acts: *locutionary*, *illocutionary* and *perlocutionary* acts. The *utterance* is the linguistic entity accomplished by the speech act. In accordance with Austin, we assume that the utterance is the reference unit for the analysis of the spoken language. In this framework, in line with Austin's approach, it receives an operative definition, that is: "every linguistic expression that can be pragmatically interpreted" that is crucial for spontaneous speech analysis.

At least two assumptions of *Language into Act Theory* depart from Austin's frame. First of all, the perlocutionary act is not conceived as the intentional accomplishment of non conventional goals, that is, as a proper act (like the illocutionary and the locutionary acts). Perlocution is rather identified as the *affective base* which is necessary to activate each speech act. We are not going to deal with this topic here. Secondly, *prosody* plays a crucial role in this frame. Prosody is mentioned by Austin only among the possible devices that, together with lexicon and syntax, can be used to perform the illocutionary act. On the contrary, prosody is assumed here to be the necessary interface between the illocutionary and

the locutionary act, and it is identified as the main objective mark for the identification of the utterance itself.

This perspective finds a direct application in the study of spontaneous speech. The long debate concerning the nature of spoken language crosses over two centuries (Weill 1844; Mathesius 1929; Bally 1932; Hockett 1958; Halliday 1967, 2004; Chomsky 1972; Blanche-Benveniste 1991; Lambrecht 1994). One of the most relevant questions regards the informational organization of speech. As it is well known, a large part of the utterances appear composed by two or more parts (*theme-rheme, known-new, topic-comment, topic-focus, head-body-tail, prefix, noyau, suffix*), often supported by *discourse markers*. However the foundation of information structure has been proposed according to very different assumptions, which consider it respectively within the domain of semantics, syntax, or directly as functional relations. The most important innovations by the *Language into act Theory* and *Informational Patterning Hypothesis* is that the information is ruled within actual spoken language use according to pragmatic principles, and that its linguistic organization is signaled by prosody (Cresti 1987).

This paper will focus on the Topic-Comment relation which is the core information pattern of spoken utterances, and will demonstrate, on the basis of evidence bootstrapped from Italian corpora, the independence of informational relations with respect to syntax. As a whole we will see that the performance of two strings in a Topic-Comment pattern determines the onset of two local syntactic and semantic domains which are not compositionality bound. We will propose that if the informational relation holds, then compositionality does not hold and vice versa. Before entering into the topic of this paper we will sketch briefly our theory, which is strongly based on both distributional and experimental research carried out at LABLITA on spoken Italian Corpora².

1.2 Information patterns and Prosody

Prosody is the unconscious mark of our attitudes in speech (emotions, feelings, affects) and because of its nature, it is an adequate medium to interface the illocutionary act and the locutionary act with their affective base (perlocution). Every language has melodic shapes conventionally codified in order to perform the various types of illocutionary values, that are much more numerous in speech than the sentence modalities of written language (*assertion, order, question, optative*)³.

From our perspective the core of the utterance corresponds to a part, called Comment, which necessarily deals with one prosodic unit (PU), and constitutes the information unit (IU) whose function is to accomplish the illocutionary force of the utterance. For this reason the Comment IU is necessary and sufficient to give rise to an utterance. More in general, this theory foresees that prosody is also the formal device devoted to the performance of the informational patterning of the utterance.

Within the utterance, various IU types, optional from an informational point of view, can surround the Comment⁴. They also correspond to prosodic units and are divided into two classes which are respectively dedicated to different types of information functions: a) the textual construction of the utterance (Topic, Appendix, Parenthesis, Locutive Introducer); b) its communicative support (Incipit, Phatic, Allocutive, Conative, Connector).

The systematic correspondence between the information pattern (IP) and the prosodic pattern (PP) is the assumption of *Informational Patterning Hypothesis*. The prosodic pattern is composed by a set of prosodic units (PU), roughly isomorphic with the set of IUs of an information pattern⁵. The following table shows a comparison schema between the most common IU types taking part to an information pattern and the corresponding PU types which necessarily perform them⁶.

Prosodic Pattern			Information Pattern	
	root	→	Comment Tag: COM	
(prefix)	(suffix)	→	(Topic) Tag: TOP	(Appendix) Tag: APC
	(parenthetical)	→	(Parenthesis) Tag: PAR	
(incipit)	(phatic)	→	(Incipit) Tag: INC	(Phatic) Tag: PHA

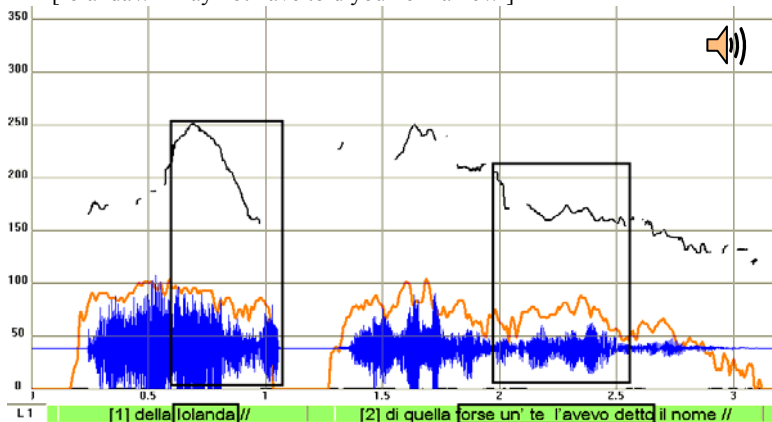
Figure 1. The correspondence between PU types and IU types of the utterance

Each prosodic unit type has specific prosodic features (F_0 movements, intensity, length, timing), and conditions on its distribution. Each PU is separated by a non terminal prosodic break within a prosodic pattern. A prosodic pattern is concluded by a terminal prosodic break. The specific prosodic form of any PU type implies conditions on distribution: for instance, a *prefix* PU must precede a *root* PU, a *suffix* PU must follow a *root* PU, a *parenthetical* PU can precede or follow a *root* PU and also be inserted in a *root* or *prefix* PU, but it cannot occur in the opening of an utterance. All of these conditions have melodic reasons. *Root* and *prefix* PUs, for instance, host a prosodic Nucleus with a primary stress, while all the other PUs do not⁷.

As a whole, in this framework, utterances correspond to patterns of information units that can consist in one only Comment IU (simple utterance) or in many IUs besides the Comment (compound utterance).

Let us see for instance one dialogical turn which contains two simple utterances, each one composed by the Comment IU only⁸:

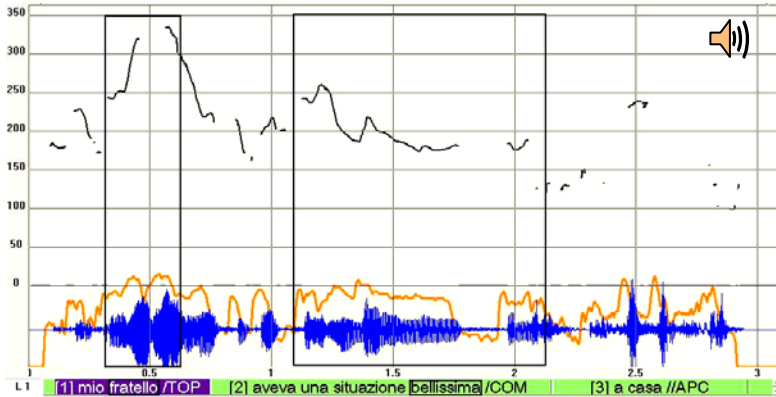
- (1) della Iolanda // di quella forse 'un te l'avevo detto il nome //
 [ifamd102, 147-151]
 [Iolanda // I may not have told you her name //]



In 1, the proper name *Iolanda* fills a Comment IU which accomplishes an illocution of answer. This expression constitutes one utterance which receives a pragmatic interpretation, even in isolation. In the second part of 1, the sentence *I may not have told you her name* also constitutes a Comment IU, accomplishing a new speech act with the illocution of hypothesis. Both utterances are accomplished through a PU of the *root* type, ending with a terminal prosodic break. The “aboutness” of both illocutionary acts is found in the dialogic context.

Compound utterances are information patterns built up around the pragmatic center constituted by a Comment IU and are more frequent in speech than simple utterances. Nearly the 58% of the utterances in the Italian section of the C-ORAL-ROM corpus are compound. For instance, 2 and 3 are compound utterances and, in accordance with this theory, their information pattern corresponds to their prosodic parsing.

- (2) *ANT: mio fratello /TOP aveva una situazione bellissima /COM in casa //APC
 [ifamd105, 95]
 [my brother / used to have a wonderful deal / at home //]



- (3) *ANT: io /TOP proprio /PAR provinciale al massimo //COM [ifamd105, 30]
[I / really / small-town girl at all //]

In 2, a Topic IU, a Comment IU and an Appendix IU, occur in sequence. The Comment develops a narration illocution. In 3, a Topic IU, a Parenthesi IU and a Comment IU occur in sequence, and the Comment develops an expressive illocution of irony.

Corpus data make it easy to verify that the pragmatic interpretation of the Comment unit is always ensured independently from its relation with the other IUs and from its syntactic type, since prosody conveys the illocutionary force of the utterance. In all the above utterances the interpretability of the Comment is ensured, even if the other IUs are erased.

Let us see more examples. 4 is composed by a Comment IU and an Appendix IU, while 5 by a Topic IU, a Comment IU and a Parenthesis IU.

- (4) *DIN: piazzale Attilio Luzzatto /COM l'era proprio //APC [ifamn0, 192]
[Attilio Luzzatto square / that is what it was //]
- (5) *MIR: soldi e cellulare /TOP se li può pure tene' /COM se è stata lei //PAR
[itelpv13, 329]
[money and mobile / she can even keep them / if she did it //]

Each IU is performed through a PU, specific for its formal character in accordance to its information function type and is concluded by a non terminal break. The identification of the Comment as the interpretable IU within the utterance strictly depends on the illocutionary force conveyed by the *root* PU type and is not connected to its syntactic form. For instance, in 2, the Comment is a VP, an ADJ in 3, one single NP in 4, and a Sentence in 5.

The most common information pattern in spontaneous spoken language is Topic-Comment that records nearly 40% of compound utterances. The pattern is built on the relation between the expression of the Comment IU, accomplishing the illocutionary force of the utterance, and the expression in Topic IU. Both the Topic and the Comment units necessarily bear a *semantic focus*; i.e. a semantic prominence identified by perceptively relevant prosodic features on a semantic word. The Topic Comment pattern is performed through a *prefix-root* pattern, whose prosodic properties can be summarized as follows.

The essential feature of the root unit is its pragmatic interpretability which is clear to perceive also in isolation. It is characterized by a well determined spectrogram, and by a strong intensity. The root unit records a prosodic Nucleus, which can be preceded by a Preparation, and followed by a Tail, in accordance with its illocutionary value. The Nucleus corresponds to a Semantic Focus. Roughly 30 root unit types have been found (Firenzuoli 2003; Cresti and Firenzuoli 1999), each characterized by a Nucleus with a specific prosodic form (depending by the illocutionary value it conveys). The seat of the Nucleus within the *root* unit can occur on the left or right side, in accordance with its illocutionary value.

Also, the *prefix* PU, must record a prosodic Nucleus, which can be preceded by a Preparation, but cannot be followed by a Tail. The *prefix* Nucleus is frequently more prominent from a prosodic and phonetic point of view than the Comment's one; however, it is always perceived when suspended, non-concluded, and it cannot be interpreted in isolation. The seat of the Nucleus is necessarily on the right end of the *prefix* unit and also corresponds to a Semantic Focus. It is characterized by the frequent lengthening of the last syllable (stressed or unstressed), by a well determined spectrogram, by a strong intensity and it cannot be speedy. The above prosodic properties, however, do not define a prosodic form. Possible variants depending on the speaker and language context may occur⁹.

The information function of the Topic is to identify the domain of relevance for the Comment illocutionary force selecting, through linguistic means, its pragmatic domain of application. In other words the Topic specifies the pragmatic *aboutness* of the Comment¹⁰. Providing the domain of relevance for the illocutionary act, the Topic allows to *distantiate* the Comment from the direct context of the utterance and, in so doing, it makes the interpretation of the utterance autonomous from the context itself. This information structure seems to be very "primitive" and occurring in every language, and as we will demonstrate in this paper, because of its pragmatic nature it is *independent from semantic and syntactic relations*.

1.3 The relation between Information patterns and syntax

The relation between syntax and prosody is one of the most relevant questions on the nature of spoken language. The discovery of the informational organization of

the utterance demonstrates that in speech there is not a direct relation between syntax and prosody, but that they are mediated by informational patterning. The prosodic parsing of the utterance and the form of prosodic unit types cannot be considered at the level of the performance of the syntactic structure, or as style variations of the locutors, since they are conventional marks of the information pattern.

Generally speaking, *Informational patterning theory* claims the dominance of information patterns on syntax. The different IUs display their informational value with respect to the Comment, which accomplishes the illocutionary act, and prosody works as an interface between the illocutionary act, which is patterned in different IUs, and the locutionary act. Syntax is a level inside the locutionary act, such morphology, lexicon or modality. Therefore, in this frame the domain of syntax is local, holding within any IU identified by prosody, while the overall structure of the utterance and its main prosodic properties are governed by informational relations.

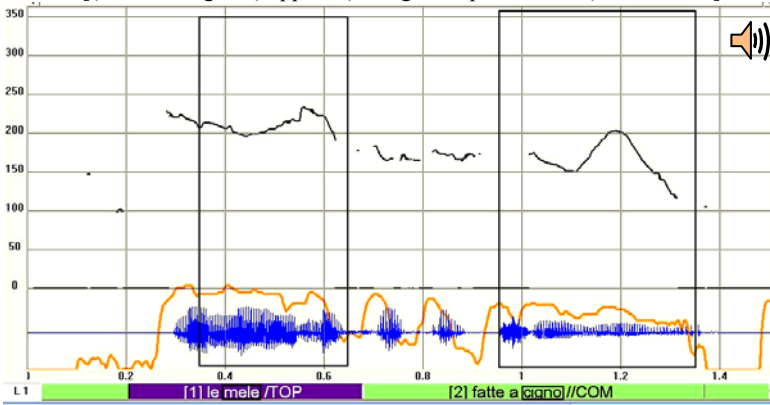
Looking at specific language contexts derived from Italian corpora we will argue that the main compositional relations are not compatible with Topic-Comment interpretation. In paragraph 2 we will see that a NP cannot be compositionally settled if its lexical filling is performed through a Topic-Comment informational structure. In paragraph 3 the hypothesis that the Topic-Comment structure impedes the onset of a sentence structure will be supported by looking into the main syntactic distribution of Topic-Comment structures in spoken corpora; i.e. a) anacolutes, b) circumstantial arguments, c) subject- predicate relation. In parallel we will also argue that the informational function specified by the Topic IU belongs to the domain of pragmatics. To this end the classical definitions of Topic in terms of semantic “aboutness” will be challenged to face the one proposed by *Informational patterning theory*. We will show that the pragmatic interpretation has higher descriptive adequacy. In paragraph 4 we will support the hypothesis that Topics are syntactic *islands* on the base of various independent arguments derived from corpus data: a) semantic restrictions on Topic; b) distribution of clitics in Topic; c) modal compositionality.

2. The Topic-Comment pattern and the NP structure

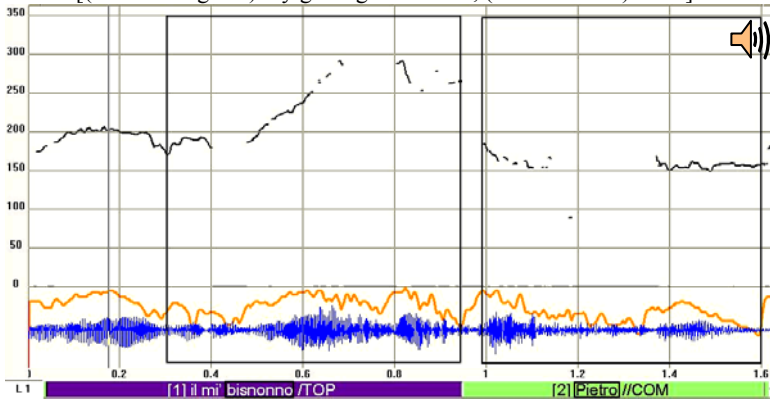
A string corresponding to a compositional NP structure cannot be interpreted as a syntactic constituent if it has the positive prosodic properties of a *prefix-root* pattern and corresponds to a Topic-Comment information structure.

Spontaneous speech corpora provide wide evidence for this. In spontaneous speech, verb-less utterances, lacking a verbal form with a finite mood, such as 6 and 7 are frequent (nearly 38%) and well-formed:

(6) *VER: le mele /TOP fatte a cigno //COM [ifamd14, 41]
 [(for what regards) apples, (the right shape should be) like a swan]



(7) *LID: il mi' bisnonno /TOP Pietro //COM [ifamd102, 75]
 [(for what regards) my great-grand father, (his name was) Peter]



A bare transcription of 6 and 7, without prosodic marks and tags for information functions, will suggest a compositional NP interpretation, based on the syntactic relation (attribution or modification) between a Noun and AdjP or between a Common Noun and a Proper Name.

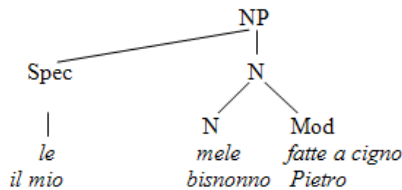
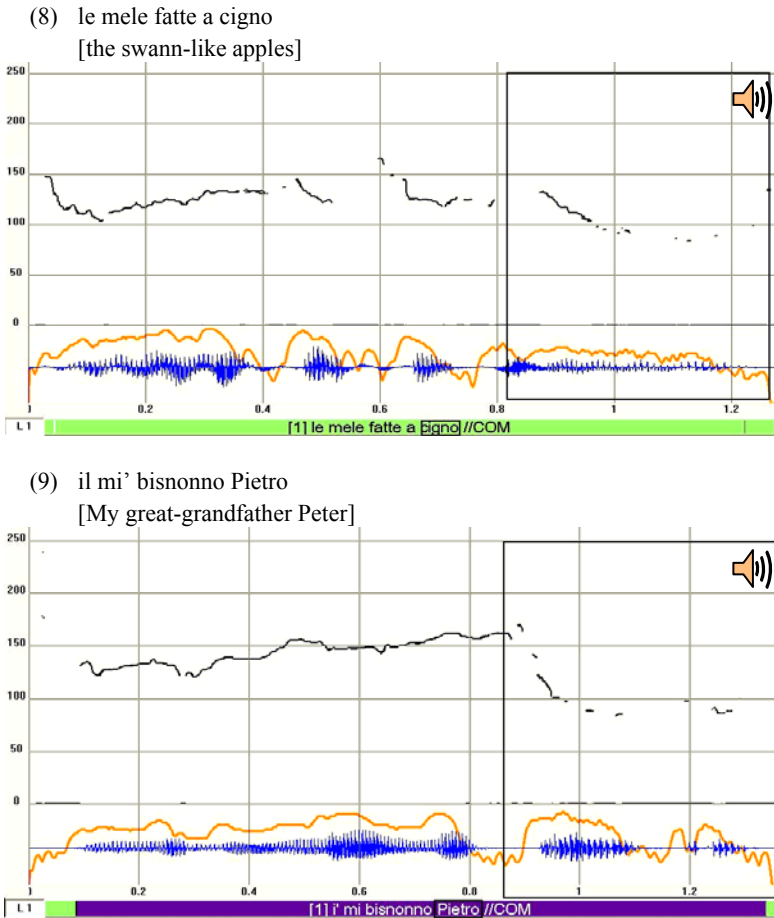


Figure 2. NP structure

This interpretation of actual speech data is obviously refused by competent speakers. However *Informational Patterning Theory* provides an explicit reason for this judgment. No NP can arise from 6 and 7 because of the prosodic patterning of the utterance. The NP interpretation should had been appropriate only if the two constituents were *linearized* by prosody within one PU, corresponding to an only IU.

For instance, 8 and 9 (laboratory utterances, performed by a male speaker, with the same lexical filling of 6 and 7) have been performed through one *root* PU hosting only one primary stress and without any internal prosodic patterning:



In this case 8 and 9 show an NP structure, i.e. they are semantically compositional, respectively denoting “some like swan apples” and “the great-grandfather Peter”. As a whole both 8 and 9 in the above laboratory utterance are verb-less speech acts with a force of answer.

6 and 7 differ from 8 and 9 for their prosodic and informational properties. The two nominal parts are performed through a *prefix-root* prosodic pattern. In 6, the

expression in Comment is accomplishing an expressive force of obviousness (it is obvious that the best shape for a table decoration is *swan like*), and the one in Topic specifies the pragmatic domain to which this act is referred (*the apples*). In 7, the illocutionary force of the answer, that the name is *Peter*, must be applied to the dialogical premise *the great-grandfather*.

Therefore, while in the linear performance (8 and 9) the syntactic relation among constituents of a NP are “alive”, in 6 and in 7 no syntactic structure is in act between the NP behaving as Topic and the AdjP or the NP behaving as Comment. The Topic-Comment interpretation obviously impedes the compositional interpretation of the two phrases as a complex NP. In other words, the informational relation between the NP in Topic and the AdjP or NP in Comment, causes the onset of two distinct syntactic domains, which cannot be compositionally linked in a syntactic constituent¹¹.

In summary, the positive presence of the informational relation between two IUs, conveyed by their prosodic pattern, impedes the onset of semantic and syntactic compositionality of nominal structures. We will see in the next sections that this property leads to much more compromising interpretations, if informational relations are taken seriously in spoken language analysis.

3. Topic-Comment pattern and Sentence structure

3.1 The linguistic filling of Topic-Comment pattern: circumstantial arguments

Some corpus based investigations (Scarano, 2004) have shown that in spoken language Comment IUs can be “filled” in principle by any kind of constituents; i.e. a sentence, any kind of phrase, a lexical entry, a pronoun. Only a very general negative condition prevents the filling by bound and free morphemes (*article, preposition, auxiliary verbal form, conjunction, clitic*) which cannot play any information function (Cresti, 2000). This is true also for Topic, although some limits must be considered (see 4.1. below). However the distribution of syntactic types within Topic and Comment shows important quantitative preferences in the speech performance, that turn out roughly complementary (Signorini, 2004a and 2004b).

Comment IUs are filled by:

Nearly 62%: VP(5% *main sentence*)

Nearly 38%: AdvP, AdjP, PP, NP

Topic IUs are filled by:

Nearly 60% NP

Nearly 40%: *if* and *when* clause, *modal* clause, AdvP, AdjP (*qualificative*), PP, *main sentence*

Let see some examples highlighting the variation of syntactic filling within the Topic-Comment structure:

Two VPs

- (10) *CLA: [13] mòre uno /TOP che te sta vicino /APT neanche te ne accorgi //COM [ifamm02]
[someone dies / near to you / you don't even realize it]

Two AdvPs

- (11) *SAB: per ora/TOP no //COM [ipubl03, 134]
[by now / no]

AdjP and VP

- (12) *APR: mensile /TOP costa un po' di più //COM [LAB -Art]
[monthly / it costs something more]

If or when clause and a main clause

- (13) *UO1: e quando un uomo politico si commuove /TOP è un cretino //COM [LAB-Gara1]
[and when a politician show himself emotional / he is an idiot]

PP and VP

- (14) *VER: poi /INP di guarnizione finale /TOP ci potrebbero essere anche le mele //COM [ifamd14, 40]
[then / like final decoration / there should be some apples too]

In the previous examples, which are all performed within a prosodic *prefix-root* pattern, the expressions behaving as Topic are external to the regency of the Verb in Comment. In the first three utterances (10-11-12), the Topic-Comment structure cannot be directly mapped onto a well-formed compositional structure, since they are *anacolutes* from a syntactic and semantic point of view.

Anacolutes, in order to receive an interpretation, strictly require a *prefix-root* prosodic pattern and will be meaningless as constituents linearized within the same PU. This is coherent with the hypothesis that the prosodic pattern conveys an informational value that is *not* included in compositional semantics, since in the given prosodic conditions the informational relation holds where the compositional relation does not.

On the contrary 13-14 are “Circumstantial structures”, which might be in principle interpreted following compositionality rules. For instance, they can receive a propositional interpretation without knowing their prosodic counterpart (i.e. as in writing). However the prosodic form of these spoken utterances is a positive property, which crucially mirrors the *prefix-root* pattern of 10, 11, and 12. In parallel, from an informational point of view, they can also receive an interpretation in line with the Topic-Comment relation, as it has been previously defined (see. the paraphrases below in 13’ and 14’).

Does it mean that the Topic-Comment interpretation and the propositional interpretation of 13-14 are equivalent? In other words, are propositional interpretations a subset of all possible Topic-Comment interpretations?

The compositionality issue is linked to the definition of the Topic informational function. In our interpretation, the Topic specifies the “aboutness” for the illocutionary act accomplished by the Comment; i.e. it is a *pragmatic relation*. It may be interesting to notice that according to the corpus of influential literature (Sperber and Wilson 1986; Lambrecht 1994) Topic is rather defined as the “*aboutness of a predication*”; i.e. it is a *semantic relation*. More specifically, the definition of Topic in terms of *semantic aboutness* leads to compositional structures, since the results of an aboutness relation is always a proposition.

If the traditional interpretation of Topic is assumed, however, various consequences can be observed in this domain of facts. Actually the role played by the illocutionary force in the construction of the utterance is not properly considered and this causes huge problems in the interpretation of the speech performance.

For instance the following paraphrases show that the interpretation in terms of *semantic aboutness* of Topic filled by circumstantial arguments (13 and 14) is indeed equivalent to a proposition, but the *pragmatic* interpretation, that better fits with natural data, means different things. Moreover *anacolutes* cannot receive at all this reading.

(13’) *UO1: e quando un uomo politico si commuove /TOP è un cretino //COM
[LAB-Gara1]
[and when a political man is moved / he is an idiot]

Semantic: *It is disapproved that the property of being an idiot is about the events in which a politician shows himself emotional.* The paraphrase is a proposition.

Pragmatic: *The act of disapproval “he is an idiot” is about the domain of relevance identified by “when a politician shows himself emotional”.* The paraphrase corresponds to the utterance, but it is not a proposition.

- (14') *VER: poi /INP di guarnizione finale /TOP ci potrebbero essere anche le mele
//COM [ifamd14, 40]
[then / as for final decoration / there should be some apples too]

Semantic: *It is made the hypothesis that the possibility of using apples is about a final decoration.* The paraphrase is a proposition.

Pragmatic: *The act of hypothesis "there should be some apples too" is about the domain of relevance identified by "for final decoration".* The paraphrase corresponds to the utterance, but it is not a proposition

The interpretation in terms of pragmatic aboutness specifies what the speech act is about in the given context, while the semantic aboutness interpretation gives rise to a propositional speech act that may be appropriate in the context, but its domain of relevance is not specified, therefore the two paraphrases are not equivalent.

If the notion of semantic aboutness is taken seriously this interpretation leads to even more embarrassing results when it is applied to anacolutes:

- (10) *CLA: mòre uno /TOP che te sta vicino /APT neanche te ne accorgi //COM
[ifamm02, 13]
[someone dies / near to you / you do not even realize it]

Semantic: *It is protested that the fact that you even do not realize it is about the death of somebody near you.* The paraphrase is a proposition, but it has little meaning.

Pragmatic: *The act of protest "you do not even realize it" is about the domain of relevance identified by "someone dies near to you".* The paraphrase corresponds to the utterance, but it is not a proposition.

- (11') *SAB: per ora /TOP no //COM [ipubdl03, 134]
[by now / no]

Semantic: * *It is refused that the quality of negation is about the present moment.* The paraphrase is a meaningless proposition.

Pragmatic: *The act of refusal "no" is about the domain of relevance identified by "by now".* The paraphrase corresponds to the utterance, but it is not a proposition.

- (12') *APR: mensile /TOP costa un po' di più //COM [LAB –Art]
[monthly / it costs something more]

Semantic: *It is asserted that the fact that something has high cost is about monthly (payments).* The paraphrase is a proposition, if some elliptical semantic material is restored, but it has little meaning.

Pragmatic: *The act of assertion "its cost is higher", is about to the domain of relevance identified by "monthly".* The paraphrase corresponds to the utterance, but it is not a proposition.

In other words, if it is made explicit that the Topic establishes a *pragmatic relation of aboutness* with the act accomplished by the Comment, the same interpretation can be assigned to all the syntactic fillings of the *prefix-root* prosodic pattern under consideration¹². This is not the case when the Topic is defined in terms of a *semantic relation of aboutness*, which generates wrong or unnatural results. For this reason the latter is at least less general with respect to the domain of facts into object.

At the same time if the Topic-Comment relation is read according to the pragmatic interpretation, then it is not equivalent to a compositional structure for obvious reasons. Indeed, compositionality requires a semantic relation to hold among components, but this is not the case if a *pragmatic* relation is established in Topic-Comment patterns. Therefore the compositionality issue relies on the definition of the Topic-Comment relation, which is not compositional as far as it does not belong to semantics.

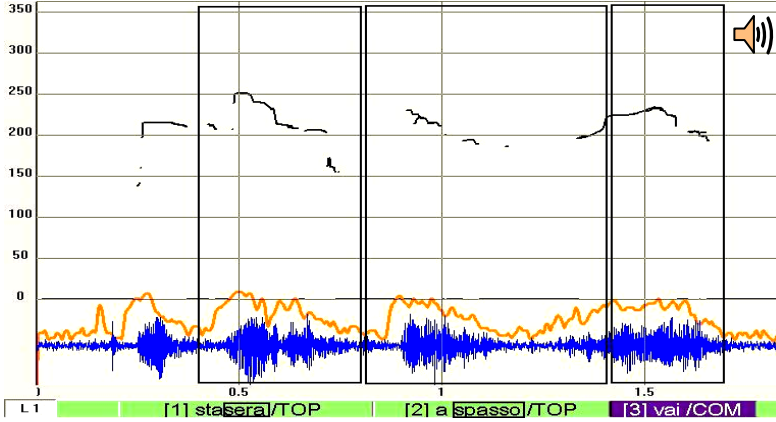
The lack of generality in the definition of Topic in terms of *semantic aboutness* and the appropriateness of the pragmatic definition turns out crucial, if not only syntactic distribution, but also *illocutionary distribution* of spoken language variation is considered.

According to a sampling of dialogues with a high interactive nature (Firenzuoli, 2003), around 40% of utterances accomplish non-assertive illocutionary acts. These acts are specified by the prosodic form of the *root* unit of the utterance copying with Comment IUs¹³. For instance, in our previous examples only 12 and 14 are assertions. This is relevant to this paper since a good lot of directive utterances show a Topic-Comment pattern that deals with a *prefix-root* prosodic pattern. These utterances further demonstrate that Topic relation belongs to the realm of pragmatics, rather than to semantics, and that predication and illocutionary acts cannot be confused.

The following examples respectively instantiate one act of *advice* and one act of *instruction* that are expressed through their prosody. In both examples the Topic specifies the domain of relevance to which the directive act accomplished by the Comment is about and allows the *Displacement* of the utterance from the extra-linguistic context¹⁴.

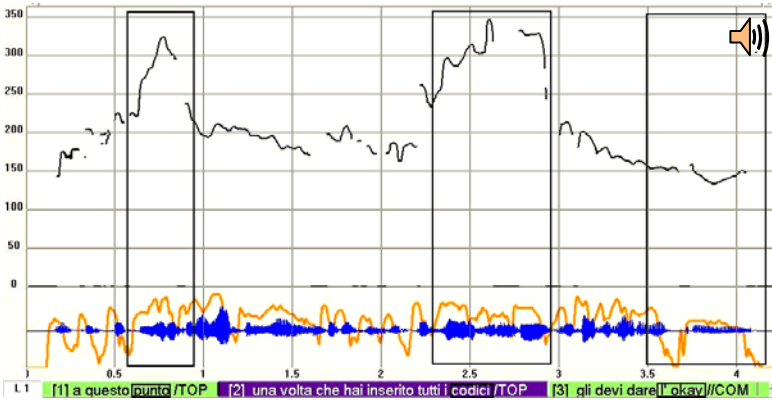
(15) DAN: stasera /TOP a spasso /TOP vai //COM [ifamcv15, 20]
[this night / walking around for fun / go]

- The advice *go* is about the two domains identified by *this night* and *walking around for fun*



(16) a questo punto /TOP una volta che hai inserito tutti i codici /TOP gli devi dare l'okay //COM [ifammn17, 64]
 [at this stage / after you have inserted all the codes / you must give your O.K.]

- The instruction *you must give your OK* is about the two domains identified by *at this point* and *after when you have inserted all the codes*



No paraphrase in terms of semantic aboutness can be proposed. For instance 15 cannot receive the paraphrase “my advice is that *the property of going* is about *this night*”, as the *semantic aboutness* will require. Again, 16 cannot receive the paraphrase “my instruction is that *giving the OK* is about *the moment when all codes are inserted*”. The proper interpretations strictly regard the pragmatic aboutness of the two directive illocutionary acts conveyed by the Comment IU¹⁵.

Formally, the *pragmatic aboutness* cannot be in the scope of the illocutionary act expressed by the Comment unit, since it is actually the illocutionary act that is said “about something”.

In summary, the study of the distribution of Topics in spoken corpora with respect to syntactic filling and illocutionary types makes clear that the Topic-Comment relation conveys an informational aboutness relation of pragmatic nature, rather than a semantic one. In parallel, the propositional interpretation determined by semantic compositionality and Topic-Comment interpretations mean two different things. *Propositions* are semantic entities which do not contain the notion of *referring one act to one domain*. Therefore, if one utterance is interpreted according to the Topic-Comment pragmatic relation, it cannot be simultaneously interpreted as a proposition. Circumstantial arguments compositionally settled within a propositional structure are not equivalent to Topics in Topic-Comment structures, since they do not specify the pragmatic aboutness of the illocutionary act, but are rather external predications.

One compound utterance with a Topic-Comment pattern can receive a compositional interpretation instead of the informational one if, by chance, the locutive content of the two IUs can be compositionally settled because of their syntactic form (as in 14 and 15). In this case the prosodic cues that positively convey the informational function may be ignored by the addressee and the utterance is interpreted as a sentence (see 3.3.1.). In the next paragraphs we will challenge this hypothesis with respect to the core of compositionality; i.e. the regency of the Verb in Comment.

3.2 The linguistic filling of Topic-Comment pattern. NP-VP relation and the “apartness” of Topic

Considering quantitative data, it is evident that the Topic-Comment pattern can also in principle correspond to a Sentence structure, given that:

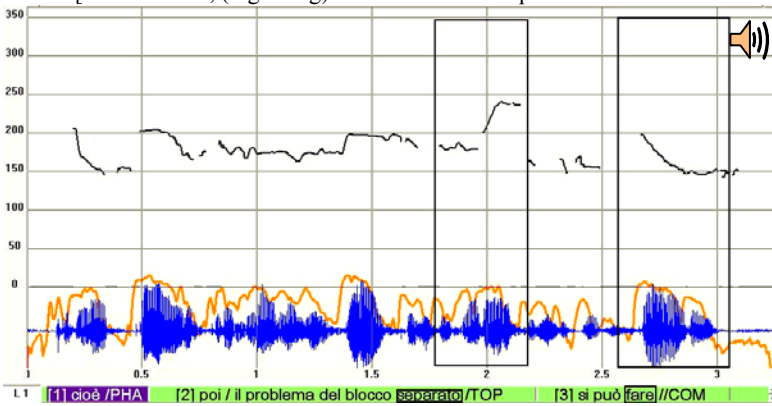
Topic = NP (nearly 60%)
 Comment = VP (nearly 62%)

However, although the most part of constituents in Topic are NP, a large part of them develop Space, Time relations, or work as anacolutes, thus still behaving as “circumstantial” constituents. See for instance 17, 18, and 19:

(17) *WOM: perché la Lampa /TOP non credo //COM [ifamcv28, 248]
 [because (at) the Lampa / I don't believe so]

(18) *LEO: la fine dell'anno /TOP sarà dura /COM magari //PAR [ifamd102, 371]
 [(at) the end of the year / it will be hard / eventually]

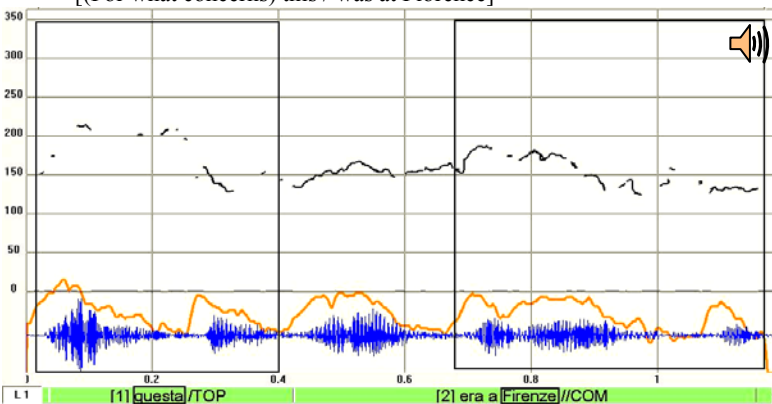
- (19) *MIK: cioè /PHA poi / il problema del blocco separato /TOP si può fare //COM
 [LAB-Art]
 [I mean / then, (regarding) the matter of the separated block / it can be done]



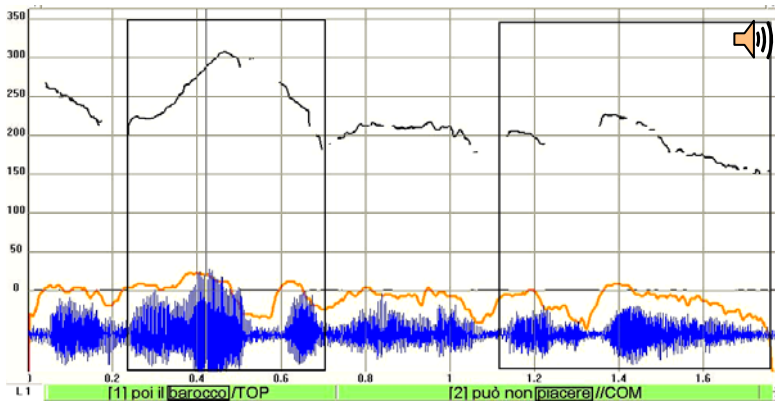
In all the above cases the Topic positively specifies the informational relation “reference domain for the illocutionary force accomplished by the Comment” and again, in parallel, in accordance with the conclusion of the previous paragraph, in this interpretation, it does not correspond to a compositional syntactic constituent.

However a certain number of NPs “resembling” Subjects have been found within a *prefix* unit (around 9%, according to Signorini, 2005). See the following examples where Subject-like NPs fill a Topic position:

- (20) *LIA: questa /TOP era a Firenze //COM [ifamcv01, 306]
 [(For what concerns) this / was at Florence]



- (21) *GAB: poi il barocco /TOP può non piacere //COM [ifamcv17, 27]
 [then (for what concerns) the baroque style / (somebody) cannot enjoy it]



In 20 and 21 the NP is performed through a *prefix* PU and fits with a Topic interpretation, i.e. it specifies the domain according to which the assertions, accomplished by the VP in Comment should be interpreted. However each of the previous NPs in Topic can be also considered a “head” assigning the inflection to the verbal form in Comment, so giving rise to a syntactic sentence structure and, from a semantic point of view, to a Subject-Predicate relation. So, is a syntactic sentence relation between the NP in Topic and the VP in Comment active? Is the NP in Topic the Subject of the VP in Comment?

While we might assume the positive informational value of Topic as evidence that no compositionality relation holds between their linguistic constituents, in the above cases it seems reasonable to assume that the NP is both a Topic and a Subject.

In the next paragraphs we will argue against this interpretation. Even if phrases occurring in Topic could be in principle considered part of the regency of a verb in Comment, they should not be interpreted according to this assumption in the spoken language performance. We will argue on the contrary, that an expression in Topic stays as an independent syntactic domain (*island*) and its relation with the expression in Comment is strictly informational rather than syntactic.

The assumption that a NP in Topic and in morphologic concordance with a Verb may not be in a direct syntactic relation with it, is not new. This assumption is widely shared in Generative Grammar (Rizzi, 1997; 2006). It is assumed that the NP in Topic is “kept apart” from the VP, while the Subject of the verb is an empty pronoun referring back to the NP in Topic. According to this conception, 21 will roughly receive the following structure:

(21') poi il barocco, /TOP 0, può non piacere //COM

However, from this perspective, the “apartness” of the NP in Topic does not imply it being part of one sentence configuration, since the Topic corresponds to a *functional category* that is part of the sentence structure. In this way the *utterance* should be

equivalent to a *sentence* with a large employ of specific functional categories, which are not identified as pragmatic concepts and entities but rather as syntactic entities specifying relations which concur to build up a proposition .

Our hypothesis is more radical than the proposed “apartness” of Topic. *Informational patterning Theory* assumes that the expression within one IU corresponds to one local syntactic domain. In accordance to this hypothesis the IUs within the informational pattern do not constitute as a whole a phrase or sentence, but a set of local syntactic configurations linked the one to the other by informational functions.

In this frame, the independence of the notion of Topic from the notion of Subject, and the consequent lack of compositionality, will be argued at two levels showing why a Subject cannot be a Topic and more intriguing, why a Topic cannot be a Subject.

3.3 Subject and Topic

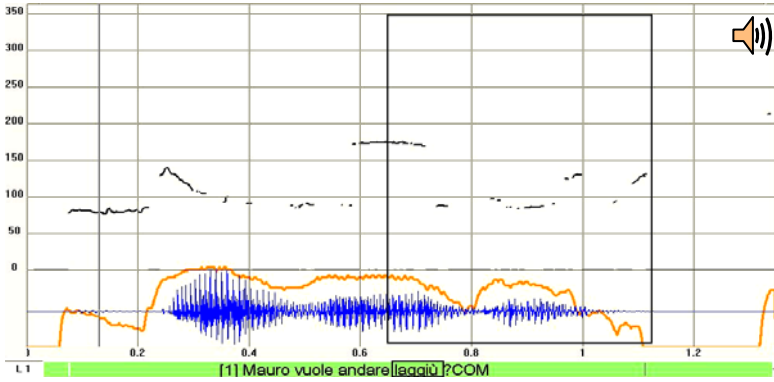
3.3.1 The demonstration of the pragmatic nature of information patterns and their independence from syntax implies various steps. One of these goes through the semantic distinction between the pragmatic function of Topic and the semantic role of Subject¹⁶.

In spoken language there are prosodic conditions to identify what a Subject is and what a Topic is. We have seen, that in order to be a Topic one expression must be performed through a *prefix* PU, while on the contrary, in order to be considered a Subject, it must be linearized in the same IU as its Predicate.

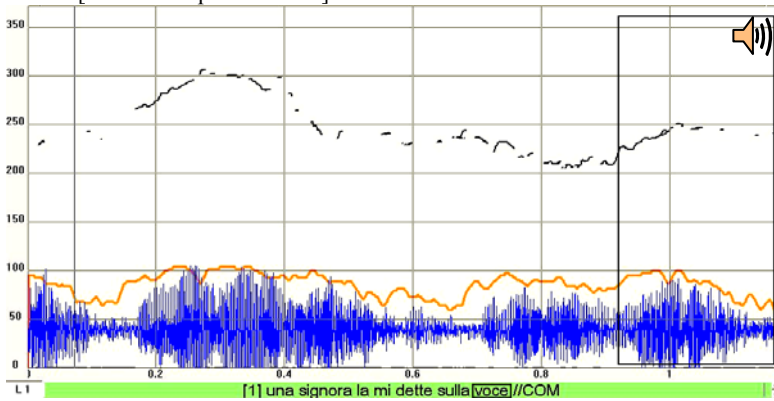
A corpus based research (Signorini 2005) has shown that nearly 40% of compound utterances (roughly corresponding to 20% of the total utterance) record a Topic. On the contrary, assuming the above prosodic constraint, less than 10% of all utterances record a NP linearized within the same IU before or after a VP which constitutes its Predicate. Therefore, according to this research, the informational strategy (Topic-Comment) appears broadly preferred in the construction of spoken texts than the semantic one (Subject-Predicate).

Considering the set of linearized Subjects, more than a half record lexical thematic NP or PP linearized in the same IU before a verb, as in 22 and 23:

- (22) *FRA: Mauro vuole andare laggiù ?COM [ifamcv06, 8]
 [(Is it confirmed that) Mauro wants to go there?]

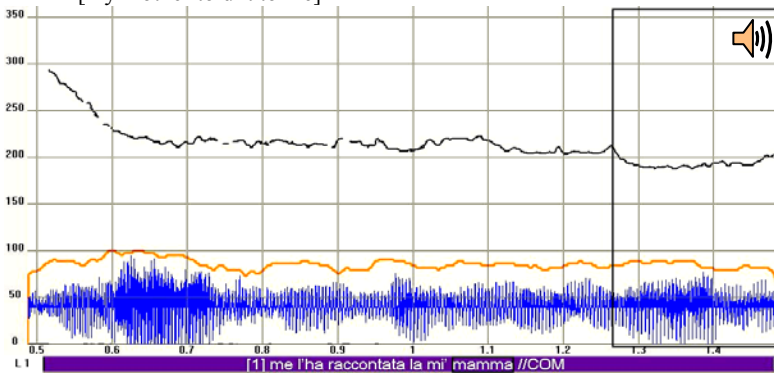


(23) *LUC: [...] una signora la mi dette sulla voce //COM [LAB-Fam1]
[a woman reproached me]

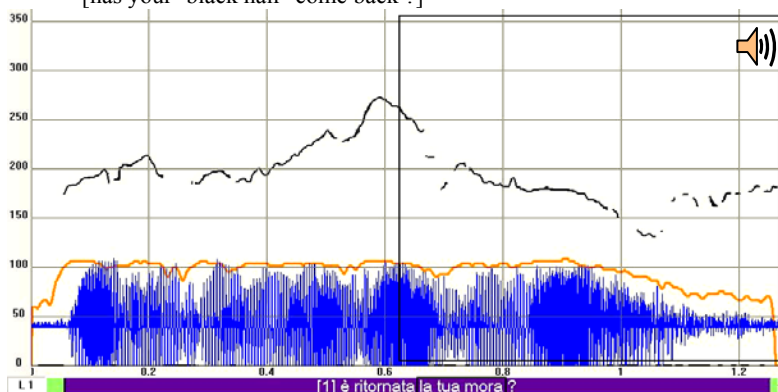


The remaining NPs linearized in the same IU, appear after the Verb, as in 24 and 25:

(24) *MIR: [...] me l'ha raccontata la mi' mamma //COM [LAB-Mir]
[my mother told it to me]



(25) *PAO: è ritornata la tua mora ?COM [LAB-Lili]
 [has your 'black hair' come back ?]



In all above examples, regardless of its position, the NP is linearized within a *root* PU. In other words the NP is co-articulated within the PU and does not give rise to any prosodic reset. The *root* PU performs a Comment IU, developing different illocutionary forces (request of confirmation, narration, question).

Comparing 22-25 with 21 (*poi il barocco / può non piacere //*), it is easy to recall the differential features characterizing the prosody of the two types: the NP and the VP are distributed on two PUs (prefix/root) in 21 while in 22-25 the NP and the VP are performed within one root PU. Two primary stresses vs. one primary stress, two prosodic Nuclei vs. one Nucleus.

3.3.2 The prosodic distinction has a straightforward result on the interpretation of these utterances at the semantic-pragmatic level. Consider for instance 22. As a matter of fact, the interpretation of the relation between *Mauro* and its VP is not consistent with a Topic Comment relation, as it was previously defined. In no circumstance 22 means that *Mauro* is the “domain of relevance” of an illocutionary act of request of confirmation. Obviously a Subject is not the domain of relevance of an illocutionary act performed by the Comment, since it is the Subject of a predication that, as a whole, constitutes the Comment itself. *Mauro vuole andare laggiù* is a sentence from a syntactic point of view and a proposition accomplishing an act of request of confirmation.

In summary, the relation between Subject and Predicate in 22-25 is strictly compositional and it is not a pragmatic relation. If the NP is bound to the VP from a syntactic and semantic point of view (i.e. it is the Subject or the Theme in the argument structure of the verb) no Topic-Comment interpretation is allowed. Linearized NP are compositional, therefore if an expression is a Subject, according to the above prosodic conditions, then it is not a Topic.

Given that obviously a Subject cannot be a Topic, we must also consider that, conversely, if one expression positively conveying the information function of Topic cannot for this reason also play the role of Subject and that this relation cannot be interpreted in the frame of sentence compositionality.

This is a much more complex question. However, assuming the formal distinction between Topic and Subject on the basis of stress and prosodic features, corpus based investigations provide results that confirm their differential nature. To this end, in paragraph 4, we will observe the properties of Subject and Topic in Italian spontaneous speech corpora, at both competence and observational levels. We will consider three main correlations of the Subject / Topic distinction showing that the relation between a Topic and a Comment lacks essential properties of compositionality that on the contrary are necessary correlations of Subjects. These evidences regard three independent phenomena. Distributional properties and Cliticization properties distinguish a possible Topic from a possible Subject for reasons of semantics (in 4.1 and 4.2). Modal properties of Topic-Comment pattern distinguish the higher level of organization of semantic information in the utterance with respect to the proposition (4.3).

4. Topic Comment structure and compositionality: negative correlations

4.1 Semantic constraints on Topic and Subject

Corpus-based investigations show that the semantic and morpho-syntactic filling of Topic and Subject diverge and record only a limited intersection (Signorini, 2005). As examples 10-14 show, Topic units can be filled by *Clauses*, *VPs*, *Quality Adjectives*, *Adverbs*. These expressions can comply properly with a displacement function for the Comment, while obviously they cannot be the Subject of any Predicate. From a general point of view it can be said that the semantics of Topic is larger than that of Subject. However, focusing on those NPs in Topic that might be considered Subjects, like 20-21, corpus based investigation allowed the discovery of a peculiar distributional divergence with linearized Subjects. *Anaphoric personal Pronouns*, *Indefinite pronouns*, *Negative NP* and *indeterminate not-specific NP* have never been found in spoken corpora in Topic position, while on the contrary they can be found as linearized Subject, as in the following examples derived from Italian corpora:

- (26) *LUC: [...] una signora la mi dette sulla voce //COM [LAB-Fam1]
[a leady reproached me]

(27) *PRF: [...] qualcosa bisogna fare //COM [inatpe02, 47]
[something should be done]

(28) *ALE: [...] esso rimbalza su una superficie e si riflette //COM [inatco01, 144]
[it jumps over a surface and reflects himself]

Given the relative small dimension of our corpora, it could be objected that the above lexical filling in Topics could have occurred in larger corpora. However this datum is confirmed by laboratory experiments that have been done in order to elicit the performance of Topics filled by *Anaphoric personal Pronouns, Indefinite pronouns, Negative NP and indererminative NP*. Expert speakers were not able to perform them with “natural” prosody, as usually they do. If the laboratory utterance is finally performed¹⁷, it sounds odd to competence based judgment, while the corresponding linearized utterance is acceptable:

(29) * nessuno /TOP è partito //COM [Nobody / left]

(30) nessuno è partito. [Nobody left]

In summary, negative, *indefinite* and *anaphoric personal pronouns*, cannot work as Topic, as seen in 29, while they are fully acceptable if they are Subjects, as seen in 30. *Indeterminative* and *not specific* NP behave in the same manner. They hardly fill a Topic, but can be found in the role of Subject (as in 23 *a woman reproached me*).

This datum asks for an explanation. Our hypothesis is that the previous set of lexical types, gathering different morpho-syntactic categories, is on the contrary homogeneous for one semantic feature; i.e. they do not specify linguistic information allowing the *individuation* of the referred entity at the cognitive levels. The reference to one entity, which is proposed by the speaker as unidentified in the world, does not seem appropriate to comply with the function of Topic.

Under this interpretation we have a straightforward informational reason for the above phenomenon. Given that Topic must specify the field of application of the illocutionary force, being the necessary intermediary with the context, the lack of the *individuation* semantic character prevents these expressions to fill a Topic position.

In further detail, the expression working as a Topic must achieve at least one cognitive representation in order to perform its informational function. If the same expression is part of a syntactic configuration, as a Subject is, this informational condition does not apply, since a proposition, with specific true conditions, is compositionally derived from the Subject- Predicate relation. Therefore, if the Topic- Comment structure is to be interpreted as a propositional structure, following the rules of compositionality, the above constraints will not emerge and nothing will impede 29 from having the same correctness and interpretation as 30.

As a whole, the semantic divergence between Topic and Subject goes hand in hand with their different syntactic status. An expression behaving as Topic is a syntactic and semantic *island*, it must stand by itself, having not only a full meaning (not being a morpheme), but also performing this meaning with a point of view (quantification, interpretation, modality) that allows a deictic, referential or at least precise cognitive representation that is required by its informational function.

4.2 Cliticizations and phoric relations in Topic-Comment pattern

The semantic difference between Topic and Subject is grounded also on the distribution of cliticization and the possible phoric relations of clitics, which have been studied in a recent corpus-based investigation (Cresti 2009, and forthcoming).

It is well known that in Italian a NP in Topic, which should be evaluated as a Theme or Direct Object of a verb in Comment, cannot develop that relation and a clitic must fill the position in the Comment IU at its place.

- (31) il pane /TOP *ho già comprato //COM (l'ho già comprato)
 [(for what concerns) the bread / I have already bought it]

Examples like 31 are neither found in corpora nor licensed by competence. However, we also never retrieve in spoken corpora examples like 32, in which the Topic guests one PPs or NPs working as Indirect Object of a verb, despite the fact that in principle, competence might license them. These structures always show the corresponding cliticization in Comment as 33¹⁸.

- (32) (a) Gianni /TOP *ho telefonato //COM (gli ho telefonato)
 [(for what concerns) Gianni / I have already called (him)]

- (33) (a) Gianni /TOP gli ho telefonato //COM (gli ho telefonato)
 [(for what concerns) Gianni / I have already called him]

In other words the distribution of clitics makes evident that in spoken Italian, when a verb occurs in Comment, its argument structure must be saturated within the same IU by a lexical item or by a clitic and cannot be saturated by a lexical NP or PP in Topic. The reason for this syntactic requirement may be found at the level of compositional semantics. Topics do not take part in the same syntactic configuration as the VP in Comment and do not saturate the argument structure of the verb. Consequently the expressions in Topics and in Comments result as independent phrases, not compositionally settled. From a semantic point of view, expressions in Topics turn out again semantically isolated.

However the discrepancy of corpus data and competence based judgments may be even more intriguing. Spoken language corpora, in connection with their informational patterning properties, show peculiar property of clitic placement, if compared to competence based grammar (Reinhardt, 1983).

NPs or clauses working as anaphoric heads in Topic with their clitic in Comment are frequent in spoken Italian corpora. See examples 34 and 34:

- (34) *UO1: e quando *un uomo politico*_i si commuove /TOP 0_i è un cretino //COM
[LAB-Gara1]
[and when a politician shows himself emotional / he is an idiot]

- (35) *MIR: *soldi e cellulare*_i /TOP se *li*_i può pure tene' /COM se è stata lei //PAR
[itelpv13, 329]
[money and mobile / she can even keep them / if she did it //]

But, contrary to our expectations, no cataphora have been observed between a clitic in Topic and a corresponding head NP in Comment. Obviously these types of cataphora are 'competence licensed', as in 36 and 37:

- (36) Quando *l*_i'ho guardato /TOP *Mario*_i ha voltato la testa //COM
[When (I) looked at *him*_i / *Mario*_i turned his head]

- (37) Quando 0_i è partito /TOP *Mario*_i ha ringraziato tutti //COM
[When 0_i left / *Mario*_i thanked everybody]

On the contrary the occurrence of cataphoric relation has been recorded within literary corpora (although they are rare and occur by preference in *when* sentences¹⁹):

- (38) Che cosa insegnasse all'allievo non lo so, ma quando 0_i è arrivato a Parigi Roberto_i faceva la sua figura [...] (Eco, *L'isola del giorno dopo*)
[What (he_i) were teaching to his pupil I don't know, but when (he_i) arrived in Paris Roberto_i cut a fine figure]

The structure of information in writing and in speech seems to be quite different. In speech the information pattern of an utterance is characterized by a pragmatic value that is absent in writing. According to its overall definition, when a Topic is performed, it provides a link with the pragmatic context to the Comment. As we saw in the previous paragraph the informational role of Topic requires an identified referential expression. Given that the reference domain for the Comment relies on the identification of a Topic, holding a cataphora in Topic seems to be contradictory from an informational point of view, since the clitic should find its identification in the Comment itself.

This explains the differential distribution of clitics in written and spoken language given that only the latter is governed by the informational structure conveyed by prosody.

More specifically, the correlation between the informational function of Topic and the lack of cataphora in Topic in speech performance is an independent argument in favor of the hypothesis that Topic is external to the syntactic configuration of the sentence. Topic behaves as a pragmatically isolated component of the utterance, in accordance with the pragmatic function conveyed by prosody. Due to pragmatic reasons, the barriers that split it from the Comment impede the onset of some semantic relations, like cataphora, with elements in the syntactic configuration that are in principle licensed by the grammar.

4.3 Modality and Topic-Comment information pattern

The syntactic and semantic independence of the expressions in Topic from those in Comment is confirmed by one other property, which has emerged from corpus-based investigations concerning modality in speech. According to a traditional ballyan view (Bally 1932; 1942) Modality is defined as the *subjective evaluation of the semantic content* by the speaker; i.e. “Modus on Dictum”.

Following this tradition, Modality is defined in our frame as *the attitude of the speaker toward the locution*. Therefore it must be clearly underlined that this kind of attitude regards the locutive act and has nothing to do with pragmatics²⁰. Modality is the higher semantic level within the locutive act and in accordance with the logical tradition, it is *a property of a proposition*. When a proposition is in the scope of more than one modal index the modal indexes are compositional, and the proposition receives the modality of the higher modal index (Huges and Creswell 1996). For instance:

(39) “A professor of linguistics *must*₁ *be able*₂ to teach linguistics”
is a deontic proposition (deontic- index₁ over alethic index₂)

(40) “A professor of linguistics *may*₁ *have been forced*₂ to teach philosophy”
is an alethic proposition (alethic-index₁ over deontic-index₂)

This compositional principle is strictly followed in writing, but not in speech. A systematic research on lexical indexes of modality (Tucci 2007) has pointed out that in spoken language the use of explicit modal lexicon is more common than believed (nearly 14% of total utterances²¹). However the most important result of this research is the discovery that if a compound utterance is multimodalized (modal indexes placed in different IUs) each of these IUs keeps its own modality, which is not compositionally solved. In other words, the scope of modality is not the entire

utterance, as in the sentence, but rather it regards the location of each textual information unit. This is a very common feature in spontaneous speech. Tucci found that 85% of utterances with more than one modal index (multi-modalized) host each index in a different IU (Tucci 2009).

The interpretation of Modality is among strongest semantic evidence supporting the proposal that, in spoken language, Topic is not a compositional part of the overall syntactic structure of a sentence. The reader can focus on evidence bootstrapped from corpora in which modal values placed in different IUs, and specifically in Topic and Comment, are not compositionally settled. Let see 41 and 42:

(41) *MAX: *secondo me* /TOP ne dimostrava di più //COM [ifamcv01, 191]
[in my opinion / she looked older]

(42) *una soluzione probabile* /TOP è che *devi* pagare tutto //COM
[(regarding) the most probable solution / it is that you must pay everything]

In 41 and 42, each IU of the Topic-Comment information pattern is characterized by a different modal value. 41 is a very common case in spoken language, where Topic holds an epistemic expression (*in my opinion*) and Comment has an alethic interpretation. Its illocutionary force is roughly equivalent to the performative *to claim*. In 42, the Topic still records an epistemic modality while the Comment bears a deontic one and its illocution seems rather an advice. Therefore both utterances are multi-modalized.

If the two modal values in 41 had to be compositionally bound according to propositional rules, a resulting paraphrase could be *I subjectively evaluate that I claim that she looked older*, with an epistemic general dominance, but it makes no sense. For what concerns 42, the paraphrase should be *It is probable that the advice that you must pay everything will be the solution* with an epistemic general evaluation, but this periphrasis is not coherent with the actual interpretation of the utterance.

Both the previous paraphrases force the Topic-Comment pattern in one proposition, but as a consequence of this, they show a meaningless or misleading modal value with respect to the actual one. The acceptable paraphrase for 41 could be *I claim that she looked older, but it is my present evaluation*. For 42 it can be *I suggest that you must pay everything and this seems to be a probable solution*. Both acceptable paraphrases are compound by two coordinative modalized clauses and therefore each syntactic entity maintains its own modality. The two modal indexes are not compositionally settled.

If modality is the higher level of a syntactic/semantic structure (a property of a proposition), given that Topic and Comment show independent modality that cannot be compositionally solved, they cannot be parts of the same syntactic domain and they do not reflect together the form of a proposition. In other words,

compositionality requires the synthesis of modal values within a proposition. The fact that modal values stay within the IUs hosting their indexes is a direct evidence that Topic-Comment informational patterning is not consistent with compositionality.

In patterned spoken utterances modality is a property of the semantic content of each IU that from this point of view behaves like one island. Again, the relation between IUs follows informational principles rather than a compositional semantic one.

5. Some general conclusions regarding compositionality and Information patterns

In this paper various independent evidence bootstrapped from Italian corpora has shown that the performance of two strings in a Topic-Comment pattern determine the onset of two local syntactic and semantic domains which are not compositionality bound; i.e. the information pattern is neither equivalent to a syntactic configuration nor to a proposition. On the contrary, expressions that are linearized within one IU strictly follow compositionality rules and, by converse, their relation cannot be bound within an information pattern.

While the following syntactic and semantic relations hold within linearized constituents they cannot involve expressions in a Topic Comment pattern:

- Modification (NP: Noun head- Modification)
- Regency (VP: Verb – Direct and Indirect Object)
- Predication (Sentence: Subject-Predicate)
- Modalization (Proposition: Compositionality of modal indexes)

The non-terminal prosodic break between a *prefix* PU and a *root* PU represents a barrier for syntax and semantics. This is due to the nature of the Topic Comment informational relation, which identifies the pragmatic domain of relevance for the illocutionary force, displaced from the context. To displace an act is not equivalent to a predication and does not determine the onset of a propositional structure.

Notes

¹ E. Cresti managed the corpus based research and selected the relevant arguments, M. Moneglia designed the ratio of the empirical demonstration.

² The framework of informational patterning is being applied also to other romance languages such as Brazilian Portuguese (Raso et al. 2007, Raso and Ulisses 2008) and Spanish (Nicolas 2006).

³ For detailed studies on the expression of the illocutionary force in spontaneous speech see Firenzuoli (2003), Cresti, Moneglia and Martin (2003), Cresti (2006).

⁴ For a different point of view on the same domain of facts see Roulet (2002).

⁵ In spontaneous speech, prosodic patterns do not correspond to information patterns in two major cases: a) retracting and fragmentation phenomena; b) scanning phenomena; i.e parsing of the locution in separate prosodic envelopes with no perceptively relevant prosodic movements (Cresti 2000; Cresti and Moneglia, in press).

⁶ Prosodic types are defined in accordance to the IPO's approach ('t Hart et al. 1990) which is based on the perceptively relevance of prosodic cues.

⁷ For the prosodic cues of various IUs see respectively Cresti (2000), Cresti and Firenzuoli (2000), Cresti and Firenzuoli (2002), Firenzuoli and Signorini (2003), Firenzuoli and Tucci (2003), Frosali (2008).

⁸ All examples in this paper come from the Italian section of the C-ORAL-ROM corpus (Cresti and Moneglia 2005) and from the LABLITA Corpus of Spontaneous Spoken Italian. Transcriptions are orthographic and follow a variant of the CHAT format (Mac Whinney, 1991) in which prosodic breaks are marked (double slashes “//” marks terminal and single slash “/” marks non terminal). Examples are followed by the filename and by the ranking number of the utterance in the corpus. Tags for IU types are in capitals aside prosodic breaks, as indexes for each information unit. The English translation is in square brackets and tentatively try to save the original informational structure of the original example. For this reason some phrasing may sound odd to the English reader. Figures under the examples identify F₀, Intensity and Timing of the wave. The acoustic signal is aligned to the transcription on bottom. The transcription is parsed into information units whose limits correspond to their start and end points on the wave. The perceptively relevant prosodic movement of each PU is highlighted in a rectangle. The corresponding phrase in the IU on bottom is also highlighted.

⁹ In Italian nearly 80% of Topic correlates with tree allomorphic types of the *prefix* Nucleus that are represented in accordance with the IPO system and instantiated in the figures of this paper: Type “[1] [A]”: *raising-falling*, see example 7, 16, 19, 20, 21 below ; Type “[1]”: *raising* see example 2; Type “[A] [0] [1]” *falling-flat-raising*, see example 6 below. See Firenzuoli and Signorini (2003) for a more detailed description.

¹⁰ See Cresti (2000) and Cresti and Moneglia (in press)

¹¹ The interpretation of the above facts in terms of ellipsis of predicative elements, is frequently considered in the linguistics literature (According to this point of view 7 should be mapped onto the same structure of the sentence “*The name of my grandfather is Peter*” and 6 onto “*Apples should have the shape of swans*”. However ellipsis requires ad hoc solutions that are not empirically motivated. For instance the need for ellipsis is still determined by prosodic cues, that is strictly required in order to assign an interpretation to the above examples. Moreover and for this reason it opens more problems than it solves (Scarano 2004). We will not discuss this topic here.

¹² With no need of any ad hoc elliptical structure.

¹³ The sampling is a sub-corpus of the LABLITA spoken corpus annotated with illocutionary tags (Firenzuoli 2003).

¹⁴ In 15 and 16 the informational patterning presents two Topic IUs, as it is frequently the case in spontaneous spoken Italian. This does not change the argument.

¹⁵ A possible compositional paraphrase is available when a performative verb expressed through lexical means what the Comment expresses through prosody, as it is the case in 15 (*I advise you to go walking around for fun this night*). However this paraphrase strongly changes the form of the utterance, which is no more a Topic-Comment pattern and, accordingly, it does not specify any aboutness relation for the illocutionary act. Moreover, in spontaneous speech the illocutionary act expressed by the Comment frequently does not find any equivalent performative verb in the dictionary. For instance 16 cannot receive a performative paraphrase (*I give to you the instruction that you must give your OK once all codes are inserted* does not perform an act of instruction).

¹⁶ See Li (1976) and Rizzi (2006) for an overall discussion of the Subject-Topic distinction.

¹⁷ For instance this has been achieved in 29 considering “Nobody” as a proper name (alias Ulysses).

¹⁸ Grammars claim this is possible for Italian, however we cannot be very confident on the prosodic requirements, that are not clearly stated. We refer here specifically to a *prefix-root* pattern.

¹⁹ 51% of phoric relations regards a Clitic in Topic or Comment with an anaphoric relation with a NP or a clause in a Focus position of a previous utterance. No cataphora among utterances have been retrieved (Cresti 2009).

²⁰ Modality and illocution must be clearly distinguished. Illocution performs the conventional communicative activity of the speech act, and belongs to pragmatics. Modality specifies the attitude of the speaker toward the locution and belongs to semantics. The independence of the two notions in spoken language have been formally demonstrated (Cresti 2002; Tucci 2008; Tucci and Moneglia in press).

²¹ On 37.289 utterances, 5.152 are lexically modalized (Tucci 2007).

References

- Austin, J.L. 1962. *How to do things with words*. Oxford: Oxford University Press.
- Bally, Ch. 1932. *Linguistique générale et linguistique française*. Berna: Francke Verlag.
- Bally, Ch. 1942. Syntaxe de la modalité explicite. *Cahiers Ferdinand de Saussure* 2: 3-13.
- Blanche-Benveniste, C. (ed.) 1991. *Le français parlé. Études grammaticales*. Paris: Editions du CNRS.
- Blanche-Benveniste, C. 1997. *Approches de la langue parlée en français*. Paris: Ophrys.
- Chomsky, N. 1971. Deep Structure, Surface Structure, and Semantic Interpretation. In D. Steinberg and L. Jakobovitz (eds), *Semantics*. Cambridge: Cambridge University Press, 183-216.

- Cordin, P. and A. Calabrese. 1998. Il pronome. In L. Renzi, G.P. Salvi and A. Cardinaletti (eds), *Grande Grammatica di Consultazione. Vol. 1. La frase. I sintagmi nominale e preposizionale*. Bologna: Il Mulino, 535-594.
- Cresti, E. 1987. L'articolazione dell'informazione nel parlato. In AA.VV., *Gli italiani parlati*. Firenze: Accademia della Crusca, 27-90.
- Cresti, E. 1994. Information and intonational patterning. In Ph. Martin, B. G. Ferguson and H. Gezundhajt (eds), *Accent, intonation and modèles phonologiques*. Toronto: Edition Mélo die, 99-140.
- Cresti, E. 2000. *Corpus di italiano parlato*, 2 voll., CD-ROM. Firenze: Accademia della Crusca.
- Cresti, E. 2000a. Critère illocutoire et articulation informationnelle. In M. Bilger (ed.), *Corpus. Méthodologie et applications linguistique*. Paris: Champion, 350-367.
- Cresti, E. 2002b. Illocuzione e modalità. In P. Beccaria and C. Mare llo (eds), *La parola al testo. Scritti per Bice Mortara-Garavelli*. Torino: Ed. Dell'Orso, 133-145.
- Cresti, E. 2006. Some comparison Between UBLI and C-ORAL-ROM. In Y. Kawaguchi, S. Zaima and T. Takagaki (eds), *Spoken Language Corpus and Linguistics Informatics*. Amsterdam: Benjamins, 125-152.
- Cresti, E. 2009. Clitics and anaphoric relations in informational patterning: a corpus driven research in spontaneous spoken italian (C-ORAL-ROM). In L. Mereu (ed.), *Information structures and its interfaces*. Berlin-New-York: Mouton de Gruyter, 171-203.
- Cresti, E. and V. Firenzuoli. 1999. Illocution et profils intonatifs de l'italien. *Revue française de linguistique appliquée*, 4, 2: 77-98.
- Cresti, E. and V. Firenzuoli. 2002. L'articolazione informativa topic-comment e comment-appendice: correlati intonativi. In A. Regnicoli (ed.), *La fonetica acustica come strumento di analisi della variazione linguistica in Italia* (Atti delle XII giornate del GFS). Roma: Editrice Il Calamo, 153-16.
- Cresti, E. and M. Moneglia (eds). 2005. *C-ORAL-ROM. Integrated reference corpora for spoken romance languages*, DVD + vol. Amsterdam: Benjamins.
- Cresti, E. and M. Moneglia. In press. *Specifications for the annotation of the informational patterning in spontaneous spoken Italian*. Firenze: Firenze University Press.
- Cresti, E., Ph. Martin and M. Moneglia. 2003. L'intonation des illocutions naturelles représentatives: analyse et validation perceptive. In A. Scarano (ed.), *Macrosyntaxe et pragmatique: L'analyse linguistique de l'oral*. Roma: Bulzoni, 243-26.
- C-ORAL-ROM. <http://lablita.dit.unifi.it/coralrom/>
- Firenzuoli, V. 2003. *Le forme intonative di valore illocutivo dell'italiano parlato: analisi sperimentale di un Corpus di parlato spontaneo (LABLITA)*. PhD diss., University of Florence.
- Firenzuoli, V. and S. Signorini. 2003. L'unità informativa di topic: correlati intonativi. In G. Marotta and N. Nocchi (eds), *La coarticolazione* (Atti delle XIII Giornate del GFS). Pisa: ETS, 177-184.
- Firenzuoli, V. and I. Tucci. 2003. L'unità informativa di inciso: correlati intonativi. In G. Marotta and N. Nocchi (eds), *La coarticolazione* (Atti delle XIII Giornate del GFS). Pisa: ETS, 185-192.
- Frosali, F. 2008. Il Lessico degli ausili dialogici. In E. Cresti (ed), *Prospettive nello studio del lessico italiano* (Atti del IX Congresso della SILFI). Firenze: FUP, 417-424.
- Halliday, M.A.K. 1967. *Intonation and grammar in British English*. The Hague: Mouton.

- Halliday, M.A.K. 2004. *An introduction to functional grammar*. London: Edward Arnold.
- 't Hart, J., R. Collier and A. Cohen. 1990. *A perceptual study on intonation*. Cambridge: Cambridge University Press.
- Hockett, F. 1958. *A Course in Modern Linguistics*. New York: The Macmillan Company.
- Huges, G. E. and M. J. Cresswell. 1996. *A new introduction to modal logic*, London: Routledge.
- LABLITA. <http://lablita.dit.unifi.it/>
- Lambrecht, K. 1994. *Information structure and sentence form*. Cambridge: Cambridge University Press
- Li, C. (ed.) 1976. *Subject and Topic*. New York: Academic Press.
- Mathesius, V. 1929. La linguistica funzionale. In R. Sornicola and A. Svoboda (eds, 1991), *Il campo di tensione. La sintassi della scuola di Praga*. Napoli: Liguori, 97-112.
- Mac Whinney, B. 1991. *The CHILDES project. Tools for analyzing talk*. New York: Hillsdale.
- Moneglia, M. 1994. The ontogenetic foundation of informational patterning. In B. Ferguson, H. Gezundhajt and Ph. Martin (eds), *Accent, intonation et modèles phonologiques*. Toronto: Editions Mélodie, 155-186.
- Moneglia, M. and E. Cresti. 2006. C-ORAL-ROM. Prosodic boundaries for spontaneous speech analysis. In Y. Kawaguchi, S. Zaima and T. Takagaki (eds), *Spoken Language Corpus and Linguistics Informatics*. Amsterdam: Benjamins, 89-114.
- Nicolás-Martínez, C. 2007. Primeros resultados del estudio de nada, niente y nulla en C-ORAL-ROM. In C. Nicolás Martínez (ed.), *Ricerche sul Corpus del parlato romanzo C-ORAL-ROM*. Firenze: Firenze University Press, 49-66.
- Raso, T., H. Mello, H.L. Deus and A. Jesus. 2007. Uma aplicação da Teoria da Língua em Ato ao PB. *Revista de Estudos da Linguagem* 2: 147-166.
- Raso, T. and A. Ulisses. 2008. Tópico e Apêndice no português do Brasil: algumas considerações. *Revista de estudos da linguagem* 1: 45-60.
- Reinhardt, T. 1983. Coreference and bound anaphora: a restatement of the anaphora questions. *Linguistics and Philosophy* 6, 1: 47-88.
- Rizzi, L. 1997. The fine structure of the left periphery. In L. Haegeman (ed.), *Elements of grammar: A handbook of generative syntax*. Dordrecht: Kluwer, 281-337.
- Rizzi, L. 2006. On some properties of subjects and topics. In L. Brugé, G. Giusti, N. Munaro, W. Schweikert, and G. Turano (eds), *Proceedings of the XXX Incontro di Grammatica Generativa*. Venezia: Cafoscarina, 203-224.
- Roulet, E. 2002. Le problème de la définition des unités à la frontière entre le syntaxique et le textuel. *Verbum* 24, 1-2: 161-178.
- Scarano, A. 2003. Les constructions de syntaxe segmentée: syntaxe, macro-syntaxe et articulation de l'information. In A. Scarano (ed.), *Macrosyntaxe et pragmatique: L'analyse linguistique de l'oral*. Roma: Bulzoni, 183- 203.
- Scarano, A. 2004. Enunciati nominali in un corpus di parlato. Appunti per una grammatica corpus based. In F. Albano Leoni, F. Cutugno, M. Pettorino and R. Savy (eds), *Atti del convegno nazionale "Il parlato italiano"* (CD-ROM). Napoli: M. D'Auria Editore, 1-18.
- Scarano, A. 2009. A The prosodic annotation of C-ORAL-ROM and the structure of information in spoken language. In L. Mereu (ed.), *Information structures and its interfaces*. Berlin and New York: Mouton de Gruyter, 51-74.

- Signorini, S. 2004a. L'unità di topic. Caratteristiche e frequenza in un corpus di italiano parlato. Il topic complesso. In P. D'Achille (ed.), *Generi, architetture e forme testuali* (Atti del VII convegno internazionale SILFI). Firenze: Franco Cesati, 227-238.
- Signorini, S. 2004b. Il Topic: criteri di identificazione e correlati morfosintattici in un corpus di italiano parlato. In F. Albano Leoni, F. Cutugno, M. Pettorino and R. Savy (eds), *Atti del convegno nazionale "Il parlato italiano"* (CD-ROM). Napoli: M. D'Auria Editore, 1-24.
- Signorini, S. 2005. Topic e soggetto in corpora di italiano parlato spontaneo. PhD diss., University of Florence.
- Sperber, D. and D. Wilson. 1986. *Relevance: communication and cognition*. Oxford: Basil Blackwell.
- Tucci, I. 2007. La modalizzazione lessicale nel parlato spontaneo. Dati dal corpus C-ORAL-ROM Italiano. PhD diss., University of Florence.
- Tucci, I. 2008. Lessico della modalità e illocuzione in un corpus di italiano parlato (C-ORAL-ROM). In E. Cresti (ed.), *Prospettive nello studio del lessico italiano* (Atti del IX Congresso della SILFI). Firenze: Firenze University Press, 471-478.
- Tucci, I. 2009. The scope of lexical modality and the informational structure in spoken Italian". In L. Mereu (ed.), *Information structures and its interfaces*. Berlin and New York: Mouton de Gruyter, 203-226.
- Tucci, I. and M. Moneglia. In press. Modality and Illocutionary force in spoken Italian. In C. Push (ed.), *Corpora and Pragmatics* (Proceedings of the 3rd Freiburg Workshop on Romance Corpus Linguistics). Tübingen: Gunter Narr Verlag.
- Weil, H. 1844. De l'ordre des mots dans les langues anciennes comparées aux langues modernes. In *The order of words in the ancient languages compared with that of the modern languages translation*, by C.W. Super (1978). Amsterdam: Benjamins.

LANGUAGE-TEXT INTERFACE: THE EXAMPLE OF THEMATIC PROGRESSION*

Angela Ferrari, Anna Maria De Cesare
University of Basel

1. Introduction

The concept of “thematic progression” has been developed by František Daneš (cf. in particular Daneš 1970, 1974) to define an important textual (macro-syntactic) effect of the sentence-related phenomenon known as “Functional Sentence Perspective”, identified and defined by the Prague School around 1930. In this paper, the concept of thematic progression has been chosen both to present a model of the semantic-pragmatic organization of the written text developed at the University of Basel and to illustrate its heuristic power¹. This model pays particular attention to the role played by the linguistic component in text organization and is based on two main hypotheses: 1. the semantic component of lexical items, syntax and punctuation codes textual information; 2. the communicative actualization of such information is mediated by the informational articulation of the Utterance. The informational articulation of the Utterance therefore functions as the interface between two systems of organization of verbal communication that are governed by very different principles: the linguistic system and the textual system.

In this paper we try to show that the model of text organization we have developed allows for a more rigorous definition of “thematic progression”. Thematic progression, as we know, has to do with textual organization, and in particular with the selection of the successive Topics in the text (whenever possible we will therefore use the more appropriate term of “Topic progression”). Since, however, the organization of texts is based on a modular system, the selection of the Topics in the message interacts with other dimensions of text organization. More specifically (as we will illustrate in more detail below), Topic progression interacts with what we call “the hierarchical-illocutionary organization of the Utterance”: choosing between the Topics available within each Semantic Proposition, the hierarchical-illocutionary level of text structuring selects the Topics that are most significant in the development of the message.

Our paper is organized as follows. After a brief section in which we recall the main aspects of the concept of “thematic progression” (paragraph 2), we will discuss some limitations of this notion and show that, within the informational structuring of the Utterance content, a distinction should be made between the level related to the concept of “thematicity” (in the sense of *aboutness*) on the one hand, and the level related to the concept of Communicative Dynamism on the other (paragraph 3). The informational organization of the Utterance is composed of (at least) two levels: a. the level related to the notion of *aboutness*, based on the informational functions of Topic and Comment; and b. the hierarchical-illocutionary level, organized according to the illocutionary and textual functions performed by the Utterance in the message. In order to address their complexity better, the two levels mentioned will be described separately (paragraph 4). The last part of this paper will address Topic organization in the (written) text according to the model we have developed (paragraph 5). In this section we will show how the interplay between the two informational levels described above – the Topic-Comment-related level on the one hand, and the hierarchical-illocutionary level on the other – defines Topic progression in the paragraph (or in a subpart of it). The hierarchical-illocutionary level is responsible for choosing which Topic is to function as the Utterance Topic; further, it defines whether the Utterance Topic also functions as a macro-Topic, i.e. whether it coincides with the Topic of a group of Utterances or of a whole paragraph.

2. The concept of “thematic progression”

2.1 As we know, the concept of “thematic progression” is based on the notions of *Theme* and *Rheme* (cf. *infra*). The Theme of the Utterance is, according to Firbas 1974, the semantic element associated with the lowest degree of Communicative Dynamism, while the Rheme is the element with the highest degree of Communicative Dynamism. From an informational perspective, it is therefore the sequence of rhematic elements in the text that ensures its development, whereas thematic elements participate in creating text cohesion, and therefore semantic continuity. Thematic progression – i.e. the successive choices of the Theme within each Utterance and the concatenation of such Themes – defines the “semantic skeleton” of the text, i.e. the supporting framework for its stability. In Daneš’s words (1974, 114):

By this term [thematic progression] we mean the choice and ordering of utterances themes, their mutual concatenation and hierarchy, as well as their relationship to the hyperthemes of the superior text units (such as the paragraph, chapter, etc.), to the whole text, and to the situation. Thematic progression might be viewed as the skeleton of the plot.

2.2 As soon as it was defined, the concept of thematic progression was immediately used in Text Linguistics, where it came to denote an important condition of text coherence (Combettes 1988). Utterances forming a coherent text are typically characterized by the presence of thematic contents with a low degree of Communicative Dynamism, and these thematic elements are responsible for linking more dynamic information to the immediate and less immediate contexts (the phenomenon involves the preceding as well as the succeeding context, cf. Givón 1983). More specifically, in a coherent text, the Utterance Theme relates to the preceding context either directly – by reintroducing information previously given – or indirectly – through semantic or contextual inference. As Mortara Garavelli 1979 suggests in the title “La continuità del discorso: la struttura tematica” (*Discourse continuity: thematic structure*) (p. 93), the thematic structure of the text becomes a symptom of its continuity, which is in turn an important ingredient of textuality. To understand this idea better, let us look at the following example (note that here and in the following examples the elements we are discussing are shown in boldface; for the purposes of our analysis, we will also occasionally number each Utterance of the text we are looking at):

- (1) 1. Come in un sogno stava d’innanzi a noi la casa. 2. **Su la facciata rustica, per tutte le cornici, per tutte le sporgenze, lungo il gocciolatoio, sopra gli architravi, sotto i davanzali delle finestre, sotto le lastre dei balconi, tra le mensole, tra le bugne, dovunque le rondini** avevano nidificato. 3. **I nidi di creta** innumerevoli, vecchi e nuovi, agglomerati come le cellette di un alveare lasciavano pochi intervalli liberi. 4. **Su quegli intervalli e su le stecche delle persiane e su i ferri delle ringhiere gli escrementi** biancheggiavano come spruzzi di calcina. 5. **Benché chiusa e disabitata, la casa** viveva. 6. [sogg. nullo = la casa] Viveva d’una vita irrequieta, allegra e tenera. 7. **Le rondini fedeli** l’avvolgevano dei loro voli, dei loro gridi, dei loro luccichii, di tutte le loro grazie e di tutte le loro tenerezze, senza posa. 8. **Mentre gli stormi s’inseguivano per l’aria in caccia con la velocità delle saette, alternando i clamori, allontanandosi e riavvicinandosi in un attimo, radendo gli alberi, levandosi nel sole, gittando a tratti dalle macchie bianche un baleno, instancabili,** ferveva dentro ai nidi e intorno ai nidi un’altra opera. (D’Annunzio, *L’innocente*, pp. 155-156, in Ferrari 1994, 58.).

The information marked in bold in the example helps to link the non-thematic (central) elements of the Utterance (not in bold face in our example) to the preceding context. For instance, the omitted subject of Utterance 6 ensures semantic continuity with the preceding Utterance (as well as with the initial Utterance) through anaphoric resumption of the nominal phrase *la casa*. Semantic continuity is also insured by the extensive sequence of text appearing in bold face in Utterance 8: this sequence links the rhematic information of the Utterance, i.e. *ferveva dentro ai nidi*

e intorno ai nidi un'altra opera, to the preceding context, in particular to Utterance 7. More specifically, what has been said in Utterance 7 is resumed, through linguistic variation, at the beginning of Utterance 8: *le rondini fedeli* → *gli stormi*; *l'avvolgevano dei loro voli* → *s'inseguivano per l'aria in caccia con la velocità delle saette ... allontanandosi e riavvicinandosi in un attimo*; *dei loro gridi* → *alternando i clamori*; *dei loro luccichii* → *gittando a tratti dalle macchie bianche un baleno*; *senza posa* → *instancabili*.

2.3 There are different types of thematic progression, which in a text typically alternate, cross each other, and overlap. The typology proposed by Daneš 1974 – and adopted without any substantial variation for instance by Combettes 1988 – has been slightly changed under the influence of the first contributions of Mortara Garavelli (particularly in the light of her 1979 work). Basically, according to the nature of the preceding informational unit being thematized (i.e. whether it is a thematic and/or a rhematic element), three types of thematic progression can be distinguished (cf. Ferrari and Zampese 2000):

1. “constant Theme progression”: a preceding Theme (or subpart of it) is thematized
2. “linear Theme progression”: a preceding Rheme (or subpart of it) is thematized
3. “thematization of a Theme+Rheme sequence”, or “thematization of more than one Theme+Rheme sequence”.

In order to illustrate the three different thematic progressions, let us first look at the following example:

- (2) **1. Il microscopio** permette di osservare oggetti molto piccoli, tanto piccoli da non essere visibili ad occhio nudo. **2. (sogg. nullo)** È uno strumento che funziona così: **3. un primo gruppo di lenti, l'obbiettivo**, ingrandisce l'oggetto da vedere; **4. un secondo gruppo di lenti, l'oculare**, ingrandisce l'immagine creata dall'obbiettivo. **5. Il segreto del microscopio** è dunque l'ingrandimento dell'ingrandimento. **6. Con il microscopio, (sogg. nullo)** possiamo osservare le parti delle piante. **7. (sogg. nullo)** Iniziamo con una pelle di cipolla: **8. essa** infatti è sottile e si osserva con facilità. **9.** Vista attraverso il microscopio, **la pelle di cipolla** appare formata di piccoli 'mattoni' detti cellule. (in Ferrari and Zampese 2000, 354).

In the text reproduced in (2), Constant Theme progression can be observed between Utterances 1 and 2: through an omitted subject, the Theme of Utterance 2 resumes the nominal phrase *il microscopio*, which is the Theme of the preceding Utterance. Utterances 7 and 8 show an instance of Linear Theme Progression: the Theme of Utterance 8, *essa*, resumes the nominal phrase *una pelle di cipolla*, which is part of

the Rheme of Utterance 7. In turn, example (3) illustrates the thematic progression involving the thematization of a Theme+Rheme sequence: the pronoun *ciò* encapsulates the content of the entire preceding Utterance:

- (3) I raggi del sole che giungono sui monti sono più caldi dei raggi di sole che arrivano in pianura. **Ciò** è noto a tutti coloro che sono stati in montagna e che si sono scottati la pelle malgrado le temperature molto basse (in Ferrari/Zampese 2000, 345).

The Theme of a Proposition may also be cataphorically linked to the Theme of another Proposition belonging to a following Utterance (see, for example, Utterance 8 of text (2), where the Theme *essa* is linked to the Theme *la pelle di cipolla* of Utterance 9).

3. Some limitations of the Prague School conception of “thematic progression”

3.1 An attempt to apply the concept of thematic progression to texts showing a certain degree of complexity (typically pertaining to non-descriptive text types) immediately proves its inability to define the “communicative felicity” of written discourse. This limitation was observed from the very beginning by František Daneš himself, who proposed the concept of “thematic leap” to address the problem. This concept refers to the sequences of text which cannot be considered incoherent (that is, non-texts), even if there is no connection between the Themes of these sequences and their co-text. As we now know, and as has been shown in recent studies of Discourse Analysis (cf. for example Roulet, Filliettaz and Grobet 2001; Ferrari (ed.) 2004, Ferrari (ed.) 2005, Ferrari and De Cesare in press; Ferrari et al. 2008), the semantic-pragmatic structure of the text is defined by a modular system, i.e. by the interaction of a number of conceptually independent dimensions of text organization (in our research we account principally for two major dimensions of text organization: the logic-argumentative dimension and the thematic dimension). Nevertheless, there are sequences of texts in which one of these conceptually independent organizing dimensions – for instance, the logic-argumentative dimension – prevails over the others. In those cases, the thematic dimension of text organization becomes inactive or secondary, and can therefore present instances of “thematic voids”. This is, for instance, the case in the short text given in (4):

- (4) Piove. Non esco.

Although the sequence in (4) lacks a thematic link between the two Utterances, the text is perfectly interpretable. The unexpressed logical relation between the two

parts of the texts – Utterance 2 is a consequence of what is said in Utterance 1: *It rains. (Therefore) I will not go out* – is sufficient to produce textual coherence.

This very interesting phenomenon is not to be underestimated. It has, however, no impact on the validity and applicability of the concept of thematic progression as it was originally defined. It simply shows that the thematic dimension is but a single aspect of text structuring.

3.2 Let us now consider a more important limitation of the concept of thematic progression proposed by the Prague School. In her 1986 paper (which is part of the 1988 collection of essays referred to in the bibliography of the present study), Maria-Elisabeth Conte points to the overlapping of notions that should be kept separate:

Nell'ambito della Functional Sentence Perspective, tema e rema sono stati definiti sia sotto l'aspetto *tematico* ("*thematischer Aspekt*": tema è ciò su cui si comunica qualcosa; rema è ciò che sul tema si comunica), sia sotto l'aspetto *contestuale* ("*kontextueller Aspekt*": tema è ciò che è dato o noto, contestualmente o co-testualmente; rema è ciò che è nuovo).

Questi due aspetti spesso (ma non sempre) vengono a coincidere. Per i suoi tipi di progressione tematica Daneš non ritiene necessario di tenerli distinti (Conte 1988, 49).

[In the Functional Sentence Perspective framework, theme and rheme have been defined both from a *thematic* point of view ("*thematischer Aspekt*": the theme is what one communicates something about; the rheme is what one communicates about the theme), and from a *contextual* point of view ("*kontextueller Aspekt*": the theme is what is given or known, contextually or co-textually; the rheme is what is new). These two aspects often (though not always) coincide. For the purposes of his types of thematic progression, Daneš does not consider it necessary to draw a distinction between them.]

Conte's remarks are crucial insofar as, if we think carefully about them, they reveal one of the most significant weaknesses of the Prague School's conception of thematic progression. The problem (which, as we will show below, is in fact even more acute than Maria-Elisabeth Conte suggests) lies in the fact that, in the original perspective, phenomena which ought to be kept separate are forced into one and the same level. One thing is to examine the cognitive status (Given or not Given) of the entities evoked in the text; another is to observe how each Utterance of a text chooses the entity about which it conveys information (i.e. the Theme); and yet another is to assess the degree of Communicative Dynamism associated with each entity. The overlap of these notions impacts on the very intelligibility of the phenomenon of thematic progression and its possible use as an analytical tool in Text Linguistics.

3.3 Although they all define aspects of the informational structure of Utterances, the phenomena related to the concepts of Givenness, “Thematicity” (in the sense of *aboutness*) and Communicative Dynamism must be carefully distinguished (but, as shown in Lambrecht 1994, it is true that there are preferential associations between these concepts). Indeed, it is easy to show that, in coherent texts, the phenomenon of semantic continuity – described by the property known as Givenness – (may) associate with different informational “spaces”: an entity that is Given in the text can coincide with the entity about which something is being said (i.e. the Theme), with what one says about it (the Rheme), with Transitions, or with collateral information etc. (according to the “spaces” defined by Firbas 1964; 1974). For instance, the Given referent indicated by *lui* in the examples given below shows a low degree of Communicative Dynamism in (5), where it coincides with the entity being talked about (the Theme), but a maximum degree of Communicative Dynamism in (6), where it comes to coincide with the Utterance peak (also referred to as “Rheme Proper” or Focus):

- (5) *BM2: c’ è [/] c’ è la mamma di Pierino / dice // va a comprarmi mezzo chilo di maiale //” e / **lui** ci va // (Cresti 2000, *Corpus di italiano parlato*, “Barzellette”)
- (6) *LCN: perché ci sono alcuni / per esempio / che sono ladri // poi uno / si fa amico di questo tizio / che è ladro // e dopo / coll’ amicizia / diventa ladro anche **lui** // (Cresti 2000, *Corpus di italiano parlato*, “Maestra”).

This distinction between the concepts mentioned above is essential not only because it is a necessary condition for a better understanding of both the substance and the forms of the informational structuring of Utterances, but also because it is a requirement for an adequate understanding and description of text structuring.

4. The levels of the informational organization of the Utterance²

Besides an illocutionary component, the communicative meaning of an Utterance (the Utterance being defined in terms of a linguistic act) includes (at least) a denotational component, corresponding to the state of affairs evoked by the speaker/writer. The denotative meaning of every Utterance is in turn organized informationally, i.e. it is organized “as a message” according to both the context in which it is expressed and the addressee to which it is addressed. Although the informational organization of the denotative meaning of an Utterance is determined first and foremost by the context, it is partially predetermined by the linguistic structures used by the speaker/writer, i.e. by the lexical items used, by the prosody or punctuation, the morphology, and the syntax used in the Utterance to denote a state of affairs³.

As mentioned above, the informational organization of the Utterance is a complex system, which can be broken down into several different levels⁴. Two of the levels that should be identified are the following⁵:

- a. the level the notion of *aboutness* belongs to, which is expressed by the informational functions known as “Topic⁶” and “Comment”;
- b. the hierarchical-illocutionary level, organized according to the illocutionary and textual functions performed by the Utterance in the message. The central informational units expressing this level (in written texts) will be called “Nucleus”, “Frame” and “Appendix”.

Each level plays an important role in defining the organization of texts. The Topic-Comment level shapes text evolution by selecting its successive “topics”, in particular by determining constituent order and partly also by selecting the form – pronoun vs. full lexical item – those constituents should have. The hierarchical-illocutionary level defines the architecture of the text in terms of “backgrounds” and “foregrounds”, mainly by determining the distribution of circumstantial elements (adverbial, clauses) and by imposing certain choices in punctuation.

4.1 The Topic-Comment level

A Semantic Proposition, defined as the mental image of a real, supposed or imagined state of affairs, typically evokes one or more “textual referents”, i.e. (following Andorno 2003, 27-68) “conceptual objects” characterized by a number of properties and/or involved in actions, processes, or states. From an ontological perspective, conceptual objects are typically first-order entities (namely physical objects: human beings, animals and things). In an appropriate context, however, i.e. when they are treated by the language as referents to which properties, etc. are assigned (typically through phrase nominalization), conceptual objects may also coincide with entities of a higher level, particularly with second-order entities (corresponding to events, processes, situations taking place in time, which are described as “occurring” or “taking place” rather than as “existing”; cf. Lyons 1980, 78). In the following Utterance, for instance, the complex phrase *la partenza di Gianni*, expressing the event of leaving by the individual called *Gianni*, is treated by the language as an entity to which the property *mi ha molto sorpresa* is assigned:

- (7) **La partenza di Gianni**_{referente testuale} mi ha molto sorpresa.

4.1.1 The notions of “Topic” and “Comment”. Within the Proposition, one (or more) textual referent has the function of Topic if, in the terms of Lambrecht,

in a given situation the proposition is construed as being about this referent, i.e. as expressing information which is relevant to and which increases the addressee's knowledge of this referent (Lambrecht 1994, 131).

As this definition suggests, the relation of *aboutness* defining the Topic is not to be understood generically: otherwise every referent⁷ of a Proposition would be involved in such a relation. The *aboutness* relation applies to one (or more) “communicatively special” referent, i.e. to the referent for which semantic enrichment is required by the context for communicative reasons. For instance, in the following text, only the referents appearing in boldface function as a Topic of the Semantic Proposition in which they appear; the other referents – *Elena Pistolesi, Antonella Benucci, le caratteristiche dei segnali discorsivi di e-mail e SMS* – are not Topics:

- (8) **L'italiano delle chat**_{Topic} è stato studiato da Elena Pistolesi (*L'italiano nella rete*), **quello della pubblicità televisiva**_{Topic} da Antonella Benucci (*La pubblicità televisiva e l'italiano non standard*), mentre **Carla Bazzanella**_{Topic} (*Nuove forme di comunicazione a distanza, restrizioni contestuali e segnali discorsivi*) ha analizzato le caratteristiche dei segnali discorsivi di *e-mail* e *SMS*. (Maraschio and Poggi Salani 2000, IV-V).

A topical element may involve one or more textual referents, as shown in:

- (9) **L'inaugurazione in Palazzo Vecchio nel Salone dei Cinquecento e il saluto del sindaco di Firenze Leonardo Domenici, la presenza in qualità di relatore del Ministro della Pubblica Istruzione Tullio De Mauro in apertura e quella del Presidente della Camera dei Deputati onorevole Luciano Violante in chiusura**_{Topic} hanno conferito al XXXIV Congresso della SLI un carattere di ufficialità in sintonia con la ricorrenza millenaria. (Maraschio and Poggi Salani 2000, VII).

In some cases, as in text (10), two referents belonging to the same Proposition share the property of being “communicatively special”, i.e. are topical:

- (10) Umberto cresce con la madre, che gestisce un negozio di oggetti usati, e con due zie, una delle quali, la zia Regina dalla “dolce anima di formica”, gli sarà prodiga di attenzioni e di aiuti. **A lei**_{Topic1}. **Saba**_{Topic2} dedicherà affettuosamente le prose raccolte nel volume *Ricordi-Racconti* nel 1956. (Lavezzi et al. (eds) 1992, 630).

As all the previous examples show, the Topic comes to function naturally as “point of departure of the message” (Halliday 1985, 38). In addition, it is usually particularly “evident” in the communicative context, because its content is typically associated with direct or indirect Givenness. In turn, the referential and functional profile of the Topic determines its preferred linguistic manifestations. The Topic is usually expressed by a bound personal pronoun or by a pre-verbal nominal or

prepositional phrase; in the sentence with normal word order, the Topic coincides with the syntactic subject (even when it is not phonetically realized). “Topicality”, however, is ultimately granted to a referent based on the context of usage. The context may select as Topic the referent of a semantically full element in post-verbal integrated position (although this is more typically the case in spoken than in written communication); or it may select one of several potential Topics made available by the linguistic component.

The semantic element functionally related to the Topic is the Comment. The Comment follows the linguistic expression of the Topic, and coincides with the predicate of a Proposition (it can also include circumstantial information):

- (11) Umberto **crece con la madre**, che gestisce un negozio di oggetti usati, **e con due zie**_{Comment}, una delle quali, la zia Regina dalla “dolce anima di formica”, gli sarà prodiga di attenzioni e di aiuti. A lei, Saba **dedicherà affettuosamente le prose raccolte nel volume *Ricordi-Racconti nel 1956***_{Comment}. (Lavezzi et al. (eds) 1992, 630).

4.1.2 The informational functions of Topic and Comment do not necessarily account for the whole content of a Semantic Proposition. In the following two examples, for instance, the elements in boldface opening the Utterance function neither as Topic nor as Comment; rather, they function as “circumstantial” information:

- (12) **Nell’ultimo decennio**, l’industria_{Topic} si è sviluppata accanto al porto_{Comment}. (in Ferrari and Zampese 2000, 335)
- (13) **Dato che aspettano un bambino**, Michela e Luca_{Topic} hanno deciso di cambiare casa_{Comment}. (in Ferrari and Zampese 2000, 336).

The elements provided in addition to the Topic and the Comment do not necessarily have the same informational status. The informational difference between the elements added to the Topic and the Comment can be seen in example (14), in which several pieces of information appear between the Topic (*Roger Wright*) and the Comment (*insiste sulla [...] romanza*). In this example it is clear that the content in parentheses has a different informational status from the sentence opened by the gerund *ricordando* that follows (the informational status of Utterance elements will be discussed in more details in paragraph 4.2.):

- (14) Roger Wright_{Topic} (*La periodizzazione del romanzo*) ricordando il diverso **situarsi nel tempo di tanti romanzi di natura diversa e i lunghi periodi di variabilità, e nonostante l’esistenza di alcuni noti riferimenti fondamentali**, insiste sulla difficoltà del definire date spartiacque nell’evoluzione dal latino al romanzo e nella frammentazione romanza_{Comment}. (Maraschio and Poggi Salani 2000, I-II).

4.1.3 It should be noted that there are Semantic Propositions in which it is neither possible nor relevant to identify elements that function as Topic and as Comment and that are therefore lacking a Topic-Comment articulation. These Propositions are typically the ones that grammar calls “presentative” or “eventive”, namely the clauses constructed with a zero-valent verb, i.e. not denoting any referent (*Piove*), the clauses opened with the expression *c’è/ci sono* (*al Polo Nord ci sono orsi bianchi*), impersonal constructions (*si dice che i treni svizzeri sono sempre in orario*) and sentences with post-verbal subjects (*è arrivata Stella*).

4.2 The hierarchical-illocutionary level

In formulating an Utterance with communicative intentions, the speaker/writer accomplishes an illocutionary act (assertive act, interrogative act, etc.). Simultaneously, if the Utterance is part of a co-text, the speaker/writer also accomplishes an act of textual composition (an act that can function, for instance, as explanation, reformulation, illustration of a preceding text). As we will show, not all the information expressed in an Utterance is equally relevant, i.e. has the same communicative prominence (or dynamism) in determining the illocutionary and/or the textual act that the Utterance performs in the message.

Our conception of what we refer to as the “hierarchical-illocutionary informational level”, was largely inspired by the work of Cresti (in particular, Cresti 2000) about the informational articulation of the spoken Utterance. It also diverges from it, however, in different ways. For instance, the two models use a partly different terminology (cf. Ferrari (ed.) 2004, (ed.) 2005). There are, however, more important differences. First, while Cresti’s hypothesis fails to account for the analytical tools provided by Text Linguistics, our model of text has integrated these tools. We find that it is inevitable that they must be dealt with when analyzing dialogic exchanges that are not restricted to an informationally-simple Utterance. The important role played in our model by the concepts borrowed from Text Linguistics can be measured, for instance, by our use of the concept of “act of textual composition” (see Ferrari et al. 2008 for an account of its nature and its organization in sequences). Secondly, unlike Cresti’s model, our account of the Text can be applied equally to the oral and the written language. Therefore, we strongly disagree with the idea that the sentence – even if conceived as having a semantic component (cf. Cresti in press) – should be considered the reference unit of written text (see Ferrari in press for a more detailed discussion of this question).

4.2.1 The main Information Unit. Some information is more directly linked to the illocutionary and textual function which the Utterance performs in the message. This information defines a unit which – using the terminology of Blanche-Benveniste et al. 1990 – we will call “Nucleus” or “Nuclear Unit” (It. *Nucleo* or *Unità Nucleare*). For instance, the nucleuses of the two Utterances that form the text given in (15)

coincide with the sequences appearing in bold, because it is in these two sequences that the concession relation indicated by *eppure* is based⁸:

- (15) 1. // / Leggendo qualche anno fa il bel libro di Rosetta Loy, / *La parola ebreo*, / **mi sono reso conto che questa parola mi era stata estranea assai a lungo nell'infanzia.** /^{Nucleo} // 2. / Eppure / **Napoli aveva avuto e ancora aveva una comunità ebraica non piccola,** /^{Nucleo} non irrilevante socialmente, / a cominciare dagli avvocati Foà, presso cui era Giovane di Studio il neolaureato Giovanni Leone. // (De Mauro 2006, 103).

An informational Nucleus is necessary and sufficient for an Utterance to be expressed. Normally, though – that is, if we do not deliberately choose what is known as *style coupé* – the Nucleus is accompanied by other Informational Units, which provide backgrounded information. These Units, which are optional and repeatable, will be given the names of “Frame” and “Appendix”.

4.2.2 Secondary Information Units. The “Frame Unit” (It. *Unità di “Quadro”*) linearly precedes the Nucleus. From a functional perspective, it indicates the general denotational domain of relevance of the Nucleus. The Frame Unit may be used locally. In this case, it indicates the circumstances (most commonly spatial, temporal, and modal) in which the event described in the Nucleus takes place, as is, for instance, the case in:

- (16) // / **Leggendo qualche anno fa il bel libro di Rosetta Loy,** /^{Quadro} *La parola ebreo*, / mi sono reso conto che questa parola mi era stata estranea assai a lungo nell'infanzia. // (De Mauro 2006, 103).

The Frame Unit may limit or extend the “implicature” effects of the Nucleus, as in (17): here, the indication, in the Frame Unit, of a possible cause of the accident described in the Nucleus leads one to read the rest of the article with the guilt of the driver in mind:

- (17) // / **Forse per un sorpasso azzardato,** /^{Quadro} l'Alfa Romeo 145 guidata da Alessandro Granata [...] ha urtato la Renault Clio [...] // (*Corriere della Sera*, in Zampese 2004, 175).

In addition, the Frame Unit may make explicit illocutionary components of the Nucleus, as, for instance, the source of the Utterance:

- (18) // / **Secondo i carabinieri,** /^{Quadro} si è trattato di una rapina su commissione. // (*Corriere della Sera*, in Zampese 2004, 175).

Besides its “local” motivation, the Frame Unit may also have a function that goes beyond the Utterance in which it is expressed. In this case the content of the Frame Unit is chosen to create, anaphorically, the semantic link that insures / makes explicit / underlines the relationship between the Nucleus and the preceding co-text. This function can be illustrated on the basis of example (19a) and a manipulated version of it (19b): the deletion in (19b) of the anaphoric expression *per questo*, which makes explicit the logical link between the main content (*i prigionieri hanno sfondato la porta*) and the preceding Utterance (*non li hanno però legati*) in (19a), leads to a somewhat “unnatural” link between the two Utterances:

- (19) a. Non li hanno però legati: // **per questo**, ^{Quadro} intorno alle tre, a forza di spallate / i prigionieri hanno sfondato la porta. // (*Corriere della Sera*, in Zampese 2004, 176)
 b. Non li hanno però legati: // intorno alle tre, a forza di spallate / i prigionieri hanno sfondato la porta. //

Likewise, the Frame Unit may be used cataphorically, to open semantic “spaces” that provide unity to the following co-text. In the example given in (20), the phrase expressed in the Frame (*All’epoca in cui risalgono i miei primi ricordi di lei*) provides the perspective from which to interpret the long description that follows:

- (20) // **All’epoca in cui risalgono i miei primi ricordi di lei** [mia madre], ^{Quadro} era, anche con parametri del tempo, una donna ancora giovane, trentottenne. // Aveva una carnagione bianco-latte e lunghi capelli rosso scuro. Mi pareva bellissima. Dopo essersi lavata, rivestita da un accappatoio candido pettinava la lunga chioma, che le ricadeva davanti al viso. [...] (De Mauro 2006, 103).

The “Appendix Unit” (It. *Unità di “Appendice”*), which is also repeatable, completes the Nucleus and/or the Frame Unit(s). The Appendix may be expressed within the Nucleus and the Frame, or be placed immediately after them. From a functional perspective, the Appendix has a local impact in the text: its function is typically restricted to the Utterance in which it is expressed. The Appendix can be used to repeat Given (simple or complex) information, or to reactivate Semi-Given information⁹. When it provides New (or almost New) information in the text, the Appendix may be used by the speaker/writer to specify the meaning of his/her words, or simply to add information that is relevant but not textually “vital”, i.e. that is not capable of functioning as a semantic reference framework in the following co-text, and thus is not truly connected to the preceding co-text¹⁰.

The Frame and the Appendix Units both provide backgrounded information and may express the same semantic contents. The contribution they make to textuality, however, is very different. This can be shown by the impact the transformation of the content of a Frame in Appendix or vice versa has on text coherence (cf. Ferrari 2003, (ed.) 2004, 2006). Texts (21) and (22), which deal with the comparison

between Italian and other languages, offer an example of the difficulty in transforming a content that is part of a Frame Unit into one that is part of an Appendix Unit:

- (21) Oggi sappiamo che la fissità dell'italiano è stata alquanto sopravvalutata. Non vi è dubbio, infatti, che anch'esso sia mutato nel corso del tempo; // / **rispetto alle altre lingue**, /^{Quadro} però, questo mutamento è stato per secoli più contenuto (o meno avvertibile), // tanto che sembra avvenuto quasi di colpo dalla fine dell'Ottocento, dopo il raggiungimento dell'unità nazionale, fino al duemila, l'epoca dell'informatica e della multimedialità. (in Ferrari 2006, 74).
- (22) Oggi sappiamo che la fissità dell'italiano è stata alquanto sopravvalutata. Non vi è dubbio, infatti, che anch'esso sia mutato nel corso del tempo; // / questo mutamento, / **rispetto alle altre lingue**, /^{Appendice} è stato però per secoli più contenuto (o meno avvertibile), // tanto che sembra avvenuto quasi di colpo dalla fine dell'Ottocento, dopo il raggiungimento dell'unità nazionale, fino al duemila, l'epoca dell'informatica e della multimedialità.

In this case, it is a better choice to evoke the comparison between Italian and other languages within a Frame Unit than within an Appendix Unit, because the Appendix Unit treats such information as an accessory specification.

5. The Topics' organization in the text

5.1 From Proposition Topic to Discourse Topic

Rigorously defined in terms of *aboutness*, the Topics of the (Semantic) Propositions which make up the text are part of a multi-level system, which has already been outlined by van Dijk 1977 in the initial phases of Text Linguistics. The main idea is that a text can be considered as the semantic-pragmatic expansion of a basic (central or discursive) Topic (or Theme). The expansion occurs through the development of lower level Topics down to the level of the Semantic Proposition Topics (in the sense defined in paragraph 4.1). The result is a reversed tree diagram, whose trunk is formed by the central Topic, whereas lower level Topics are distributed in the progressively "newer" branches of the tree.

The diagram described above represents the global organization of Topics in the text. Within this representation, by "Topic progression" we mean the ways in which Topics at different levels follow each other in a coherent text segment: we can thus identify the progression of chapter Topics within a book, the progression of section Topics within a chapter, that of Paragraph Topics within a section, and so on.

When a Paragraph or one of its subparts is considered – as was the case within the Prague School framework – the relevant dimension of analysis is the Utterance. Therefore, the concept of “thematic progression” developed by František Daneš concerns (partially: cf. above) the sequence of Topics in the Utterances: for instance, we have constant Theme progression when two or more contiguous Utterances share the same Topic. Thus, in order to identify the thematic progression of a Paragraph, it is necessary to define the ways in which Topics are linearly distributed in the Utterances that form the Paragraph.

Minimally, however, the identification domain of the Topic is the Semantic Proposition. It is here that the interaction between the Topic-Comment level and the hierarchical-illocutionary level comes into play. In this interaction two processes are involved: firstly, the hierarchical-illocutionary level selects which Topics made available by the Propositions forming the Utterance is/are the Topic/s of the Utterance; secondly, the hierarchical-illocutionary level defines whether the Utterance Topic also functions as a higher-level Topic, i.e. a Topic of a homogeneous group of Utterances or of a Paragraph. A similar proposal has been made by Iørn Korzen (see for instance Korzen 1998, 2001) on the basis of concepts drawn from the Rhetorical Structure Theory framework (cf. Mann and Thompson 1988). Korzen identifies two different Topics: a primary Topic, associated with the nucleus, *i.e.* with a main clause, and a secondary Topic, found in the satellite, *i.e.* in a subordinate clause. In our conception, however, there is no necessary association between Topic level and syntax.

5.2 The Utterance Topic

The Utterance Topic coincides with the Topic associated with the main Semantic Predication expressed in the Nuclear Unit. In the following example, *Maria* (and not *Carlo*, which functions as Topic of a Secondary Predication) is the Utterance Topic:

(23) // / **Maria**_{Topic} mi ha detto che Carlo non ha capito nulla. /^{Nucleo} //

The Utterance Topic may be either incorporated into the Nucleus – in this case the Nucleus contains both the Topic and the Comment of the main Semantic Proposition, as in (23) – or expressed within the Frame Unit, in which case it is followed by a predicative Nucleus, i.e. a Nucleus containing the Comment of the main Semantic Proposition, as in (24):

(24) Ieri mattina, patenti e carte d'identità dei 32 avventori rapinati sono state ritrovate a Milano, in una cassetta della posta. // / **Dei malviventi**_{Topic} /^{Quadro} invece, / **non c'è traccia**_{Comment} /^{Nucleo} // (in Zampese 2005, 209).

Given the two possible informational manifestations of the Topic, more complex cases than (23) and (24) can be observed. They are Utterances like (25), and have two main (Utterance) Topics. The first is expressed in the Frame Unit, the second in the Nucleus:

(25) // / **Per Toce**_{Topic1}, ^{/Quadro} **le manette**_{Topic2} sono invece scattate mercoledì scorso.
^{/Nucleo} // (in Zampese 2005, 209).

In cases like (25), the Comment about the first Topic coincides with the information conveyed by the entire Nucleus (*le manette sono invece scattate mercoledì scorso*), whereas the Comment about the second Topic corresponds to the central predication of the proposition that appears within the Nucleus (*sono invece scattate mercoledì scorso*).

The Topic expressed in a Frame Unit can have various linguistic forms. It can have the form of a left-dislocated referential phrase, as in example (24), or of a circumstantial textual referent introduced by expressions such as “*quanto a x*”, “*riguardo a x*”, “*in relazione a x*”, “*a proposito di x*” (i.e. “as of x”, “regarding x”, “in relation to x”). These expressions have the function of evoking one or more entities involved in the event denoted in the Nucleus:

(26) // / Quanto a **Luisa**_{Topic}, ^{/Quadro} **non mi ha ancora risposto**_{Comment} ^{/Nucleo} //;

A Topic expressed in the Frame Unit may also coincide with a pre-verbal syntactic subject:

(27) // È vero che hanno partecipato tutti in modo attivo. // / **Alice**_{Topic}, ^{/Quadro} però,
^{/Appendice} **è stata proprio decisiva**_{Comment} ^{/Nucleo} // A lei_{Topic} dobbiamo l’idea
 ecc. // (in Ferrari et al. 2008, 155).

Although they may all serve the function of expressing the Topic within the Frame Unit, the linguistic forms mentioned above – dislocated constituents, referents introduced by particular expressions, syntactic subjects – are not equivalent, and therefore not interchangeable. Firstly, because they do not belong to the same language register: while expressions like “*per quanto riguarda x*” are typical of formal written texts, in informal spoken language marked constructions such as left dislocations or “hanging topics” are more common (in written texts, a dislocated constituent has the tendency of being incorporated into the Nuclear Unit, cf. Ferrari 2003). Secondly, the linguistic forms mentioned are not interchangeable because they do not create an autonomous Topic, i.e. a Topic detached from the Nucleus, with the same ease: while expressions such as “*quanto a x*”, when opening an Utterance, impose de facto the presence of a Topic within the Frame Unit, it is much more difficult to decide whether a pre-verbal syntactic subject is part of the Frame

Unit or the Nuclear Unit of the Utterance to which it belongs. For these cases, the following general rule applies: when a topical subject is separated from its Comment by a sequence that is part of a background Informational Unit, typically an Appendix Unit, like *però* in (27), it also forms an autonomous Frame Unit.

5.2.1 As we have seen, the Utterance Topic is the Topic associated with the Nucleus, whether incorporated or non-incorporated in this Nucleus. While the incorporation degree of the Topic has no specific effect on the thematic progression described above, its degree of integration has an impact on a more general level. Compared with a Topic expressed in the Nucleus, a Topic that is part of the Frame Unit more easily serves wide-ranging textual functions.

A Topic placed in the Frame Unit can have a significant structuring function – and this is its first important contribution to the thematic progression of the text. Specifically, a Topic in the Frame Unit is ideal for signaling the topical macro-articulation of Paragraphs. In a somewhat “iconic” way (owing to its initial position), the Frame Unit comes to determine the referent which functions as macro-Topic of the Paragraph as well as, simultaneously, the referential links to the preceding co-text, and consequently the topical coherence of the whole text. The following is a straightforward example of the structuring function of the Topic expressed in the Frame Unit:

- (28) Chi dice che la nostra lingua è stata guastata dagli eccessivi innesti di anglicismi si sbaglia di grosso. Luoghi comuni. Sì, va bene, qualche prestito è inevitabile in tempi ad alto tasso tecnologico. Ma la lingua italiana è essenzialmente una lingua conservativa, refrattaria ai cambiamenti. È quanto sostiene **Luca Serianni**, autore di una Garzantina sulla grammatica e la sintassi dell’italiano che riprende un volume uscito dieci anni fa per la Utet e che sarà in libreria fra qualche giorno (pagine 609, lire 42.000).

// / **A Serianni**_{Topic} /^{Quadro} che insegna Storia della lingua all’Università di Roma, /^{Appendice} non piacciono per nulla i continui lamenti sulla corruzione dell’italiano: /^{Nucleo} // “Quel che fa testo è l’uso reale. [...] // Insomma, diciamo che per allestire una grammatica bisogna tener conto del livello medio, quello, per intenderci, di un buon **articolo di giornale**, né troppo elevato né troppo basso o familiare?”. //

// / **A proposito di giornali**_{Topic} /^{Quadro} Serianni insiste sulla tendenza conservativa: /^{Nucleo} // [...] (*Corriere della Sera*, 18.9.1997).

Of course, a Topic incorporated in the Nucleus may also coincide with a macro-Topic. This happens when the Utterance Topic expressed in the Nucleus is resumed in one or more Utterances of a text sequence, that is, when constant Theme progression occurs. Unlike the Topic that is part of a Frame Unit, however, in order to be promoted to macro-topic of a text sequence (for instance, of a Paragraph), a

Topic expressed in the Nucleus must be repeated in each main Semantic Proposition of the text (note that in Italian it can also take the form of an omitted subject).

The structuring function of a topical referent expressed in the Frame Unit is a precious textual tool because it can help the reader easily to identify chunks of texts with the same macro-Topic. This function therefore comes in handy, for instance, in examples like (29), where there is no parallelism between a Paragraph and a main, independent Topic:

- (29) 1.2.3. Passando alla sintassi, possiamo individuare principalmente tre ambiti nei quali si avverte differenza tra il parlato e lo scritto standard: la sintassi del periodo, fenomeni concernenti l'ordine dei costituenti frasali, e fenomeni riguardanti l'effetto della ridotta gittata di pianificazione sulla coesione sintattica delle frasi.

// / Quanto al **primo punto**_{Topic}, /^{Quadro} c'è ampio accordo fra gli studiosi sul fatto che il parlato preferisca la paratassi all'ipotassi: /^{Nucleo} // [...]

Inoltre, la realizzazione dei nessi di subordinazione (e coordinazione) è affidata a una gamma minore di forme e congiunzioni rispetto alle possibilità dello scritto standard, ciascuna delle quali avrà una frequenza relativa più alta: [...]

Meritano infine un rapido cenno, in tema di strutture frasali ricorrenti nel parlato, due costrutti tipici. [...]

// / Quanto a **fenomeni relativi all'ordine dei costituenti**_{Topic}, /^{Quadro} è rilevante la presenza di frasi segmentate di vario genere, /^{Nucleo} aventi in comune la funzione di sottolineare l'articolazione tema/rema e di marcare la struttura informativa della frase. /^{Appendice} // [...]

Un terzo carattere evidente della sintassi nei testi parlati è dato dalla ricorrenza di strutture sintattiche interrotte: [...] (Berruto 2003, 47-8, in Ferrari and De Cesare in press).

5.3 Secondary Topics

According to what has been said so far, the secondary Topics which may be found in Frame and/or in Appendix Units, and more precisely in main Propositions fully expressed in the Information Unit of Frame and/or Appendix (when referring to a "secondary Topic" we are therefore not speaking of the case identified, for instance, in example (27), where the Topic is expressed in a Frame Unit without the Comment that is made about it), are not directly involved in the primary Topical organization of the Paragraph. Consider, for instance, the two Utterances of the following invented text, where *Maria* functions as Utterance Topic:

- (30) // / **Maria**_{Topic} ha un pessimo carattere. /^{Nucleo} // / Quando **la situazione**_{Topic} si fa difficile, /^{Quadro} [**sogg. nullo**]_{Topic} non riesce mai, /^{Nucleo} malgrado **Carlo**_{Topic} tenti pazientemente di mediare, /^{Appendice} a mantenere la calma. /^{Nucleo} // (in Ferrari et al. 2008, 155).

In her first Utterance, the speaker/writer expresses a negative judgment about the Utterance Topic *Maria*. In the second Utterance, the speaker/writer justifies her negative evaluation by evoking a typical behavior of *Maria*. This example shows that both the Frame Topic (*la situazione*) and the Appendix Topic (*Carlo*) are not directly involved in the topical progression of the text. The text progresses by keeping *Maria* a constant Topic.

It would, however, be wrong to assume that the phenomenon of “thematic progression” is defined solely by the Nuclear Units of the text, and that secondary Topics do not play any role in determining this text organization.

5.3.1 Firstly, a Topic expressed in a Proposition included in the Frame Unit may help to identify the Utterance Topic. Let us look at example (31), a case where a nuclear Semantic Proposition (in boldface) is expressed at the very beginning of the text. Owing to its text-opening position, the Semantic Proposition expressed in the Nucleus is compatible with two interpretations: with an informational articulation in Topic and Comment and with no such articulation (the Proposition is thus interpreted as a global event):

- (31) // **Sbarbaro ebbe una vita schiva e priva di eventi esterni significativi**, ^{/Nucleo}
tutta trascorsa nella sua Liguria. ^{/Appendice} // (adapted from Lavezzi et al. (eds)
1992, 524).

Now, if the text given in (31) is preceded by a Frame Unit containing a Proposition in which the referent *Sbarbaro* is the Topic, as in (32), we are naturally led to interpret the referent *Sbarbaro* of the Proposition expressed in the Nucleus as topical, too:

- (32) // Nato [= **Sbarbaro**_{Topic}] a Santa Margherita Ligure (in provincia di Genova)
nel 1888, ^{/Quadro} **Sbarbaro**_{Topic} ebbe una vita schiva e priva di eventi esterni
significativi, ^{/Nucleo} tutta trascorsa nella sua Liguria. ^{/Appendice} // (Lavezzi et al.
(eds) 1992, 524).

In cases like (32), the following general rule applies: if the text-opening nuclear Proposition is preceded by a Frame Unit containing a clause that has a Topic-Comment articulation, and if the Frame Topic is co-referent with the first element of the nuclear Proposition, there is a good chance that this element also functions as Topic. This rule is based on the “presupposition effect”, according to which the first topic of the initial Proposition (i.e. of the Frame) also serves as Topic of the entire Utterance.

A second important contribution made by the Topic(s) expressed in the Frame Unit to the thematic organization of texts is to be defined in terms of semantic-referential continuity. Through its collocation in the Frame Unit, the Utterance Topic may have the function of linking to the preceding co-text a Nucleus that is

informationally eventive or presentative, i.e. that lacks a Topic of its own. This support of the topical continuity of the text is possible both when the Topic is expressed in a Frame Unit that does not contain the Comment and when the Frame Unit contains a Proposition with Topic-Comment articulation, as in (33). In this example, the Topic of the last Frame Unit, i.e. the referent *Croce*, ensures referential continuity of a presentational Utterance (as shown by the opening of the Nucleus with *ecco*) with the preceding text:

- (33) // Essendo descrittiva e caratterizzante, la critica crociana_{Topic} comporta anzitutto, o almeno in un primo tempo, un continuo cedere la parola all'autore, vale a dire grande abbondanza di citazioni molto ben manovrate, sfiorando quello che si potrebbe chiamare un racconto che ha per protagonisti gli autori. // E secondariamente [la critica crociana_{Topic} comporta] una grande attenzione alle psicologie e dunque ai personaggi. // / E poiché per **Croce**_{Topic} non si tratta soltanto di individuare ma di inserire con un gesto largo in categorie, /^{Quadro} **ecco gli insistenti paragoni tra testi lontani e allotrii:** /^{Nucleo} // [...] (Mengaldo 1998, 14-15).

This second contribution to the thematic organization of the text is crucial: in these cases, the Topic expressed in the Frame Unit solves the “informational-textual problem” affecting those particular logical relations – typically illustration, exemplification, consecution – which have simultaneously to manage within one and the same referential domain topic continuity and informational dynamism.

5.3.2 The Topic of a Semantic Proposition expressed in an Appendix Unit – that is, within the Information Unit most typically used to enrich the intensional description of textual referents – may also be crucial in text construction and interpretation. For instance, it may introduce a referent which can become important in the ensuing discourse. Let us consider the following example:

- (34) 1. // Umberto_{Topic} cresce con la madre, /^{Nucleo} che gestisce un negozio di oggetti usati, /^{Appendice} e con due zie, /^{Nucleo} **una delle quali**_{Topic}, /^{Appendice} la zia Regina dalla “dolce anima di formica”, /^{Appendice} gli sarà prodiga di attenzioni e di aiuti. /^{Appendice} // 2. / **A lei**_{Topic1}, Saba_{Topic2} dedicherà affettuosamente le prose raccolte nel volume *Ricordi-Racconti* nel 1956. /^{Nucleo} // 3. La sua carriera scolastica_{Topic} è piuttosto breve: // 4. [Umberto Saba_{Topic}] frequenta il ginnasio e, soltanto per pochi mesi, l'Accademia di Commercio, abbandonata quasi subito per la necessità di trovare un lavoro (si impiega presso una casa di commercio triestina). (in Ferrari and De Cesare in press).

The text from which (34) is taken is about the poet Umberto Saba, and yet in the Paragraph reproduced in (34) another important figure emerges: it is the referent *la zia Regina*, which functions as Topic of the Appendix of the first Utterance. *La zia*

Regina is introduced in the text as background information of Utterance 1, and is then promoted to Utterance Topic of Utterance 2.

When the Appendix resumes a textual referent without enriching its intensional semantics, the Topic that it contains may also contribute to rendering the semantic continuity of the text more transparent. For instance, in (35) the Topic expressed by the full lexical noun phrase *Lady D.* in the Appendix Unit helps to identify the Utterance Topic expressed by the omitted subject of the nuclear Proposition:

- (35) // [sogg. nullo_{Topic}] Avrà, /^{Nucleo-} perché **Lady D.** _{Topic} è nel cuore di tutti, /^{Appendice}
un addio da regina /^{-Nucleo} // (in Ferrari et al. 2008, 172).

Repeating a Topic by using a full lexical phrase placed in the Appendix Unit insures the correct referential interpretation of the Utterance Topic. Simultaneously, this procedure allows avoidance of undesired dynamization of the information. In the Italian written text tradition, where repetition tends to be stigmatized, the resumption in the Appendix Unit of a Given or Semi-Given referent through a full nominal phrase comes in very handy. It enables the writer to assign lower informational prominence to that referent than by placing it in the Nucleus (or the Frame Unit).

6. Conclusion

In this paper we have shown that the Topical progression of a Paragraph (or of one of its subparts) is defined by the modular association of two informational levels: the Topic-Comment level and the hierarchical-illocutionary level. Based on partly linguistic and partly contextual clues, a semantic-pragmatic coherent textual sequence indicates: a. a sequence of textual referents, defined within the Proposition, which function as Topic; and b. a sequence of illocutionary and textually-based Utterances, with an internal, hierarchical organization expressed by the Information Units of Nucleus and – optionally – Frame and/or Appendix. The Topical organization of the text selects, from among the possible Topics, the ones that are linked to the sequence of informational Nucleuses – be they a constitutive part of them (propositional Nucleus) or a dislocated element within the Frame Unit (predicative Nucleus). This, however, does not mean that secondary Topics – i.e. the ones expressed in a Semantic Proposition included in the Frame Unit or the Appendix Unit – are not related to the phenomenon of thematic progression. As we hope we have shown, secondary Topics are crucial for text construction and interpretation. They may help to identify and clarify the main Topic of the Utterance and can introduce new textual referents which can gain prominence in the following text.

Notes

* This paper is a revised and shortened English version of Ferrari and De Cesare in press. We are very thankful to Claudia Ricci for her help in translating this document into English.

¹ A more detailed account of the model of text organization that we present in this paper has been given most recently in Ferrari and colleagues 2008. Our research has been funded by the Swiss National Science foundation (project no. PP001--68675/1. Project title: *L'organizzazione informativa dell'Enunciato scritto (in italiano contemporaneo non letterario)*).

² The topic of this section is explained more extensively in Ferrari et al. 2008.

³ For a more detailed account of the semantic interpretation of linguistic acts presented here, cf. Pasch et al. 2003.

⁴ The multi-level nature of the informational organization of the Utterance has long been recognized by scholars (cf. Daneš 1964; Halliday 1967a/b, 1985; Lambrecht 1994; and, in Italy, Lombardi Vallauri 1996, 2002).

⁵ At least three more levels could be added to the two we consider in this paper: (1) the level that defines the different states of information activation of entities within Textual Memory, which can be defined, in line with Chafe 1987, as “Active”, “Semi-Active”, “Inactive”; (2) the informative level based on the status of textual referents in the Encyclopedic Memory (as opposed to the Textual Memory) – the referents can be present, absent or inferable; (3) the informational level based on the conceptual opposition between “asserted” and “presupposed” entities.

⁶ In the rest of the paper we prefer to use the term “Topic” (first defined by Mathesius 1915 and later by Hockett 1958) rather than “Theme”, because the latter is associated with too many different meanings and informational functions. Our terminological choice is consistent with the work of Lambrecht 1994, which is one of the most convincing and thorough accounts of this informational unit, and with the terminology adopted in the field of Discourse Analysis (in Germany, France, the United Kingdom, United States, etc.).

⁷ In what follows, the terms “referent” (used in Lambrecht 1994) and “textual referent” will be used interchangeably.

⁸ From now on – in line with Cresti 2000, who was the first in Italy to study this aspect of informational organization of the Utterance in depth – within the hierarchical-illocutionary analysis of the Utterance a double slash (//) indicates an Utterance boundary, while a single slash (/) indicates an Information Unit boundary. Boundary indications will not be provided in every example and throughout the example: we will provide them only when they are necessary for the discussion.

⁹ The “Appendix Unit” is different from the “Parenthetical Unit” (It. *Unità di “Inciso”*), which is illocutionary-independent and which creates an autonomous level of text (cf. Cignetti 2004).

¹⁰ As has perhaps been noted, the hierarchical-illocutionary level is the closest to the level defined by Jan Firbas 1974 on the basis of the concept of Communicative Dynamism. The Nucleus coincides with the most dynamic information (the Rheme), while the other Units

represent the textual equivalents of the properties of “Theme”, “rest of the Theme”, “Transition”, etc.

References

- Andorno, C. 2003. *Linguistica testuale. Un'introduzione*. Roma: Carocci.
- Blanche-Benveniste, C., M. Bilger, Ch. Rouget and K. van den Eynde. 1990. *Le français parlé. Etudes grammaticales*. Paris: Editions du CNRS.
- Chafe, W. 1987. Cognitive constraints on information flow. In R.S. Tomlin (ed.), *Coherence and Grounding in Discourse*. Amsterdam and Philadelphia: John Benjamins, 21-51.
- Cignetti, L. 2004. Le parentesi tonde: un segno pragmatico di eterogeneità enunciativa. In A. Ferrari (ed.), *La lingua nel testo, il testo nella lingua*. Turin: Istituto dell'Atlante Linguistico Italiano, 165-90.
- Combettes, B. 1988. *Pour une grammaire textuelle. La progression thématique*. Bruxelles and Paris: De Boeck-Duculot.
- Conte, M.-E. 1988. Determinazione del tema. In M.-E. Conte (ed.), *Condizioni di coerenza. Ricerche di linguistica testuale*. Florence: La Nuova Italia, 49-56.
- Cresti, E. 2000. *Corpus di italiano parlato*. vol. I. Florence: Accademia della Crusca.
- Cresti, E. La parataxe: articulation informative dans le parlé spontané vs juxtaposition syntaxique dans l'écriture littéraire? Paper presented at *La Parataxe*, Colloque International de Macro-syntaxe (Neuchâtel, 12-15 February 2007).
- Daneš, F. 1964. A Three-level approach to Syntax. *Travaux linguistiques de Prague* 1: 225-40.
- Daneš, F. 1970. Zur linguistischen Analyse der Textstruktur. *Folia Linguistica* 4: 72-8.
- Daneš, F. 1974. Functional sentence perspective and the organization of the text. In F. Daneš (ed.), *Papers on Functional Sentence Perspective*. Prague and Paris: Academia-Mouton, 106-28.
- Ferrari, A. 2003. *Le ragioni del testo. Aspetti morfosintattici e interpuntivi nell'italiano contemporaneo*. Florence: Accademia della Crusca.
- Ferrari, A. (ed.). 2004. *La lingua nel testo, il testo nella lingua*. Turin: Istituto dell'Atlante Linguistico Italiano.
- Ferrari, A. (ed.). 2005. *Rilievi. Le gerarchie semantico-pragmatiche di alcuni tipi di testo*. Florence: Franco Cesati Editore.
- Ferrari, A. 2006. La fonction informationnelle d'Appendice. De la dislocation à l'apposition à travers la composante informationnelle. *Cahiers Ferdinand de Saussure* 59: 55-86.
- Ferrari, A. In press. Note sulle unità di analisi dello scritto e del parlato. Convergenze e divergenze funzionali e strutturali. In A. Ferrari (ed.) *Subordinazione, coordinazione, giustapposizione* (Atti del X congresso SILFI). Florence: Franco Cesati Editore.
- Ferrari, A., L. Cignetti, A.-M. De Cesare, L. Lala, M. Mandelli, C. Ricci and E. Roggia. 2008. *L'interfaccia lingua-testo. Struttura e funzione dell'articolazione informativa dell'enunciato*. Alessandria: Edizioni dell'Orso.
- Ferrari, A. and A.-M. De Cesare. In press. La progressione tematica rivisitata. *Vox Romanica*.

- Ferrari, A. and L. Zampese. 2000. *Dalla frase al testo. Una grammatica per l'italiano*. Bologna: Zanichelli.
- Firbas, J. 1964. On defining the theme in functional sentence analysis. *Travaux Linguistiques de Prague* 1: 267-80.
- Firbas, J. 1974. Some aspects of the Czechoslovak approach to problems of function sentence perspective. In F. Daneš (ed.), *Papers on Functional Sentence Perspective*. Prague and Paris: Academia, 11-37.
- Givón, T. (ed.). 1983. *Topic continuity in discourse: a quantitative cross-language study*. Amsterdam and Philadelphia: John Benjamins.
- Halliday, M.A.K. 1967a. Notes on transitivity and theme in English: Part I. *Journal of Linguistics* 3: 37-81.
- Halliday, M.A.K. 1967b. Notes on transitivity and theme in English: Part II. *Journal of Linguistics* 3: 199-244.
- Halliday, M.A.K. 1985. *An introduction to functional grammar*. London: Arnold.
- Hockett, Ch. F. 1958. *A course in modern linguistics*. New York: Macmillan.
- Korzen, I. 1998. Anafora e testo. Su codificazione anaforica e strutturazione testuale. In M. T. Navarro Salazar (ed.), *Italica Matritensia* (Atti del IV Convegno SILFI). Florence: Franco Cesati, 279-298.
- Korzen, I. 2001. Anafore e relazioni anaforiche: un approccio pragmatico-cognitivo. *Lingua Nostra* LXII, 3-4: 107-126.
- Lambrecht, K. 1994. *Information structure and sentence form. Topic, focus and the mental representations of discourse referents*. Cambridge: Cambridge University Press.
- Lombardi Vallauri, E. 1996. *La sintassi dell'informazione. Uno studio sulle frasi complesse tra latino e italiano*. Rome: Bulzoni.
- Lombardi Vallauri, E. 2002. *La struttura informativa dell'Enunciato*. Scandicci: La Nuova Italia.
- Lyons, J. 1980. *Sémantique linguistique*. Paris: Larousse.
- Mann, W.C. and S.A. Thompson. 1987. *Rhetorical structure theory. A theory of text organization*. Marina del Rey, CA: Information Sciences Institute.
- Mathesius, V. 1915. O passivu v moderní angličtině. *Sbornik filologický* 5: 198-220.
- Mortara Garavelli, B. 1979. *Il filo del discorso*. Turin: Giappichelli.
- Pasch, R., U. Brauße, E. Breindl and U.H. Waßner. 2003. *Handbuch der deutschen Konnektoren. Linguistische Grundlagen der Beschreibung und syntaktische Merkmale der deutschen Satzverknüpfen (Konjunktionen, Satzadverbien und Partikeln)*. Berlin and New York: Walter de Gruyter.
- Roulet, E., L. Fillietaz and A. Grobet. 2001. *Un modèle et un instrument d'analyse de l'organisation du discours*. Bern: Peter Lang.
- van Dijk, T.A. 1977. *Text and context*. London: Longman.

Sources of examples

- Berruto, G. 2003. Varietà diamesiche, diastratiche, diafasiche. In A.A. Sobrero (ed.), *Introduzione all'italiano contemporaneo. La variazione e gli usi*. Bari: Laterza, 37-92.
- Cresti, E. 2000. *Corpus di italiano parlato*. vol. II. Florence: Accademia della Crusca.
- De Mauro, T. 2006. *Parole di giorni lontani*. Bologna: Il Mulino.

- Ferrari, A. 1994. La linguistica del testo. In E. Manzotti and A. Ferrari (eds), *Insegnare italiano: principi, metodi, esempi*. Brescia: Editrice La Scuola, 43-73.
- Ferrari, A. 2006. La fonction informationnelle d'Appendice. De la dislocation à l'apposition à travers la composante informationnelle. *Cahiers Ferdinand de Saussure* 59: 55-86.
- Ferrari, A., L. Cignetti, A.-M. De Cesare, L. Lala, M. Mandelli, C. Ricci and E. Roggia. 2008. *L'interfaccia lingua-testo. Struttura e funzione dell'articolazione informativa dell'enunciato*. Alessandria: Edizioni dell'Orso.
- Ferrari, A. and A.-M. De Cesare. In press. La progressione tematica rivisitata. *Vox Romanica*.
- Ferrari, A. and L. Zampese. 2000. *Dalla frase al testo. Una grammatica per l'italiano*. Bologna: Zanichelli.
- Lavezzi, G., C. Martignoni, P. Sarzana and R. Saccani (eds). 1992. *Testi nella storia. La letteratura italiana dalle Origini al Novecento*, vol. 4, *Il Novecento*. Milan: Mondadori.
- Maraschio, N. and T. Poggi Salani. 2000. Presentazione degli Atti. In *Italia linguistica anno mille. Italia linguistica anno duemila* (Atti del XXXIV Congresso Internazionale della SLI), N. Maraschio and T. Poggi Salani (eds), I-VII, Rome: Bulzoni.
- Mengaldo, P.V. 1998. *Profili di critici del Novecento*. Torino: Bollati Boringhieri.
- Zampese, L. 2004. Aspetti semantico-testuali del gerundio modale in apertura di frase. In A. Ferrari (ed.), *La lingua nel testo, il testo nella lingua*. Turin: Istituto dell'Atlante Linguistico Italiano, 79-116.
- Zampese, L. 2005. La struttura informativa degli articoli di cronaca: natura e funzioni dell'Unità di Quadro. In A. Ferrari (ed.), *Rilievi. Le gerarchie testuali di alcuni tipi di testo*. Florence: Franco Cesati Editore, 173-216.

INTERACTION ENTRE LA SYNTAXE DES LEXÈMES ET LE SÉMANTISME DES PARTIES DU DISCOURS: NOM *VS.* ADJECTIF DE COULEUR EN JAPONAIS¹

Itsuko Fujimura

GSID, Nagoya University (Japan)

1. Introduction

Cette étude porte sur l'opposition entre les usages adjectival et nominal des termes de couleur qui qualifient ou déterminent un nom en japonais. Cette alternance s'observe seulement auprès de six unités lexicales de couleur. Ce sont d'abord les quatre termes chromatiques fondamentaux du japonais : *SHIRO* (blanc), *KURO* (noir), *AKA* (rouge) et *AO* (bleu), et les deux autres *KI-IRO* (jaune), *CHA-IRO* (marron/brun). Ce sont à la fois les formes substantives indépendantes et les racines adjectivales. Pour former l'adjectif on leur ajoute un suffixe adjectival *-I*. Tout les autres, à commencer par *MIDORI* (vert), sont uniquement substantifs et ils n'ont pas de forme adjectivale.

L'objectif de cette étude est double: d'abord la mise en lumière des facteurs qui conditionnent cette opposition, qui est à peine abordée dans la littérature linguistique japonaise, en recourant aux études statistiques et introspectives des 5000 exemples recueillis dans les corpus journalistique et littéraire de très grande taille, et ensuite, l'explication de ce phénomène du point de vue de la linguistique générale.

Cet article est composé de 6 sections. Dans la section 2 suivante, après avoir donné quelques informations nécessaires sur la langue japonaise, nous expliquerons cette question qui n'est pas si simple qu'il y paraît, en montrant l'insuffisance de l'unique travail précédemment mené sur cette question, qui repose sur la caractérisation fonctionnelle de l'adjectif et du nom. Nous poserons ensuite dans la section 3, une hypothèse globale pour rendre compte de cette opposition adjectivo-nominale, en nous appuyant sur les connaissances généralement admises par les linguistes dans le domaine de la typologie et des universaux du langage : adjectif comme qualité *vs.* nom comme entité. Après avoir présenté les corpus et les données sur lesquels notre discussion sera fondée dans la section 4, nous relèverons dans la section 5 de multiples facteurs qui conditionnent ce phénomène. Nous démonterons

parmi ces facteurs une distinction nette entre la propriété syntaxique de chaque lexème et les conditions sémantico-fonctionnelles des termes dans le discours. Ces dernières proviennent du sémantisme prototypique des parties de discours : adjectif vs. nom, tandis que la première est à expliquer par les faits qui se produisent sporadiquement dans l'histoire de chaque langue. Nous soulignerons à la fin que l'interaction entre ces deux facteurs qui sont indépendants l'un de l'autre, joue un rôle primordial dans ce phénomène qui est énigmatique à première vue.

2. Aperçu préliminaire

2.1 Grammaire

Les termes chromatiques de base en japonais sont classés en trois groupes² :

- 4 lexèmes chromatiques principaux :
SHIRO (blanc), *AKA* (rouge), *AO* (bleu), *KURO* (noir)
Ils sont adjectivables et substantivables.
- 2 lexèmes avec un suffixe *-IRO* "couleur":
KI-IRO (jaune), *CHA-IRO* (marron/brun)³
Ils sont aussi adjectivables et substantivables.
- Tous les autres :
MIDORI (vert), *MURASAKI* (violet), *HAI-IRO* (gris) etc.
Ils sont toujours substantifs. Ils ne sont pas adjectivables.

-I est un des deux suffixes adjectivaux à côté de *-NA*, qui sont obligatoires l'un ou l'autre dépendant du lexème, quand les adjectifs sont employés comme épithètes. *NO*, qui est un des mots le plus fréquemment employés en japonais, est la particule dont la fonction est la liaison de deux substantifs comme "*de*" français, "*of*" anglais. Le japonais ne dispose pas de la catégorie comparable aux prépositions dans les langues européennes indiquant les relations spatiales, d'appartenance etc.. Ce *NO* est de fait l'unique moyen pour lier les substantifs si ce n'est leur juxtaposition sans faire intervenir aucune particule. Dans les exemples suivants, *NO* sera glossé en GÉN (= génitif) pour la commodité. En japonais, la tête nominale, étant toujours mise à la fin d'un syntagme nominal, est postposée aussi bien à l'adjectif qu'au nom qui la modifie. L'opposition adjectivo-nominale est donc à observer tout simplement en tant qu'alternance entre les formes *-I* et *NO*⁴, bien que ces deux constructions : <Nom déterminant + Particule *NO* (=GÉN) + Tête nominale> et <Racine adjectivale + Suffixe adjectival *-I* + Tête nominale > soient syntaxiquement éloignées. Voici des exemples :

(1) COULEUR (-suffixe ADJ /-GÉN) + NOM

- *KURO (-I / -NO) ZUBON*
‘noir (-ADJ / *de*) pantalon’
[pantalon noir / pantalon de couleur noire]
- *KI-IRO (-I / NO) ZUBON*
‘jaune (-ADJ/*de*) pantalon’
[pantalon jaune/ pantalon de couleur jaune]
- *MIDORI (*-I / NO) ZUBON*
‘vert (*-ADJ/ *de*) pantalon’
[pantalon vert]

Les substantifs japonais sont syntaxiquement et morphologiquement autonomes. Ils n’ont pas de désinences flexionnelles, ni n’exigent la présence d’un autre mot tel qu’un article. Le japonais n’a pas la catégorie grammaticale des déterminants nominaux. Il y a seulement des démonstratifs dont l’usage est facultatif. Dans cet article, toutes les gloses françaises des exemples sont données sans déterminants, parce que leur degré de détermination est ambigu par nature. Il n’en reste pas moins que la notion de détermination n’est pas exclue du japonais. On verra que notre question, le choix entre le substantif et l’adjectif, peut fonctionner comme indicateur de la détermination.

En japonais, le nombre des adjectifs est restreint. Ils sont beaucoup moins nombreux que dans les langues européennes⁵. La raison majeure en est sans doute que les adjectifs japonais sont les qualificatifs (exprimant une qualité) et non les relationnels. Cette langue a en effet un bon nombre d’adjectifs qualificatifs tels que *CHIISA-I* (petit), *YO-I* (bon), *KAWAI-I* (joli), *KIREI-NA* (beau), *JYUU-NA* (libre), *SHIZUKA-NA* (calme)⁶. Par contre, elle n’a pas d’adjectifs relationnels tels que *présidentiel* dans *élection présidentielle*, *hivernal* dans *froid hivernal*. Pour les dire, on emploie le substantif au lieu de l’adjectif comme le montrent les exemples (2) et (3)⁷. Cette construction est exactement la même que celle qu’on emploie avec les termes de couleur substantivaux.

(2) *FUYU (*-I/ NO/ *Ø) SAMUSA*

‘hiver (*-ADJ / -GÉN/ *Ø) froid’
[froid hivernal, froid d’hiver]

(3) *DAITORYO (*-I/ NO/ Ø) SENKYO*

‘Président (*-ADJ / -GÉN/ Ø) élection’
[élection présidentielle, élection du Président]

2.2 Usages

Nous avons déjà fait remarquer que tous les termes de couleur n'ont pas la forme adjectivale. Mais même avec les six premiers termes de couleur, qui sont potentiellement à la fois adjectivables et substantivables, le choix entre les deux n'est pas libre, comme l'indique l'exemple suivant.

(4) Devant un ciel bleu, des nuages blancs et des montagnes brunes à l'arrière plan, une jeune fille aux yeux verts portant un tee-shirt jaune et un pantalon noir se tient debout avec une rose rouge à la main.

ciel bleu:	<i>AO</i> (-I (ADJ)/ *-NO(GÉN)) <i>SORA</i>
nuage blanc:	<i>SHIRO</i> (-I (ADJ)/ *-NO(GÉN)) <i>KUMO</i>
montagne brune:	<i>CHA-IRO</i> (-I (ADJ)/ -NO(GÉN)) <i>YAMA</i>
yeux verts:	<i>MIDORI</i> (* -I (ADJ)/ -NO(GÉN)) <i>ME</i>
tee-shirt jaune:	<i>KI-IRO</i> (-I (ADJ)/ -NO(GÉN)) <i>T-SHATSU</i>
pantalon noir :	<i>KURO</i> (-I (ADJ)/ -NO(GÉN)) <i>ZUBON</i>
rose rouge:	<i>AKA</i> (-I (ADJ)/ -NO(GÉN)) <i>BARA</i>

L'usage nominal de *AO* (bleu) et *SHIRO* (blanc) dans “ciel bleu” et “nuages blancs” est inacceptable. À la limite, ces usages donnent l'impression de la description d'un paysage artificiel, un collage composé de feuilles de papier artificiellement colorées, tel qu'on rencontre à une école maternelle par exemple. Par contre, l'usage nominal de *CHA-IRO* (marron / brun) dans “montagne brune” ne pose pas de problème en tant que description de toutes les sortes de montagnes soit naturelles soit artificielles. En outre, tous ces termes chromatiques sauf *MIDORI* (vert) peuvent être aussi bien adjectivaux que nominaux, lorsqu'ils qualifient une rose naturelle ou un vêtement, un tee-shirt, un pantalon etc. Cette hétérogénéité demande une explication. Il n'y avait cependant jusqu'ici, autant que nous sachions, qu'un seul travail qui l'avait envisagée.

2.3 Étude précédente : Sawada (1992)

L'unique travail publié en japonais qui traite cette opposition adjectivo-nominale des termes de couleur japonais est Sawada (1992). Comme l'indique le titre de l'article : “The Indicatory Function in Nouns Compared with Restrictivity and Descriptivity of Adjectives--From an Analysis of Selectional Factors in the Basic Color Words”, l'auteur essaye d'expliquer cette opposition en étudiant la fonction des termes de couleur. Si leur fonction est la distinction de l'objet modifié, le substantif est utilisable, tandis que si elle est la description, l'adjectif est à choisir. Voici les exemples cités dans l'article. Le (5) est l'exemple de l'usage distinctif et le (6), celui de l'usage descriptif.

(5) *Sono AO (-I/-NO) KOPPU o totte-kudasai.* (+distinctif)

‘ce bleu (-ADJ/-GÉN) verre’

[Prenez ce verre bleu !]

(6) *Teeburu no ue ni AO (-I/?-NO) KOPPU ga miemasu.* (+ descriptif)

‘bleu (-ADJ/?-GÉN) verre’

[Je vois un verre bleu sur la table.]

Quand on distingue un objet d’un autre par une différence de couleur, on peut employer aussi bien le nom que l’adjectif pour indiquer cette couleur, alors qu’on préfère l’adjectif quand on qualifie un objet avec la couleur. Certes l’auteur a contribué à résoudre la question. Mais cette contribution est limitée, parce que premièrement l’auteur n’a pas étudié les usages réels, que deuxièmement, par conséquent, elle n’a jamais fait attention au problème lexical des termes de couleur, et que troisièmement elle a laissé tomber de nombreuses questions intéressantes qu’on peut observer uniquement dans l’usage épithète (c’est surtout le cas des exemples (7), (8) et (21)), en étudiant principalement l’usage prédicatif de ces termes. Nous partageons l’avis de Beck (2002, 83-85) et Wirtzbicka (1988, 484) que l’adjectif non-marqué est épithète. L’usage épithète n’est pas dérivé de l’usage prédicatif. Ce dernier est en effet beaucoup moins fréquent que le premier, un cinquantième dans nos corpus. Nous croyons que l’explication donnée dans Sawada (1992), même si elle n’est pas fautive, est loin d’être suffisante pour rendre compte de la totalité complexe de ce problème.

Voyons des contre-exemples :

(7) Portant un vêtement de deuil noir, Kyoko, l’air épuisée, est restée silencieuse.

KURO (-I/-NO) MO-FUKU (-distinctif)

‘noir (-ADJ/-GÉN) vêtement de deuil’

(8) Ne m’enlève pas mes cheveux noirs. Enlève seulement les cheveux blancs.
(+distinctif)

KURO (-I/-NO) KE*

SHIRO (-I/-NO) KE*

‘noir (-ADJ/*-GÉN) cheveux’

‘blanc(-ADJ/*-GÉN) cheveux’

Dans l’exemple (7), l’usage substantif : *KURO -NO* est tout à fait naturel, bien que le terme ne remplisse pas la fonction distinctive. Au Japon, la couleur de vêtement de deuil est toujours noire sans exception. Dans l’exemple (8), par contre, l’usage substantif, *KURO-NO*, *SHIRO-NO* est inacceptable, malgré le contexte qui indique la fonction distinctive que remplissent ces termes de couleur. La fonction des termes de couleur ne correspond pas toujours à leur forme grammaticale. Il faut intégrer ces deux contre-exemples dans la totalité de la question.

Il faut tenir compte aussi d’autres facteurs que des fonctionnels. Surtout la question de la référence et la particularité lexicale de chaque terme sont importantes,

d'autant plus qu'elles ne sont mises en évidence dans aucune étude déjà existante. Par ailleurs, l'exemple (7) dans lequel l'usage nominal des termes chromatiques les plus fondamentaux joue le rôle "descriptif" est une question épineuse.

3. Opposition adjectivo-nominale et termes de couleur

Dans cette section, nous présenterons une hypothèse qui permettrait de résoudre une moitié de notre problème. Cela concerne le rapport entre la forme et le contenu linguistiques, à savoir la relation entre l'opposition grammaticale des parties de discours : adjectif *vs.* nom et l'opposition sémantico-fonctionnelle : OBJET *vs.* PROPRIÉTÉ, qui lui correspond.

Le nombre de linguistes qui s'intéressent aux universaux linguistiques, surtout cognitivistes et typologues, s'accordent sur les sémantismes prototypiques des noms et des adjectifs dans les langues (cf. Wierzbicka 1988; Croft 1991; Goes 1999; Beck 2002; Whittaker 2002)⁸. D'après eux, le nom prototypique est un objet discret et sa fonction est la référence à cet objet; l'adjectif prototypique désigne une propriété continue et sa fonction est la modification d'un autre objet à l'aide de cet adjectif. Il va sans dire que la catégorie grammaticale est définie comme classe des mots qui partagent les mêmes comportements morphosyntaxiques, sans rapport ni avec leurs sens ni avec leurs fonctions. Mais dans le cas où le locuteur peut choisir librement une forme entre deux catégories, son choix est certainement influencé par le sémantisme prototypique des catégories.

Or, les couleurs ont un double aspect (Beck 2002, 54; Tucker 1999, 149). Primo, les couleurs peuvent être considérées comme objets, en raison de leur visibilité, leur non-gradualité sur l'échelle linéaire, leur stabilité dans le temps. Un cas prototypique de la couleur comme objet serait un échantillon chromatique dans le nuancier⁹. Dans ce cas extrême, la couleur semble exister indépendamment sans aucun support, ce qui est caractéristique prototypique de l'objet sémantique. Secundo, la couleur est une qualité ou une propriété d'un autre objet. La couleur blanche d'une "petite maison blanche" ne peut pas exister indépendamment de "la maison", comme c'est le cas pour toutes les autres qualités de "la maison" qui ne peuvent pas exister sans celle-ci. Nous supposons que la couleur comme concept linguistique doit avoir ces deux types de valeur opposés dans son sémantisme. Nous considérons également que ces deux pôles ne sont pas dissociés mais qu'il y a un continuum entre les deux.

La table 1 suivante montre le cadre théorique de départ de ce travail. La couleur conceptualisée comme objet a la tendance d'être exprimée comme nom parce que le nom correspond prototypiquement à l'objet et que la fonction du nom est la référence à un objet. Par contre, la couleur conceptualisée comme qualité a la tendance d'être exprimée comme adjectif, parce que l'adjectif correspond

prototypiquement à une qualité et que la fonction de l’adjectif est la modification d’un autre objet. Nous devons les deux dernières parties de cette table à Croft (1991). On peut dire que ce schéma est plus ou moins accepté par les linguistes universalistes.

Notre hypothèse de départ est donc que le terme de couleur a plus de probabilité d’être exprimé en tant qu’adjectif, s’il a plus de caractéristiques sémantico-fonctionnelles dérivées de la notion de Propriété, et qu’il a plus de chances d’être exprimé en tant que nom, s’il a plus de caractéristiques sémantico-fonctionnelles dérivées de la notion d’Objet. Toutefois, nous verrons à la suite des recherches de corpus que les facteurs sémantico-fonctionnels qui y jouent sont beaucoup plus nombreux et variés que ceux qui sont cités dans la table 1. Nous vérifierons qu’ils sont toujours les facteurs provenant de l’opposition entre l’Objet et la Propriété et non les facteurs sporadiquement intervenants. Nous proposerons enfin une rectification à notre hypothèse de départ, dans le but de mieux appréhender la totalité de ce phénomène.

Table 1 : Opposition adjectivo-nominale et couleurs (Croft 1991, 55 et 65)

Couleur	comme objet	comme qualité
	↑ ↓	↑ ↓
	Noun	Adjective
Semantic class	object	property
Pragmatic function	reference	modification
	↑ ↓	↑ ↓
	Objects	Properties
Valency	0	1
Stativity	state	state
Persistence	persistent	persistent
Gradability	nongradable	gradable

4. Corpus et données

Pour mener à bien cette étude, nous avons établi une base de données constituée d’environ 5000 exemples des termes de couleur recueillis dans les corpus journalistique et littéraire suivants :

- *Journal Mainichi (Mainichi shimbun)*¹⁰, janvier au juin 1999 : 125.6 millions de morphèmes au total¹¹.
- 42 romans contemporains (après 1950) : 23.5 millions de morphèmes au total.

Nous avons d'abord automatiquement recueilli les exemples en cherchant *KURO* (noir), *AKA*(rouge), *SHIRO* (blanc), *AO* (bleu), *KI-IRO* (jaune), *CHA-IRO* (marron/brun) et *MIDORI* (vert) avec *-I* et *-NO*. Nous avons ensuite manuellement éliminé tous les exemples qui ne sont pas pertinents suivant les critères morphosyntaxique, sémantique et syntaxique suivants, afin de ne conserver que ceux qui sont appropriés à notre recherche:

a. condition morphosyntaxique

Les séquences suivantes ont été retenues :

(*KURO* (noir), *AKA*(rouge), *SHIRO* (blanc), *AO* (bleu), *KI-IRO* (jaune), *CHA-IRO* (marron/brun)) *MIDORI* (vert) + (*-I* / *-NO*) + (un ou plusieurs mots) + NOM (= objet ayant de la couleur)

b. condition sémantique

Nous avons gardé seulement les exemples indiquant une couleur proprement dite. On a donc exclu des expressions idiomatiques : *KI-IRO-I KOE*, jaune-ADJ voix, (voix jaune) “voix perçante”, des métaphores ou des métonymies : *KURO-I WARAI*, noir-ADJ rire, “rire noir”, *KURO* (noir (nominal)) signifiant “criminel”, *AKA* (rouge (nominal)) “communiste”, *SHIRO-NO SHORI*, blanc-GÉN victoire, (victoire du blanc) “victoire des blancs dans l'échec”, des noms propres : *AKA-NO HIROBA*, rouge-GÉN place, “Place Rouge (à Moscou)”, expressions métachromatiques : *KURO-I IRO*, noir-ADJ couleur, “couleur noire” etc.

c. condition syntaxique

Nous avons retenu seulement des exemples dans lesquels l'alternance entre le nom et l'adjectif n'est pas morphosyntaxiquement contrainte. Il a été donc exclu des adjectifs de couleur composés: *AO-JIRO-I*, bleu-blanc-ADJ, “pâle”, des noms de couleur composés: *SHIRO-KURO-NO*, blanc-noir-GÉN “noir et blanc”, des noms de couleur modifiés par un adjectif : *FUKAI AO-NO HITOMI*, profond bleu-GÉN prunelles, “prunelles bleu foncé” et des noms de couleur coordonnés : *AKA TO SHIRO NO HATA*, rouge et blanc-GÉN drapeau, “drapeau rouge et blanc”.

5. Résultats

5.1 Occurrences générales

Nous montrons tout d'abord la figure indiquant les occurrences des termes de couleur et celles de l'adjectif et du nom dans les deux genres de texte. Nous voyons premièrement que l'usage adjectival est majoritaire excepté dans les cas de *CHA-*

IRO (marron/brun) et de *MIDORI* (vert). Nous observons deuxièmement que dans le texte littéraire, les termes de couleur sont 10 fois plus fréquents que dans le texte journalistique. Nous relevons troisièmement que la même tendance lexicale règne dans les deux genres de texte : *SHIRO* (blanc) y est le plus fréquemment employé et les autres termes suivent dans l'ordre décroissant :

SHIRO (blanc) > *KURO* (noir) > *AKA* (rouge) > *AO* (bleu) > *KI-IRO* (jaune) > *MIDORI* (vert) > *CHA-IRO* (marron/brun)

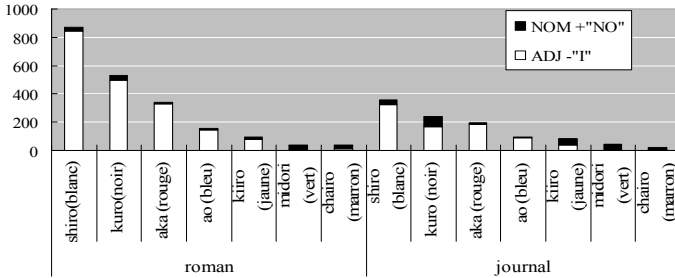


Figure 1. Occurrences des termes de couleur dans les deux genres de texte et ADJ vs. NOM (en nombre réel)

On peut supposer que cette tendance est la caractéristique de l'usage des termes de couleur de base en japonais. Nous faisons également remarquer que la disparité de fréquence entre les termes de couleur est plus importante que notre attente intuitive. Le pourcentage de chaque couleur perçue dans le monde réel nous paraît loin de correspondre à celui de chaque terme de couleur dans le texte. On trouve une discordance entre le monde de référence et le monde linguistique même dans ce domaine relativement objectif.

5.2 Propriété syntaxique des lexèmes

La figure 2 montre avec plus de clarté le taux de l'adjectif et du substantif auprès de chaque terme de couleur dans l'ensemble des corpus.

Avec les 4 premiers termes, dits fondamentaux suivant les critères de Berlin and Kay (1969), l'usage nominal est rare. Pour le dernier : *MIDORI* (vert), l'usage adjectival n'existe pas, étant donné que cet usage est structurellement absent. Les deux autres termes sont entre les deux pôles. Pour *CHA-IRO* (marron/brun), l'usage nominal est majoritaire, tandis que pour *KI-IRO* (jaune), il est minoritaire. Cette hétérogénéité de répartition entre les 6 termes de couleur, qui n'a jamais été examinée dans la littérature linguistique japonaise, demande une explication. La recherche de cette question n'est d'ailleurs possible qu'avec la méthode quantitative basée sur les grands corpus, qu'est la nôtre.

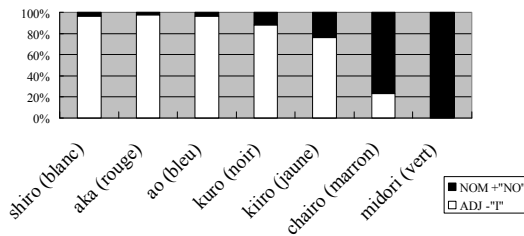


Figure 2. ADJ vs. NOM et lexèmes de couleur

Notre hypothèse énoncée dans la section 3 et représentée dans la table 1 n'est pas suffisante pour aborder ce phénomène, parce qu'il n'y a pas de raison de supposer que les quatre premières couleurs : blanc, rouge, bleu et noir sont enclines à être conceptualisées comme qualité et la dernière : marron comme objet. L'hypothèse émise dans Sawada (1992) fondée sur les fonctions distinctive et descriptive ne l'explique pas non plus. Suivant son hypothèse, on devrait aboutir à la conclusion invraisemblable que les premiers quatre termes de couleur jouent la fonction descriptive dans presque tous les cas et que *CHA-IRO*, la fonction distinctive dans la plupart des cas. D'après nous, c'est la question de particularité morphosyntaxique de chaque lexème qui n'est pas attribuable à ses caractéristiques sémantico-fonctionnelles¹².

5.3 Conditions sémantico-fonctionnelles des mots dans le discours

5.3.1 *Référent chromatique : couleur dans la nature vs. couleur artificielle*

Parmi les conditions sémantico-fonctionnelles, une des plus remarquables est que l'usage substantival des quatre premiers lexèmes est très limité et même impossible en cas de référence à une couleur dans la nature, tandis qu'il est possible pour désigner une couleur artificielle. Concernant ces quatre premiers termes de couleur, il y a seulement 5 occurrences nominales contre 1313 adjectivales pour les couleurs naturelles, tandis que 125 nominales contre 965 adjectivales pour les couleurs artificielles. L'affinité entre la couleur naturelle et l'adjectif et celle entre la couleur artificielle et le nom sont statistiquement significatives ($X^2=143.631$, $dl = 1$, $p < .01$). L'étude introspective faite par nous-mêmes indique la même tendance¹³. L'usage nominal de *AKA* (rouge) est acceptable et naturel dans (9), alors qu'il ne l'est pas dans (10). Voir aussi les exemples (4), (7) et (8).

- (9) AKA (-I / -NO) KOOTO ga miemasu
 'rouge (-ADJ / -GÉN) manteau'
 [Je vois un manteau rouge]

- (10) AKA (-I / *-NO) TAIYO ga miemasu
 ‘rouge (-ADJ / *-GÉN) soleil’
 [Je vois le soleil rouge]

Cette tendance n’est cependant pas la règle. Par exemple, dans l’exemple (11), l’usage substantif n’est pas complètement exclu, afin d’indiquer la couleur des roses naturelles. On verra plus bas que si les noms modifiés ont une extension moins large, ils ont une affinité plus importante avec les termes de couleur substantifs.

- (11) AKA (-I / !?-NO) BARA ga miemasu
 ‘rouge (-ADJ / -GÉN) rose’
 [Je vois des roses rouges]

Pour les deux autres termes : *KI-IRO* (jaune) et *CHA-IRO* (marron/brun), l’usage substantif est plus fréquent et plus acceptable en référant à la couleur naturelle que les autres couleurs de base, comme l’indique l’exemple (12) relatif aux feuilles d’arbre naturelles.

- (12) Des feuilles marron, jaunes, noires, rouges, vertes
 (CHA-IRO/ KI-IRO/ *KURO/ *AKA/ *AO/ MIDORI)-NO HAPPA
 ‘(marron/ jaune/ *noir/ *rouge/ *bleu/ vert) -GÉN feuille’
 (CHA-IRO / KI-IRO / KURO / AKA / AO/ *MIDORI)-I HAPPA
 ‘(marron/ jaune/ noire/ rouge/ *vert) -ADJ feuille’

Cependant dans certains contextes, tout en indiquant la couleur naturelle, le *CHA-IRO* (marron/ brun) nominal est plus acceptable que le *KI-IRO* (jaune) nominal. C’est le cas dans l’exemple (13), où l’on parle des yeux naturels des êtres humains. Le substantif n’est accepté que pour *CHA-IRO*.

- (13) Des yeux bruns, jaunes, noirs, rouges, bleus, verts
 (CHA-IRO/ ?KI-IRO/ *KURO/ *AKA/ *AO/ MIDORI)-NO ME
 ‘(brun/ ?jaune/ *noir/ *rouge/* bleu/ vert) -GÉN yeux’
 (CHA-IRO / KI-IRO / KURO / AKA / AO/ *MIDORI)-I ME
 ‘(brun/ jaune/ noir/ rouge/ bleu / *vert) -ADJ yeux’

Si l’on parle des dents naturelles des êtres humains comme dans l’exemple (14), *CHA-IRO* et *KI-IRO* nominaux sont cette fois tous les deux inacceptables. Si l’on emploie la forme nominale, les dents ne sont plus naturelles mais ce sont de fausses dents artificiellement colorées en jaune ou brun comme dans l’exemple (15).

- (14) Yorikuni a ri en montrant ses dents jaunes/ brunes.
 (KI-IRO / CHA-IRO) (-I / *-NO) HA
 ‘(jaune / brun)-ADJ dents’

- (15) Yorikuni a ri en montrant des dents colorées jaunes/ brunes de sa poche.
 (KI-IRO / CHA-IRO) -NO HA
 '(jaune / brun)-GÉN dents'

Il est certain que le *CHA-IRO* (marron/brun) a une affinité la plus forte avec l'usage substantival parmi les termes ci-examinés. Malgré cela, cet usage n'est pas accepté dans (14), lorsqu'il réfère à une couleur naturelle non homogène mais complexe telle que la couleur des dents non esthétiques, brunies par des cigarettes.

Voyons la figure 3, qui indique la relation entre la distinction de couleurs naturelle et artificielle et le taux de NOM auprès de chaque lexème de couleur. On y trouve, d'une part, une tendance claire de la préférence du substantif ou de l'adjectif suivant le lexème de couleur¹⁴. On y relève d'autre part une autre tendance évidente que sont l'affinité entre l'adjectif et la couleur naturelle et celle entre le substantif et la couleur artificielle. L'isolé dans la figure est *MIDORI* qui est constamment substantif sans exception. Ces deux tendances sont indépendantes d'une de l'autre. Avec les quatre premiers termes de base, la couleur naturelle indiquée par le nom est tout à fait exceptionnelle : 2 exemples sur 628 pour blanc, 1 sur 220 pour rouge, 1 sur 149 pour bleu et 1 sur 319 pour noir. La moitié de ces exceptions sont relatives aux couleurs des espèces de fleurs. C'est le cas des exemples (11) et (22).

La raison pour laquelle la couleur artificielle est exprimée plutôt comme substantif nous semble qu'elle a les caractéristiques qui permettent de la conceptualiser comme objet. C'est la couleur simple, plate, statique et définie, qu'on peut trouver comme repère dans les échantillons chromatiques. Par contre, la raison pour laquelle la couleur dans la nature est exprimée plutôt comme adjectif est qu'elle a les caractéristiques permettant de la conceptualiser comme qualité. C'est la couleur complexe, profonde, dynamique, instable, et difficile à définir sans repère précis.

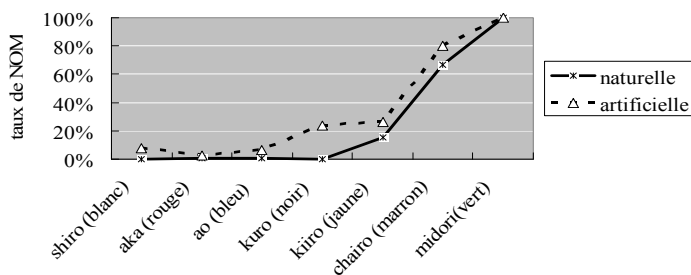


Figure 3. Taux de NOM, couleur naturelle vs. artificielle et lexèmes de couleur

5.3.2 Relation entre le terme de couleur et l'objet modifié

Dans cette section, nous examinerons la relation entre la couleur et l'objet ayant de la couleur. La figure 4 montre les pourcentages de l'adjectif et du substantif suivant la nature des objets.

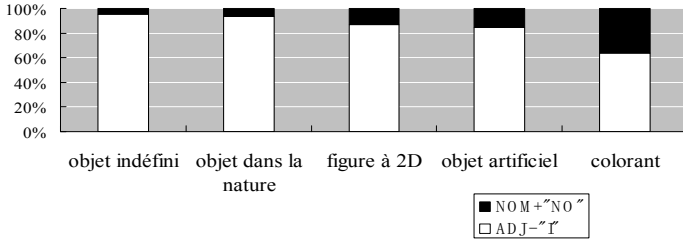


Figure 4. ADJ vs. NOM et la nature de l'objet ayant de la couleur

On remarquera d'abord que le substantif est le plus fréquemment employé, quand la couleur réfère à celle du colorant, à savoir celle du contenu de l'objet et non à celle de son apparence. Le substantif est surtout préféré, lorsque l'objet est un outil concret dont la surface est artificiellement colorée, stylo, feutre, spray par exemple. Probablement, parce que l'adjectif chromatique est interprété comme expression de la couleur de la surface et que le nom chromatique, comme celle de la couleur du contenu qui est invisible de l'extérieur.

- (16) Examiner en entourant d'un carré avec un stylo rouge.

AKA (??-I /-NO) BALL PEN

'rouge (??-ADJ/ -GÉN) stylo'

- (17) Rectifier la couleur des cheveux avec un spray colorant noir.

KURO (-I /-NO) KAMIZOME-SPRAY

'noir (??-ADJ/ -GÉN) spray colorant noir'

Il nous semble que ce phénomène est explicable par le fait que la relation entre le contenant et le contenu est dans l'ordre plus objectif que la description de l'apparence chromatique d'un objet.

Ensuite, les noms indiquant les figures à deux dimensions : ligne, figure etc. ont plus d'affinité avec les substantifs, comme les exemples (18) et (19). Même si ces noms réfèrent aux figures constituées de la matière naturelle comme de l'eau, l'usage substantival est acceptable, comme l'indique l'exemple (20). Ce qu'on peut observer dans ce genre de relation est la combinaison de la forme et de la couleur. Ce sont les deux propriétés saillantes des objets quand on les qualifie. Elles n'ont pas de relation de subordination entre elles et sont indépendantes l'une de l'autre. La combinaison de ces deux propriétés est donc dans l'ordre abstrait, voir logique. Ce serait la raison pour laquelle dans cette relation la couleur est plus fréquemment exprimée en tant que substantif que dans la relation entre un objet substantiel et son apparence chromatique.

- (18) Le logo blanc apporte une note de discrétion.

SHIRO (-I/-NO) LOGO
 ‘blanc (-ADJ/-GÉN) Logo’

- (19) Sur la couverture grise de ce livre, sont imprimés le titre et des traits noirs.

KURO (-I/-NO) SEN
 ‘noir- (-ADJ/-GÉN) ligne’

- (20) une ligne blanche de l’eau vivement jaillie

SHIRO (-I/-NO) SEN
 ‘blanc (-ADJ/-GÉN) ligne’

5.3.3 Extension des noms modifiés

Nous démontrons par la suite que l’adjectif de couleur a plus d’affinité avec des objets exprimés par des termes ayant plus d’extension, à savoir termes plus vagues, plus génériques. À l’opposé, le substantif de couleur en a avec des termes ayant moins d’extension, à savoir des termes plus spécifiques, plus informatifs. Dans la figure 4 ci-dessus, cette question est indiquée avec la colonne “objet indéfini”. Au cas où le nom modifié est un pronom indéfini, le nom de couleur n’est pas en principe acceptable.

- (21) On appelle ce jour-là “le jour blanc” car les hommes offrent aux femmes quelque chose de blanc ce jour-là.

*SHIRO (-I / *-NO) MONO*
 ‘blanc (-ADJ / *-GÉN) chose’

Dans l’exemple (21), pour un nom générique : *MONO* (chose), l’usage du substantif est inacceptable dans ce contexte, malgré les conditions qui favorisent l’usage du substantif. D’abord, cette “chose” se réfère à un objet artificiel, parce que le “jour blanc” a été inventé pour une raison commerciale pour mieux écouler des marchandises. Et ensuite le terme de couleur remplit la fonction non descriptive mais déterminative parce que les cadeaux offerts aux femmes doivent être blancs et non d’une autre couleur. Parmi les termes de couleur, comme prévu, l’usage substantival est acceptable pour *CHA-IRO (CHA-IRO-NO MONO* (marron-GÉN chose)), à côté de l’usage adjectival (*CHA-IRO-I MONO* (marron-ADJ chose)) qui est également acceptable. Et si *MONO* (chose) est remplacé par *SCARF* (foulard), l’usage du *SHIRO-NO SCARF* (blanc-GÉN foulard) est parfaitement acceptable.

Il nous semble que moins l’extension de nom modifié est grande, plus l’usage substantif des termes de couleur est facilement accepté. Les exemples (22) et (23) indiquent qu’avec les noms génériques tels que fleur, vêtement, l’usage de l’adjectif est exigé en tant que modifiant de ces noms. Si le nom est plus déterminé avec plus de compréhension, l’usage du substantif comme modifiant est plus facilement accepté. Cela nous semble explicable en disant que l’adjectif a en général plus

d'extension et moins de compréhension que le substantif. Quand le nom tête n'est pas déterminé, on est obligé d'employer un modifiant avec une certaine marge qui doit s'adapter à l'indétermination de ce nom tête.

- (22) ?? *AKA-NO* (-GÉN) *HANA* [fleur rouge]
SHIRO-NO (-GÉN) *YURI* [lis blanc]
AKA-NO (-GÉN) *DARK LADY* [The dark lady rouge (nom d'espèce de rose)]

- (23) ?*AO-NO* (-GÉN) *FUKU* [vêtement bleu]
AO-NO (-GÉN) *SEI-FUKU* [uniforme bleu]
SHIRO-NO (-GÉN) *WEDDING DRESS* [robe de mariage blanche]

5.3.4 *Genre de texte (journal/ roman) : intention de l'énonciateur*

Nous allons enfin examiner la relation entre le genre de texte et l'opposition adjectivo-nominale. Quant aux 6 termes de couleurs qui ont deux formes nominale et adjectivale, il y a seulement 95 occurrences nominales contre 1882 adjectivales dans les romans, tandis que 143 nominales contre 803 dans le journal. Bien que dans les deux textes, l'usage adjectival soit majoritaire, l'affinité entre les textes littéraires et l'adjectif et celle entre les textes journalistiques et le nom sont statistiquement significatives ($X^2=90.9512$, $dl = 1$, $p < .01$)¹⁵.

Dans les journaux, la fonction des termes de couleur est souvent la détermination de l'objet modifié. Ils ajoutent une information de plus pour préciser et mieux identifier le référent de l'objet évoqué par le nom. On pourrait dire que la fonction principale et fondamentale des journaux est de transmettre des informations précises aux lecteurs. Les termes de couleur qui fonctionnent comme classificateur de l'objet sont bien appropriés à cet objectif. Par contre, dans les romans, la fonction des termes de couleur est plus souvent la qualification de l'objet modifié. Il transmet l'attitude du locuteur à l'égard du nom dénoté par le nom. D'après Wierzbicka (1988, 484), "Adjectives, which stand for single features, can be freely used to enrich the image evoked by the noun". Elle caractérise la fonction prototypique de l'adjectif comme suit (Wierzbicka 1988, 488):

wanting to cause you to thinking of it
 in the way I am thinking of it
 I say: imagine [ADJ NOUN]

On pourrait dire approximativement que dans les romans l'écrivain décrit un monde imaginaire et la transmet les émotions afférentes. Les termes de couleur qui fonctionnent comme enrichissant l'image sont bien conformes à ce but.

Voyons les exemples recueillis à partir des corpus. Les exemples (24) et (25) parlent tous les deux d'une modèle de voiture japonaise appelée Skyline. Dans (24), qui est journalistique, la fonction du terme de couleur est la sous-catégorisation de la voiture. À l'opposé, dans (25), qui est littéraire, celle-ci est la description non

restrictive de la voiture. On peut trouver la même différence entre (26) et (27), et entre (28) et (29).

- (24) Comprenant que le beau-frère était l'assassin d'après la plaque d'immatriculation de la Skyline noire échappée, la police s'est lancée à sa poursuite. (journal)
KURO-NO SKYLINE : 'noir-GÉN Skyline'
- (25) Lorsque le feu est passé au vert, la Skyline blanche s'est éloignée de ma vue dans un très fort bruit de pot d'échappement et de musique de Bob Dylan. (Murakami, H.. 1985. *La fin du monde et le pays des merveilles sans merci*)
SHIRO-I SKYLINE: 'blanc-ADJ Skyline'
- (26) Selon l'enquête de la police, c'était le cadavre d'un homme adulte. Il portait un imperméable, une combinaison marron, des bottes noires, et des gants de caoutchouc jaunes sur lesquels apparaissaient des inscriptions en coréen. (journal)
CHA-IRO-NO SAGYO-FUKU : 'marron -GÉN combinaison'
KURO-NO NAGA-GUTSU : 'noir -GÉN bottes'
KI-IRO-NO GOMU-TEBUKURO: 'jaune-GÉN gants de caoutchouc'
- (27) À ce moment-là, le chef est arrivé portant un nœud papillon, une toque blanche, une chemise blanche, une veste blanche et un pantalon blanc, et des bottes noires; il a poussé la table à roulettes et nous a questionnés avec politesse. (Inoue, H.. 1970. *Bun et Fun*)
SHIRO-I KOKKUBO : 'blanc-ADJ toque'
SHIRO-I SHATSU : 'blanc -ADJ chemise'
SHIRO-I UWAGI: 'blanc-ADJ veste'
SHIRO-I ZUBON: 'blanc-ADJ pantalon'
KURO-I NAGA-GUTSU: 'noir-ADJ bottes'
- (28) On a appelé cette fleur Zades (plante de méditation zen) du fait que l'apparence de la fleur jaune dans la corolle de couleur rouge violacé rappelle la position de méditation des bonzes dans le temple. (journal)
KI-IRO-NO HANA : 'jaune-GÉN fleur'
- (29) Est-ce à cause de la douceur exceptionnelle de cet hiver, les fleurs qui fleurissent au printemps se sont trompées de saison et ont fleuri à la fin de l'année. C'était le cas du fuchsia, dont plusieurs fleurs jaunes étaient écloses dans la haie de bambou devant la maison. (Tachihara, M.. 1969. *Voyage en hiver*)
KI-IRO-I HANA : 'jaune-ADJ fleur'

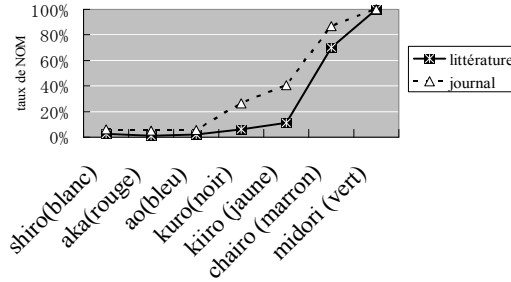


Figure 5. Taux de NOM, genre de texte et lexèmes de couleur

La figure 5 montre la relation entre le genre de texte et le taux de NOM auprès de chaque lexème de couleur. Elle indique d'une part une tendance d'affinité entre l'adjectif et le littéraire et entre le substantif et le journalistique, et d'autre part une autre tendance de préférence du substantif ou de l'adjectif suivant le lexème de couleur. Cette figure est semblable à la figure 3 ci-dessus. Dans la plupart des cas, les quatre premiers termes de couleur sont employés en tant qu'adjectif sans rapport direct avec le style textuel. Nous trouvons tout de même plus d'usages substantivaux dans les journaux que dans les romans. Le *CHA-IRO* (marron/brun) est pour la plupart des cas substantif, mais il prend cette forme plus souvent dans les journaux que dans les romans. Le *MIDORI* (vert) ne connaît qu'une seule forme substantivale.

6. Interaction entre la syntaxe des lexèmes et le sémantisme des parties du discours

L'étude statistique basée sur les corpus de grande taille et l'étude introspective nous ont permis tout d'abord de montrer parmi les conditions déterminant le choix entre l'adjectif et le nom, une distinction entre la propriété syntaxique des lexèmes (= type) et les conditions sémantico-fonctionnelles des termes dans le discours (= token), comme l'indique la table 2. *SHIRO* (blanc), *AKA* (rouge), *AO* (bleu) et *KURO* (noir) sont le plus couramment utilisés en tant qu'adjectifs, *CHA-IRO* (marron/ brun) l'est en tant que nom et *KI-IRO* (jaune) occupe une place intermédiaire. Ce qui est bien représenté dans les figures 3 et 5.

Les conditions sémantico-fonctionnelles relevées jusqu'ici sont aussi présentées dans la table 2. Ce sont les conditions relatives à la référence (§5.3.1), à la relation entre le modifiant et le modifié (§5.3.2), à l'extension du modifié (§5.3.3), à l'intention de l'énonciateur (§5.3.4) et à la fonction syntaxique (§2.3). Les flèches indiquent la continuité¹⁶. Nos résultats ne contredisent pas à l'adjectif prototypique et au substantif prototypique. Mais comme prévu dans la section 3, les conditions sont plus nombreuses et variées que celles qui sont présentées dans la table 1. Ces dernières, directement déduites de l'opposition entre la Propriété et l'Objet, sont,

d'après nous, trop strictes pour saisir la totalité de ces problèmes. Nous relâchons donc notre hypothèse en recourant à une autre opposition ayant moins de précision, qui est celle entre le CONTINU et le DISCRET, ou celle entre l'ANALOGIQUE et le DIGITAL, indiquées à la dernière ligne de la table 2. Nous nous servirons dans ce qui suit les termes ANALOGIQUE et DIGITAL plutôt que Continu et Discret, puisque ceux-ci sont des termes fréquemment employés en linguistique, ayant déjà des définitions fixes. Les notions d'analogique et digital et leur distinction ont été à l'origine discutées dans les sciences de l'information et de la communication¹⁷.

Table 2. Facteurs déterminant l'alternance ADJ vs. NOM des termes de couleur

		ADJ en -I est plus utilisé ou de règle	↔	NOM -NO est plus utilisé ou de règle
Propriété syntaxique des lexèmes de couleur (§5.2-§5.3.4)		SHIRO (blanc) AKA (rouge) AO (bleu) KURO (noir)	↔	CHA-IRO (marron/brun) MIDORI □vert□
Conditions sémantico-fonctionnelles dans le discours	Référent chromatique (référentielle, §5.3.1)	Couleur dans la nature + Complexe + Profonde + Dynamique - Référentiel	↔	Couleur artificielle + Simple + Plate + Statique +Référentiel
	Relation entre le nom modifié et le modifiant (relationnelle, §5.3.2)	Objet et son apparence chromatique	↔	Croisement de deux notions : forme et couleur, contenant et contenu etc.
	Extension du nom modifié (déterminative, §5.3.3)	+ Large	↔	+Étroit
	Intention énonciative (énonciative, §5.3.4)	Littéraire + Transmission de l'émotion + subjectif	↔	Journalistique + Transmission de l'information + objectif
	Fonction des termes de couleur (fonctionnelle, §2.3)	Description	↔	Distinction Sous-catégorisation
	SOMMAIRE	PROPRIÉTÉ CONTINU ANALOGIQUE	↔	OBJET DISCRET DIGITAL

L'analogique indique des signes ou de l'information représentés par une quantité variable d'une manière continue, alors que le digital désigne ceux qui sont représentés par des valeurs discrètes. L'analogique désigne donc l'information continue et subjective qui peut prendre une infinité de valeurs, tandis que le digital désigne l'information discontinue et logique qui peut prendre une valeur précise. Selon nous, toutes les conditions sémantico-fonctionnelles dans la table 2, bien

qu'elles soient indépendantes l'une de l'autre, peuvent être interprétées de ce point de vue, ANALOGIQUE vs. DIGITAL. Notre hypothèse est ainsi révisée comme suit : le terme de couleur a plus de probabilité d'être exprimé en tant qu'adjectif, s'il a plus de caractéristiques sémantico-fonctionnelles "ANALOGIQUES" et il a plus de chances d'être exprimé en tant que nom, s'il a plus de caractères sémantico-fonctionnels "DIGITAUX".

Nous voudrions maintenant insister sur l'indépendance de la caractéristique syntaxique des lexèmes et la propriété sémantico-fonctionnelle des catégories grammaticales. Alors que, par exemple, la forme substantivale des quatre premiers termes ne désigne pas en principe la couleur naturelle, ce n'est pas le cas pour le jaune et le marron. Avec ceux-ci, la désignation d'une couleur naturelle est bien possible sous forme de substantif et plus précisément, le marron substantif est plus fréquent et plus acceptable que le jaune substantif dans cette condition. Il s'agit d'une caractéristique immanente des lexèmes qui n'est pas attribuable à d'autres facteurs généraux. Cette caractéristique lexicale de chaque lexème est une condition prédéterminée dans la langue, étant donné que leur comportement est spécifique et difficilement prévisible. On pourrait dire qu'il est déterminé socio-culturellement d'une manière sporadique dans l'histoire de chaque mot¹⁸. Il faut cependant rendre compte aussi du fait que les quatre lexèmes majoritairement utilisés comme adjectif sont les termes de couleur dits fondamentaux et le plus fréquemment employés en japonais. À l'opposé, les conditions sémantico-fonctionnelles sont explicables en termes généraux. Ces conditions sont déterminées dans le discours, dans la relation entre le locuteur et le monde référentiel et suivant l'intention de l'énonciateur.

Nous faisons enfin remarquer avec la table 3, l'interaction entre la syntaxe des lexèmes et le sémantisme des parties du discours. En ce qui concerne les quatre termes de base, l'usage adjectival étant non marqué, cet usage est réalisé sous toutes les conditions sémantico-fonctionnelles ANALOGIQUES et NON-ANALOGIQUES. Même dans la condition hautement DIGITALE, l'usage de l'adjectif n'est pas exclu. Suivant la table 2, cette condition est réalisée dans le cas où la référence chromatique est artificielle, l'extension du nom modifié est étroite, la relation entre le nom modifié et le terme modifiant est logique, et l'intention énonciative est la sous-catégorisation de l'objet modifié.

Par exemple, pour distinguer une Mercedes blanche d'une autre bleue et d'une Jaguar blanche, on peut très bien employer la forme adjectivale : SHIRO-I MERCEDES (blanc-ADJ Mercedes, "Mercedes blanche") par opposition à AO-I MERCEDES (bleu-ADJ Mercedes, "Mercedes bleue"), et à SHIRO-I Jaguar (blanc-ADJ Jaguar, "Jaguar blanche") dans le contexte : "Cette fois, il a acheté une Mercedes blanche". Mais si les conditions sont encore plus DIGITALES, l'usage substantival est sans doute préféré. C'est le cas par exemple de la classification des voitures faite par des professionnels par marque et couleur dans le catalogue de voitures. En ce qui concerne CHA-IRO (marron/ brun), étant donné que l'usage nominal est non marqué pour ce lexème, cet usage peut être réalisé sous une

condition sémantico-fonctionnelle NON-DIGITALE : référence naturelle, extension large du nom modifié, intention descriptive de l'énonciateur. *CHA-IRO-NO KE* (brun-GÉN cheveux, “les cheveux bruns”) ne pose pas de problème. Mais on peut trouver quelquefois une différence sémantique entre les deux formes. *CHA-IRO-NO KE* demande une interprétation plus DIGITALE que *CHA-IRO-I KE* (brun-ADJ cheveux) dans le contexte : “une jeune fille aux cheveux bruns /brunâtres”. Le premier indique une couleur de cheveux qui n'est ni noir ni roux mais brun homogène, tandis que le second peut désigner les cheveux abimés par le soleil d'été, dont la couleur n'est pas homogène. *KI-IRO* (jaune) se situe entre *CHA-IRO* et les quatre premiers. L'usage substantival de *KI-IRO* (jaune) demande une condition plus DIGITALE que celui de *CHA-IRO* (marron). Et enfin, *MIDORI* (vert) et tous les autres termes de couleur non dérivationnels que nous n'avons pas étudiés dans cet article sont automatiquement nominaux sans aucun rapport avec les conditions sémantico-fonctionnelles, puisqu'ils n'ont qu'une seule forme dans la norme de la langue japonaise.

Table 3. Interaction entre la syntaxe des lexèmes et le sémantisme des parties du discours

lexème	Syntaxe de lexème		Conditions sémantico-fonctionnelles	
	ADJ en -I	NOM -GÉN	ADJ en -I	NOM -GÉN
<i>AKA</i> (rouge) <i>AO</i> (bleu) <i>SHIRO</i> (blanc) <i>KURO</i> (noir)	Non marqué	+++marqué	±ANALOGIQUE	++++DIGITAL
<i>KI-IRO</i> (jaune)	+ marqué	+marqué	+ANALOGIQUE	++DIGITAL
<i>CHA-IRO</i> (marron / brun)	++marqué	+marqué	++ANALOGIQUE	+ DIGITAL
<i>MIDORI</i> (vert)	Impossible	Seul choix	Sans conditions	

7. Conclusion

Nous avons démontré dans cette étude le caractère multifactoriel de l'opposition adjectivo-nominale des termes de couleur japonais. Le facteur le plus important est lexical. Chaque lexème de couleur a sa tendance syntaxique qui détermine globalement son comportement grammatical. Mais il y a aussi les conditions sémantico-fonctionnelles qui interviennent dans la décision de la forme du mot. Si ces dernières sont plus ANALOGIQUES, la forme a plus de chance d'être adjectivale et si elles sont plus DIGITALES, la forme a plus de chance d'être nominale. Les

conditions sémantico-fonctionnelles, à savoir fonctionnelle, référentielle, relationnelle, énonciative et déterminative sont elles aussi multifactorielles, chaque condition étant indépendante l'une de l'autre¹⁹.

Revenons enfin aux exemples (7), (8) et (21). La fonction du nom chromatique n'est pas seulement distinctive mais aussi descriptive si la condition lexicale et d'autres conditions sémantico-fonctionnelles le permettent comme on l'a vu dans l'exemple (7). À l'opposé, le nom n'est pas accepté bien qu'il joue le rôle distinctif lorsque les conditions lexicale, référentielle et déterminative exigent l'usage de l'adjectif comme dans l'exemple (8). Dans l'exemple (21), bien que les conditions référentielle et fonctionnelle soient adéquates à l'usage du nom, la condition déterminative concernant l'extension nominale l'empêche. Mais il faut avant tout faire remarquer que les lexèmes de couleur dans ces trois exemples ont la propriété syntaxique encline à l'usage adjectival.

Si cette conclusion paraît floue, c'est en grande partie en raison des caractéristiques de l'objet de recherche, c'est-à-dire surtout à cause de l'ambiguïté de la détermination nominale dans la langue japonaise. C'est d'ailleurs en raison de cette difficulté que la question traitée ici est restée intacte jusqu'à ce jour. Une linguistique de corpus pourra aider à mieux l'élucider.

Notes

¹ Les recherches ont été en partie financées par une subvention du Ministère de l'Éducation et des Sciences du Japon (Grant-in-Aid for Scientific Research (C) 20520379). J'ai plaisir à remercier Denise Malrieu et Claire Dodane qui m'ont beaucoup aidée pour accomplir ce travail. Mais tous les défauts de cet article sont les miens.

² En japonais, il existe en plus un vaste domaine des termes chromatiques qui ne sont pas basiques. Ils sont dérivés des termes de base ou empruntés de l'anglais. Leurs occurrences sont cependant très limitées dans nos corpus. Voir Molinier (2006) pour cette question en français.

³ *CHA-IRO* (marron/ brun) , *HAI-IRO* (gris) proviennent de "la couleur de thé" et de "la couleur de cendre" respectivement. L'origine de *KI-IRO* (jaune) est obscure.

⁴ Lorsque l'adjectif est épithète, le suffixe adjectival est invariable. Lorsqu'il est prédicatif, le suffixe se conjugue suivant le contexte étroit qui le suit. Les substantifs japonais sont syntaxiquement et morphologiquement autonomes. Ils n'ont pas de désinences flexionnelles, ni n'exigent la présence d'un autre mot tel qu'un article.

⁵ Dans le domaine de la figure géométrique, *MARU*(rond) et *SHI-KAKU* (quatre-angle : carré) ont la forme adjectivale *MARU-I*, et *SHI-KAKU-I*. Mais *SAN-KAKU*(trois-angle: triangle) ne l'a pas.

⁶ Le choix entre les deux formes suffixales *-I* ou *-NA* est lexicalement déterminé. Les adjectifs de couleur qu'on traite ici sont tous en *-I*. On dit que le morphème *-I* n'a plus actuellement de

productivité dérivationnelle, mais que dans les périodes anciennes, il pouvait former de nouveaux adjectifs à partir des substantifs (c'était *-KI* autrefois). À l'opposé, le suffixe *-NA* garde toujours la capacité à créer des néologismes adjectivaux à partir des substantifs.

⁷ Un autre moyen fréquent est la juxtaposition des deux substantifs sans particule, qui produit une sorte de nom composé. Par exemple, le "vin rouge" s'exprime par ce procédé : *AKA-WINE* (Rouge-VIN). L'occurrence de ce moyen idiomatique n'est pas aussi régulière que l'emploi de la particule *NO*. On observe ce cas dans l'exemple (3), l'absence de particule étant indiquée par \emptyset , mais non dans l'exemple (2).

⁸ La notion de prototype a suscité de nombreuses discussions. Voir Kleiber (1990) entre autres.

⁹ Les fameuses discussions faites par des cognitivistes tels que Berlin and Kay (1969) sur la catégorisation des couleurs est fondée sur la recherche de ce genre de couleurs qui sont des "objets".

¹⁰ *Le Journal Mainichi* est un des cinq quotidiens nationaux japonais.

¹¹ Le japonais ne met pas d'espaces entre les mots dans le texte. Nous avons donc calculé le nombre des morphèmes des corpus, après avoir automatiquement fait l'analyse morphologique avec l'étiqueteur : ChaSen. Pour recueillir des exemples, nous avons employé les corpus bruts sans recourir à l'étiqueteur.

¹² Voir nos travaux sur la question de la syntaxe des lexèmes dans l'opposition de *DE* vs. *DES* devant les noms précédés d'épithète en français : Fujimura et al. (2004), Fujimura et al. (2007). Une forte cooccurrence entre *des* et *petits* (*des petits effectifs*) est à aborder du point de vue lexical.

¹³ Il est d'ailleurs intéressant, du point de vue méthodologique, que la relation entre le choix des formes et la différence des couleurs – naturelle et artificielle – n'est pas quantitativement manifeste comme on le voit dans la figure 3, bien que statistiquement significative. Qualitativement par contre, cette relation est évidente, l'usage nominal étant interdit pour indiquer les couleurs naturelles dans la plupart des cas.

¹⁴ *KURO* (noir) a une particularité que nous ne pouvons pas aborder dans ce travail. Dans le cas de la référence à la couleur artificielle, le comportement de *KURO* est différent des autres trois couleurs de base. Voir les figures 1, 3 et 5.

¹⁵ Une tendance semblable est observée en somali et en anglais dans Biber (1995 : 79). L'adjectif épithète est plus fréquemment employé dans la littérature que dans la presse, par rapport à d'autres moyens qui modifient les noms, dans ces deux langues.

¹⁶ Ce modèle a été inspiré de l'hypothèse de transitivité proposée par Hopper and Thompson (1980). Selon cette hypothèse fréquemment citée, la transitivité est définie en tant qu'une notion complexe composée de nombreux facteurs, qui n'ont pas de relation de dépendance entre eux, dans de différents domaines sémantico-fonctionnels. Voir Fujimura (1989) pour la transitivité en japonais.

¹⁷ Voir Chandler (2007) et Wilden (1987) pour les détails de ces deux notions : ANALOGIQUE et DIGITAL.

¹⁸ Il pourrait y avoir également l'intervention des facteurs phonétiques. Je dois cette remarque à Andrée Borillo (communication personnelle).

¹⁹ Dans les haïkus, poèmes classiques japonais de dix-sept syllabes, un petit contraste comme le nôtre est souvent décisif pour déterminer leur valeur littéraire. C'est le cas de ce haïku fameux : *AKA-I TSUBAKI, SHIRO-I TSUBAKI TO OCHINI-KERI* (Rouge-ADJ Camélia, Blanc-ADJ Camélia, Sont Tombés) "Il est tombé une fleur de camélia rouge et une autre de camélia blanc, l'une après l'autre". Selon nous, les multiples facteurs examinés conditionnent la littérarité de ce haïku.

Références

- Beck, D. 2002. *The typology of parts of speech systems: The markedness of adjectives*. New York and London: Routledge.
- Berlin, B. and P. Kay. 1969. *Basic Color Terms*. Berkeley and Los Angeles: University of California Press.
- Biber, D. 1995. *Dimensions of register variation: a cross-linguistic comparison*, Cambridge: Cambridge University Press.
- Chandler, D. 2007. *Semiotics: the basics*. New York and London: Routledge. <http://www.aber.ac.uk/media/Documents/S4B/semiotic.html>.
- Croft, W. 1991. *Syntactic categories and grammatical relations: the cognitive organization of information*. Chicago and London: University of Chicago Press.
- Fujimura, I. 1989. Un cas de manifestation du degré de transitivité l'alternance des relateurs O et GA en japonais. *Bulletin de la Société de Linguistique de Paris*, 84, 1, 235-257.
- Fujimura, I., M. Uchida and H. Nakao. 2004. *De vs des* devant les noms précédés d'épithète en français: le problème de *petit*. In G. Purnelle, C. Fairon and A. Dister (eds), *Le Poids des mots 1*. Louvain-la Neuve: Presse Universitaire de Louvain, 456-467.
- Fujimura, I., M. Uchida and H. Nakao. 2007. Opposition entre *de* et *des* devant les noms précédés d'épithète en français : portée du "poids ". *Texte et corpus 2003: acte des JLC 3*, 131-144.
- Goes, J. 1999. *L'adjectif : Entre nom et verbe*. Paris and Bruxelles: Duculot.
- Hopper, P.J. and S.A. Thompson. 1980. Transitivity in grammar and discourse. *Language* 56: 251-299.
- Kleiber, G. 1990. *La sémantique du prototype, Catégories et sens lexical*. Paris: PUF.
- Molinier, Ch. 2006. Les termes de couleur en français. Essai de classification sémantico-syntaxique. *Cahiers de Grammaire* 30: 259-275.
- Sawada, N. 1992. The indicatory function in nouns compared with restrictivity and descriptivity of adjectives. From an analysis of selectional factors in the basic color words. *Journal of the Linguistic Society of Japan* 102: 1-16.
- Tucker, G.H. 1999. *The lexicon grammar of adjectives: a systemic functional approach to lexis*. London: Cassell.
- Whittaker, S. 2002. *La notion de gradation: application aux adjectifs*. Berne: Peter Lang.
- Wierzbicka, A. 1988. What's in a noun? (or: how do nouns differ in meaning from adjectives?). In *The semantics of grammar*. Amsterdam and Philadelphia: J. Benjamins, 463-496.

Wilden, A. 1987. *The rules are no game: the strategy of communication*. London and New York: Routledge & Kegan Paul.

THE SEMANTIC VARIATION OF
VERB *ESSERE* IN ITALIAN
THEORETICAL CONSEQUENCES OF CORPUS-BASED STUDIES

Alessandro Panunzi
University of Florence

1. The verb *essere*: theoretical frameworks

1.1 Introduction

This study contains the results of a research on the semantic variation of the verb *essere* (*to be*) in Italian. The specific object of research is the use of such a verb within the Italian section of C-ORAL-ROM spontaneous speech corpus (Cresti and Moneglia 2005). Data from the lemmatized corpus show that the verb *essere* counts 13,831 occurrences, which correspond to 4.66% of the total tokens. The quantitative relevance of verb *essere* in the corpus is even more valuable if we consider its presence within the utterance, considered as the reference unit of spoken language (Cresti 2000). In this respect, data show that *essere* occurs in 10,862 utterances of 38,593 (28%). If we discard the *verbless* utterances (those in which there is not a finite verb form), the datum rises to 45.5% of the *verbal* utterances (10,862/23,873).

All the corpus examples have been extracted and classified following a general schema. The criteria for the building of such a schema are reported in detail in paragraph 2.1. At a first level, three macroclasses of use have been distinguished: (a) the auxiliary uses; (b) the uses within *esserci* (*there is*), considered as an autonomous lemma (Panunzi 2005); (c) the properly verbal uses, in which *essere* occurs as an independent lexical entry.

The specific focus of this study is the set of properly verbal uses of *essere*. The classification of the corpus occurrences presupposes the identification of the taxonomic criteria. With the aim of selecting these criteria, in this first paragraph we will take into account different interpretations of the semantic values of verb *essere*, and namely the remarks on the identity predicate pointed out by logicians, the taxonomy of the copular sentences developed in the framework of the generative

grammar and the interpretation of the meaning of the verb *to be* in the cognitive semantics.

In the second paragraph, the criteria adopted for the corpus analysis will be introduced: the most relevant perspectives among those examined in the literature will be integrated in a single and original theoretical framework. Moreover, such a framework will be compared to the Higgins' (1979) proposal of "taxonomy of copular sentences".

In the third paragraph, the results of the *corpus-based* classification will be showed, and, within a stricter *corpus-driven* perspective, an inherent line of variation will be identified for each type of use.

In general, the choice to work from a spoken corpus reflects the need to compare linguistic theory with spontaneous data, which are best represented in oral collections. Moreover, corpus analysis proves essential to the aim of the observational adequacy of the theory. This way only, indeed, it is possible to take into account the positive evidence that can be observed from massive data. On the contrary, competence judgments tend to focus on evidence that is basically negative.

In this study, both directions for the analysis of corpus linguistics are integrated: the *corpus-based* one and the *corpus-driven* one (Tognini-Bonelli 2001). On the one hand, the whole set of occurrences of the verb *essere* in the corpus constitutes, once it is assumed as a statistical universe representing the language, the matter that requires an explanation, the field on which the linguist estimates the adequacy of grammatical theories (*corpus-based* direction). On the other hand, the corpus provides data for the induction of the semantic variation of the verb, and integrates the theory highlighting structures that are not-predictable on the basis of sole competence (*corpus-driven* direction).

1.2 From the classical tradition to the analytic philosophy of language

The classical logic tradition conceived the verb *essere* as the linguistic equivalent of the concept of copula, particularly referring to the value that it assumes within the structure of the predication. Aristotle found the copula to be an element that is able to reveal the twofold value of the *rhema*, considered both as "predicate" and as "tense specification". Such values, usually associated within verbal lemmas, are distinct in the structure of nominal predication. In these contexts, the verb *essere* would correspond to the expression of the "tense" without the "predicate": a free morpheme, we would say, that carries purely grammatical feature, whereas the lexical predicate is constituted by a nominal element, that lacks tense information¹.

The Scholasticism, and particularly Peter Abelard (who introduced the term "copula"), pointed out the function of the verb *to be* as a "link", a linguistic objectification of the relation that governs the tripartite structure of predication (subject - tense - predicate). Such a perspective would be further developed by the

Port-Royal grammarians, for which the partition among subject, copula and predicate is set at the basis of the two fundamental activities of the spirit: *to conceive* (to establish names for “substances” and “accidents”) and *to judge* (to set relations among names).

The reflections of logicians between the 19th and 20th centuries came to brake this classical tradition of thought², in which the lemma *essere* and the concept of copula are actually identified. First Frege, then Russel, and eventually the debate about the referential ambiguity of the definite descriptions (Strawson 1950; Donnellan 1966) pointed out the polysemic nature of *to be*: beside its value of copula, the verb can be used as identity predicate.

Such a theoretical turning point is in line with the semantic distinction between two different types of logical entities: *concepts* and *objects*. According to Frege (1892b), the *concept* (*Begriff*) is a predicative entity, while the *object* (*Gegenstand*) is a referential entity, which cannot correspond to a predicate³. This permits us to account for the fundamental difference between the pair of propositions:

- (1) The morning star is a planet
- (2) The morning star is Venus

The proposition in (1), in fact, could be paraphrased as:

- (1') the concept referred by the predicate “planet” can be applied to the object referred by the singular term “morning star”

It is the case, following Frege, of a relation that can be defined as “the falling of an *object* under a *concept*”, which corresponds to the copular relation.

Otherwise, a similar paraphrase is not possible for the proposition in (2). This happens because of a very simple reason: “Venus” is a proper name that is a singular term *par excellence*. Thus, its reference cannot correspond to a *concept* under which the mentioned *object* (“the morning star”) can fall. The reference of “Venus” is necessarily an *object* itself. The proposition in (2) must be represented in another way, with respect to what we’ve seen for (1), and namely:

- (2') the singular terms “morning star” and “Venus” refer to the same object

We are in front of a completely different kind of relation: an “equation” between two distinct “ways of referring” to the same object, that is the identity relation.

In other words, while in (1) there is a nominal predication, in (2) this does not happen, because a predicative value is not applicable to the noun phrase at the right of the verb *to be*. On the contrary, there is a distinction strictly on the semantic-referential level.

Using the terms of the Fregean paradigm that distinguishes the components of the linguistic meaning (*Sinn/Bedeutung, sense/reference*; see Frege 1892a), we could conclude that the copular relation is situated on the reference level, because it regards the inclusion of an *object* (the reference of a singular term) in a *concept* (the reference of a predicate). On the contrary, the identity relation is situated on the level of the sense, setting the equivalence of two distinct senses, that lead to the same reference.

The Fregean thesis about the polisemy of the verb *to be* is also shared by Russell (1911), who expressly proposed in the *Introduction to the Mathematics Philosophy* to distinguish two semantic values of such a verb: a copular one, as a grammatical mark, and a properly predicative one, being able to express an identity relation between two nominal constituents.

A second central point of the analysis carried out by logicians is the interpretation of the referential properties of *definite descriptions* (Russel 1905), about which a very important debate has been carried out. In this respect, Strawson (1950) made a distinction between *referential* and *attributive* uses of a definite noun phrase (e.g. *the greatest French soldier*), mostly based on the semantic function that such phrases can assume in a sentence:

- (3) The greatest French soldier died in exile
- (4) Napoleon was the greatest French soldier

The same definite description has a *referential* use in (3), as it is the sentence's subject; on the contrary, it has an *attributive* use in (4), since it plays the role of predicate. In other words, Strawson introduced the idea that a definite description is referentially ambiguous, but he claimed that its logical function can tell us what kind of use is selected within a certain syntactic configuration.

Donnellan (1966) developed a radical criticism of the Strawson's notion of referentiality, and proposed a different way to account for the referential properties of a definite description. First of all, he made a distinction between the *denotation*, defined on a semantic basis ("a definite description denotes an entity if that entity fits the description uniquely"), and the *reference*, defined as a pragmatic relation between the speaker and the thing he wants to talk about. With this respect, he placed the opposition between *referential* and *attributive* uses on a strictly pragmatic level, introducing a terminological shift. So, both *referential* and *attributive* uses of the same definite description (e.g. *Smith's murderer*) are possible even within the same sentence, depending on the sole properties of the utterance:

- (5) Smith's murderer is insane

In this framework, "Smith's murderer" has a *referential* use if the speaker says (5) having in mind the referred person. On the contrary, it has an *attributive* use if the

speaker utters (5) without having any particular person in mind (e.g. basing his claim on the sole particularly brutal manner in which Smith has been murdered). The proposals of Strawson and Donnellan have been determinant for the most part of further linguistic reflections. In the following paragraph, we will introduce the interpretation of the “copular sentences” within the generative grammar framework, which are mainly based on the analysis of the referential traits of predication constituents.

1.3 The referential taxonomy of copular sentences

After Halliday (1967), who introduced the phrase, most linguistic studies analyzed the utterances in which the verb *to be* occurs followed by an adjectival or nominal complement, labeling them as *copular sentences*. Within the different interpretative frameworks for the analysis of such sentences, at least two main types are commonly distinguished, on the basis of the referential values of the involved constituents (and particularly of the post-copular one). Several terms have been proposed for the distinction of these two types⁴; the terminology became more stable after the Akmajan (1979) proposal. In such a proposal, the copular sentences have been distinguished in: (a) predicational sentences, in which the post-copular phrase is not referential, and constitutes a predicate (e.g. *Carl is bald*); (b) specificational sentences, in which the post-copular phrase is properly referential, and *specifies* an individual (e.g. *my best friend is Carl*).

Following the original definition in Akmajan, the distinction between two typologies is based on the role played by the *complement* (the phrase at the right of *to be*), with respect to the detection of the predication referent: while in a predicational sentence the complement says something *about* the referent, in a specificational one it says *who* is the referent.

These two classes of sentences constituted the core of Higgins’ (1979) “referential taxonomy of the copular sentences⁵”, that comprehends two additional categories: (c) *identificational* sentences, in which the subject has a stronger referential value than the one expressed by the complement (which is, however, referential; e.g. *this is my best friend*); (d) *identity statements*, that *grosso modo* correspond to the sentences discussed by Frege (e.g. *The Slim is Carl*).

The four types of copular sentences have been defined on the basis of the degree of referentiality of both the subject and the complement of the verb *to be*. The following table summarizes the Higgins’ criteria:

Table 1. Criteria of the referential taxonomy of copular sentences (Higgins 1979)

Type of copular sentence	Subject referential property	Complement referential property
Predicational	Referential	Predicational
Specificational	Superscriptional	Specificational
Identificational	Referential	Identificational
Identity statement	Referential	Referential

The more prominent distinction is the one between the *specificational* type and all the other ones: this category, in fact, is the only one that has, in the subject position, a non properly referential element (*superscriptional*, in Higgins' terms). In this kind of sentence, the subject is considered as an empty element, a generic label whose referential value is specified by the complement of the structure.

We will not take into account further details about the definition of types within the referential taxonomy of copular sentences. Nevertheless, here we will consider some of the most problematic points of this proposal. The main criticisms regarding Higgins' model concerned the weakness of the referential classes that are used in the taxonomy. Such classes are on the one hand insufficiently definite (e.g. the distinction among *referential* / *specificational* / *identificational* is not so clear), and on the other hand heterogeneously identified (Van Peteghem 1991).

The adopted criteria take place, indeed, at different linguistic levels: first, strictly referential properties of the constituents are taken into account (opposition between referential and predicative uses, *à la Strawson*); second, the definition of the *specificational* class is substantially based on the informative properties of the sentence subject (opposition between definite reference and non definite one, *à la Donnellan*); finally, the *identity statement* class tries to occupy the original Fregean distinction.

Thus, a great number of copular sentences are ambiguous with respect to the classification criteria, and can belong to different types, depending on their context and their interpretation. Higgins himself points out that the sentence *the girl who helps us on Friday is my sister* has, following his criteria, three possible interpretations: (a) predicational, when it answers to the question: *what kind of relation do you have with the girl who helps us on Friday?*; (b) specificational, when it answers to the question: *who is the girl who helps us on Friday?*; (c) identificational, if it is used in a deictic context, to identify *that girl (the one who helps us on Friday) as my sister*.

The weak decidability of Higgins' taxonomy gave rise for further reinterpretation of the original criteria (Declerck 1988; Mikkelsen 2005). The distinction between predicational and specificational sentences has been widely discussed within the generative grammar framework. Different interpretations of the movements involved in the production of a copular sentence surface structure have been supposed. However, it is possible to identify a common assumption, mostly

accepted by all the generative grammarians: *specificational* sentences are *inverse copular sentences* (see Verheugd 1990).

This paradigm comes from the analysis of Blom and Daalder (1977)⁶, which claims that there is the same deep structure that underlies the constituent of both predicational and specificational sentences. According to their proposal, in all copular sentences there is a noun phrase that plays a predicative role (hypernymy) with respect to another one⁷.

The consequences of such a conception, that is, the idea that a specificational sentence is nothing but a copular sentence with inverse order, deeply influenced the following generative grammar literature. For what regard the deep structure underlying the copular clauses, there is a rather wide agreement among different studies (see Heggie 1988a, 1988b; Moro 1997; Heycock and Kroch 1999; Rothstein 2001). The deep structure assumed for the copular sentences is the one proposed by Stowell (1978), who claims that the copula is a *raising verb*⁸ that has a *small clause* as an argument⁹:

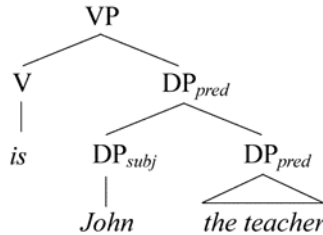


Figure 1. Hypothesis of the deep structure of copular sentences¹⁰

In predicational sentences, the subject of the *small clause* moves to the position of main sentence subject (specifier of IP), with the movement that is usually expected in the case of raising verbs; in parallel, the verb *to be* raises to the position of head of IP:

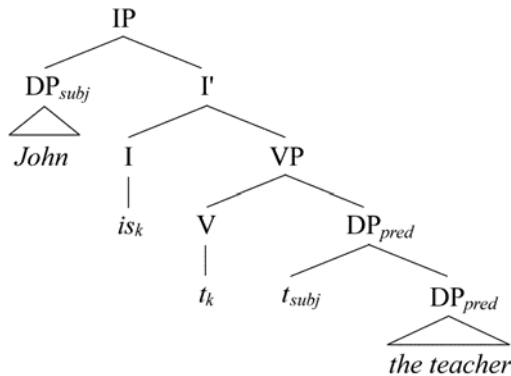


Figure 2. Hypothesis of the surface structure of predicational sentences

For what regards the structure of *specificational* sentences, several interpretations have been proposed within the generative grammar. Here we will present only one of the most shared hypotheses, the one provided by Moro (1997), namely *predicate raising*¹¹.

Moro accepts the idea to treat *predicational* and *specificational* sentences from a shared deep structure, as the one offered in figure 1. He claims that the difference between the surface structures is based on the choice of the element that rises to the position of the main sentence subject. The following is the surface structure that he proposes for the *specificational* sentences:

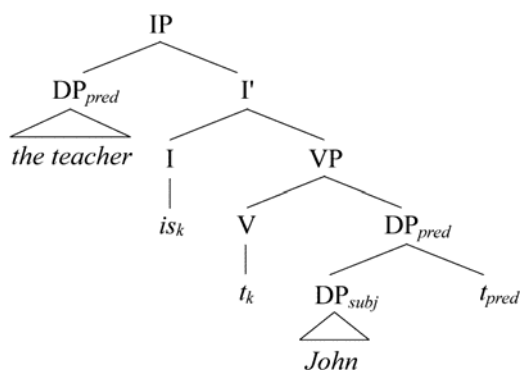


Figure 3. Hypothesis of the surface structures for the *specificational* sentences (Moro 1997)

Contrary to what occurs in the predicational sentences, in this case the predicative noun phrase (DP_{pred} , that is inside the *small clause* in the deep structure) moves to the surface position of the specifier of IP, while the noun phrase that plays the role of subject in the *small clause* (DP_{subj}) remains in its original position.

In brief, according to Moro's proposal the subject of a *specificational* sentence would represent the deep predicate of the structure, while the constituent that follows the verb *to be* would correspond to the deep subject. Such a configuration would invert the canonical predicative relation shown by a *predicational* sentence, with an unmarked constituent order. Thus, all the verbal uses of *to be* are assumed to be copular (see Moro 1988, who proposes an "unified theory of copular sentences").

Whereas the classes defined by Akmajan and the Higgin's taxonomies try to take into account the different values of the verb *to be*, most of the further studies in generative grammar substantially eliminate the problem related to the semantic interpretation of the involved structures, by means of the hypothesis that the differences are only bound to the constituent movement at the performance level, and do not correspond to any actual semantic variation. On the contrary, from our point of view this variation constitutes the basis for the correct interpretation of the use of the verb *to be*. For this reason, the proposal carried out by generative

grammar cannot be adequate in order to account for the semantic variation of the verb *to be*.

1.4 Semantic Fields and Thematic Relations

A further point of view on the semantic analysis of the verb *to be* has been developed within the framework of the Cognitive Semantics. The fundamental contribution for these reflections comes from the analysis of spatial relations, considered as the primary semantic field. On this basis, the semantic relations model is extended on further semantic fields.

Beginning with the reflections of Gruber (1976), Jackendoff identifies the basic components of a spatial relation as: (a) the *theme*, which corresponds to the localized object; (b) the *reference object*, which corresponds to the referential element used for defining the space involved in the relation. As an example, for what regards the analysis of the verb *to be*, we could say that the sentence in (6):

(6) John is at the bus stop

establishes a spatial relation between two referential elements: “a person” and “a bus stop”. More in detail, the verb *to be* assigns the thematic role of location to the *referring object* (“the bus stop”); in this way, the space of the location is defined, and so the *theme* (“Gianni”) can be localized through it.

Such a model is put at the basis of the meaning of the verb: the locative semantic core of *to be* is extended to different (nonspatial) semantic fields though the Thematic Relations Hypothesis (TRH), that can be summarized as follows: (1) in each semantic field, the functions performed by the constituents of a relation belong to a subset of those that are used for the spatial relation analysis; (2) semantic fields are defined by: a) the types of entities that can occur as the *theme*, b) the types of entities that can occur as reference object, c) the type of the established relation, which assumes the role that the localization plays within the semantic field of spatial expressions.

Table 2 shows the schema of Jackendoff’s proposal for the analysis of the verb *to be*. In Jackendoff’s proposal the copular value of *to be* is treated (as well as its value of identity predicate) as an extension of the spatial relation to the “identificational” semantic field.

Anyway, it has to be noticed that this interpretation is different with respect to Gruber’s original one, which places a distinction between two cognitive primitives, *positional* and *identificational*, that are characterized by different semantic properties and syntactic behaviors¹². In Gruber, the term *identificational* does not indicate a semantic field among the others, but a more general cognitive schema, to which it is possible to ascribe different semantic types and verb classes.

Table 2. Synthesis of Jackendoff's (1983) proposal

Semantic field	Theme	Reference object	Type of established relation (<i>to be</i>)	Example
Spatial	object	place	localization of the theme in the place of reference	<i>John is in the room</i>
Temporal	event or state	time	localization of the event/state expressed by the theme in the time of reference	<i>The meeting is at 6.00 PM</i>
Possessive	object	object	possessive relation between the theme (x) and the object of reference (y) (as a localization: "x is at y")	<i>The doll is yours (the doll belongs to you)</i>
Identificational	object	object-type or property	<i>to be</i> an instance of a class or <i>to have</i> a property play the role of localization	<i>Elise is a pianist The light is red</i>
Circumstantial	object	event or state	the theme (x) takes part to the event of reference (y) ("x is at/in y")	<i>Ludwig is composing quartets</i>
Existential	object or state	existence	the theme is placed in the referential space of existence	"be in existence"

On the other hand, the irreducibility of the copular value to a mere field of variation of the locative one turns out evident if we consider that, in the copular uses, the complement lacks *a fortiori* of the referential feature, since it has a predicative value. As the assignment of a *theta* role to a syntactic constituent has the necessary condition that it is a referential element, it seems impossible that the predicative complement of a copular relation can play such a role.

Beyond the possible criticism against Jackendoff's theorization, the treatment of the meaning variation of the verb *to be* within the TRH is still an original intuition that has a good explanatory adequacy with respect to several semantic fields in which the verb is used, particularly for the spatial, temporal and possessive ones.

This proposal will be then assumed for the definition of the variation model of the verb *essere* in Italian, even if the copular and identity uses will be considered separately.

2. Taxonomic criteria

2.1 General taxonomy of verb *essere* uses

The verbal uses of *essere* have been distinguished in three main categories that constitute the “central variation” of the verb: the copular use; the identity use and the predicative-locative use (defined on the basis of Jackendoff’s model).

Aside from the “central variation”, a “marked variation” class (Panunzi 2006), negatively identified with respect to the others, has been considered: it is the class of uses in which *essere* does not mean its proper sense. Such a sector of variation collects all the uses in which *essere* occurs within grammaticalization and lexicalization phenomena or within phraseological and stereotypic uses, and will not be treated in his study.

Figure 4 shows the schema of the complete structure of the taxonomy, as it emerges from the choices done:

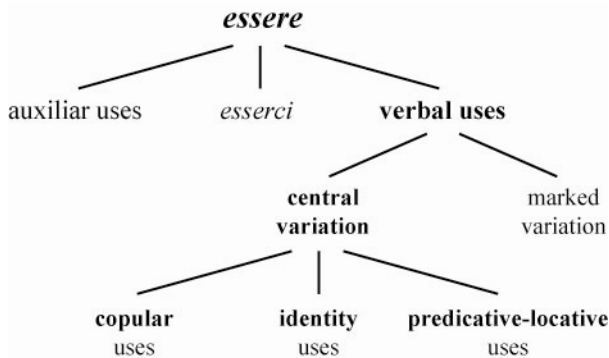


Figure 4. General schema of the classification of verb *essere* uses

Assuming a descriptive perspective, it is possible to define a surface structure that is able to represent the whole set of the verbal uses of *essere*, as in (7)¹³:

(7) subject - *essere* - “complement”

It is then possible to establish a strict correlation between the semantic properties of the complement and the selection of a certain use of *essere*. On the one hand, the distinction among the types within the verb central variation takes primarily into account the partition between copular and non-copular uses, which is based on the referentiality feature of the complement. This is true both for the identity and the predicative-locative uses of *essere*.

On the other hand, The distinction between identity and locative uses is based on the assignment of a *theta* role to the complement. The locative-predicative value,

in fact, needs the assignment of a clear thematic role to the element that constitutes the reference object of the relation. Otherwise, none of the classically identified thematic roles (*agent/experiencer*; *theme/patient*; *locative*; *possessive*; *cause*; *benefactive*; *goal*; *source*; *destination*; see Fillmore 1968; Gruber 1976; Jackendoff 1972, 1983) is assigned to the complement of an identity use. This is motivated by the fact that the semantic relation established is a relation between two sense, i.e. two ways to identify the same object: it is then an intensional relation, and not the representation of a state (or an event). Thus, it is not possible to assign a thematic role to its complement¹⁴.

As with the semantic features of the complement, we can define a set of rules that allow the classification of the uses of verb *essere*: (a) if the complement is not referential, it constitutes a nominal predicate, and the verb *essere* has the value of a copula; (b) if the complement is referential, but it doesn't carry any thematic role, *essere* is used as an identity predicate; (c) if the complement has a thematic role (and so it is, *a fortiori*, referential), *essere* has a predicative-locative value. Such rules project a matrix of features that can be used for classifying the verbal uses in the corpus:

Table 3. Correlation between the uses of *essere* and the complement semantic features

		Value of <i>essere</i>		
		<i>Copula</i>	<i>Identity</i>	<i>Predicative-locative</i>
Complement features	[referentiality]	-	+	+
	[thematic role]	-	-	+

The classification schema and the matrix of the complement features do not take into account Higgin's criteria for the referential taxonomy of the copular sentences, which have been considered inadequate for the corpus-based classification of the occurrences, for the yet mentioned problems of coherence and decidability.

2.2 Comparison between *copular/identity* uses and *predicational/specificational* classes of sentences

The main types of the referential taxonomy of the copular sentences (*predicational* and *identificational*) will be here compared to one of the main theoretical points regarding the criteria adopted in this study: the Fregean principle which distinguishes the identity relation that is the intentional relation between two noun phrases, from the copular relation.

At a first glance, it could seem possible to hypothesize a parallelism between the predicational sentences and the copular value of *essere*, and, in parallel, between the specificational sentences and the identity uses of the verb¹⁵. As a matter of fact, the

class of the *predicational* sentences includes the larger part of the copular uses of *essere*. The *specificational* class, *viceversa*, tends to identify a set of uses in which *essere* seems to have an identity value. However, such a correlation is not consistent. In certain contexts of use, the criteria that allow the distinction between the copular and the identity uses do not match with the partition between *predicational* and *specificational* sentences.

Let's take into account, for first, the *specificational* sentences, like the one in (8)

- (8) La cosa che preferisco è rossa
[the thing that I prefer is red]

According to Higgins, this sentence would be considered as a *specificational* one, on the basis of the degree of referentiality of the subject (*superscriptional*). On the contrary, from the point of view of our taxonomy, such an utterance would be treated as a copular use of *essere*: it is a property attribution to a subject, independent of its individuation¹⁶.

From the complementary point of view, a sentence like (9):

- (9) Gianni è l'insegnante di mio figlio
[Gianni is my son's teacher]

would be classified as *predicational*, because it contains a potentially predicative complement and a prototypically referential subject (a name). The sentence in (10) represents the correspondent *specificational* sentence, in which the constituent order is inverted:

- (10) L'insegnante di mio figlio è Gianni
[my son's teacher is Gianni]

Nonetheless, although the verb *essere* in the sentence (9) is preferentially read as a copula, it is possible to interpret it as an identity predicate.

Let's suppose that a man is talking with a friend about his son's teacher, saying that he's an eccentric person who likes parachuting and used to go to school on a big American motorbike. His friend knows a certain Gianni, that is a parachutist and owns a Harley Davidson; but he doesn't know (and, for some reason, he excludes) that Gianni is a teacher. In this case, if the friend would say: "my friend Gianni is also a parachutist, and he owns a big motorcycle", the other could conclude with the sentence in (9), which would have the value of an identity predication, as in the following paraphrases:

- (9') "Gianni" e "l'insegnante di mio figlio" hanno in realtà lo stesso riferimento
["Gianni" and "my son's teacher" have the same reference]

The ambiguity about the interpretation of these kinds of sentences has direct consequences on one of the main arguments that would support the hypothesis of the predicative function of the subject in *specificational* sentences. In Italian, it is possible to pronominalize the complement of a *predicational* sentence (as it is part of the VP), but this is not possible with the complement of a *specificational* one (Salvi 1991), as shown by the following examples:

(9.b) Gianni è l'insegnante di mio figlio → Gianni *lo* è
 [Gianni is my son's teacher → John *is it*]

(9.b) L'insegnante di mio figlio è Gianni → *L'insegnante di mio figlio *lo* è
 [My son's teacher is Gianni → *My son's teacher *it is*]

Nevertheless, the possibility foreseen in (9.b) is not verified in all cases. In particular, it is excluded by the identity predicate reading of the verb *essere* value. If we suppose to utter this sentence in the situation described above, which holds to the interpretation in (9'), the mentioned cliticization is no longer possible.

Thus, the possibility of pronominalization depends on the semantic and referential reading of the complement, with respect to the ambiguity between a copular interpretation and an identity one, and it is not a discriminant argument for the distinction between *predicational* and *specificational* sentences.

In general, the effort to make the generalization, starting from a pure syntactic point of view that ascribes all the occurrences of *essere* to a copular predicative structure leads to a lack in semantic adequacy of the taxonomy principles.

The distinction between copular and identity uses depends on the different relation that is established between the nominal constituents of the structure governed by *essere*. As it has been pointed out, the identity relation can be viewed as a relation on the field of *Sinn*, that is, an intensional relation between two expressions that identify the same referent. Parallely, the copular relation is an extensional one, in which the predication is about the belonging of an object to a set defined by a certain property, or by a class definition.

The distinction between predicational and specificational sentences does not adequately represent the primary semantic issue involved in the identity relations, which correspond, as we will see in the next paragraph, to a quantitatively relevant part of the linguistic use.

3. Corpus Analysis

3.1 Quantitative data from the corpus based classification

Starting from the schema in Figure 1, it has been possible to classify the whole set of occurrences of the verb *essere* in the reference corpus, with a very reduced amount of underdeterminacy (less than 1%). If we do not consider the uncategorized cases, it is possible to obtain the main data regarding the distribution of the *essere* occurrences in the categories of auxiliary uses, *esserci* and verbal uses.

The auxiliary uses, with a purely grammatical value, correspond to the 20% of the verb occurrences. Excluding these uses, it is interesting to notice that the ratio between the verbal uses of *essere* and the ones of *esserci* is 4:1. 62% of the total occurrences of *essere* correspond to proper verbal uses. The following figure shows the percentages of each subcategory within this sector of use:

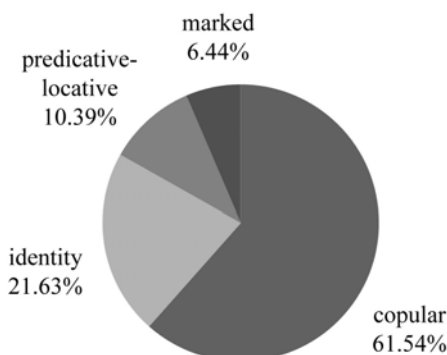


Figure 5: Distribution of the occurrences of *essere* within verbal uses

From a quantitative point of view, data show that the copular use corresponds to nearly 2/3 of the total verbal occurrences of *essere*, and turns out to be the more extensive sector of its variation. Nevertheless, also the identity use (23% of the occurrences) and the predicative-locative one (11%) have to be considered as highly frequent and productive fields. Beyond the percentage within the variation of *essere*, in fact, the absolute data of such uses (more than 1700 occurrences for the identity use, and more than 800 for the predicative locative one) place them within the fundamental lexicon of spoken language (Panunzi 2007).

Therefore, it seems that the distribution of occurrences clashes with Jackendoff's hypothesis, for which the copular and identity values would derive from the locative semantic core. It seems indeed unlikely, also from a cognitive point of view, that so frequent semantic values would derive from an extension of

the relation defined on the locative basis, which represents a much less frequent type of use. In this sense, corpus data are coherent with the theoretical description that has pointed out semantic differences among these three types: differences about the referential properties of the complement, differences about the assignment of thematic roles, lack of state representation in the identificative use.

3.2 Lines of variation

According to data, each primary use of *essere* among those identified by the proposed typology is well attested in the corpus. At this level, the most interesting result coming from the corpus-driven analysis is that the taxonomical types allow for the use of classes with a high semantic variation, that comprehends cases not included in descriptive grammars. Moreover, such a use variation can be explained, since it is theoretically predictable on the basis of the semantic definition of the types.

The concept of copula, identified through the semantic nucleus of the attribution of a property, shows a coherent internal variation: what varies is, properly, the type of the property assigned by the predication (Panunzi 2008). The copular use then represents a “semantic type”, since it projects a regular variation of forms and structures, as shown in the following table:

Table 4. Variation of the copular use

Property type	Example
Quality	<i>'un è mica brutto //</i> [he isn't ugly at all]
Class	<i>[...] gl'era un falegname / gl'era un artigiano //</i> [he was a joiner / he was a craftsman]
Quantity	<i>gli arcani minori e quelli maggiori / sono 52 / [...] //</i> [the minor arcana and the major ones / are 52]
Measure	<i>era 95 chili / quando ci si sposò / i mi marito //</i> [he was 95 kilos / when we married / my husband]
Comparison / evaluation	<i>mio fratello è così //</i> [my brother is this way]
Judgment	<i>è giusto e corretto / che parli l'Istituto Superiore di Sanità / [...] //</i> [it is right and correct / that the ISS should speak ...]

The identity semantic nucleus shows indeed a line of variation related to the ontological class of the entities that constitute the predication arguments. The identity relation can then be established at different ontological orders, from the primary one of the objects, to the one, more complex, of the events. The following table shows the independence of the identity uses variation:

Table 5. Ontological variation in identity uses

Type of entities	Example
Objects	<i>il mercato è questo //</i> [the market is this //]
Places	<i>e qui è San Gottardo / eh //</i> [and here is Saint Gottardo / eh //]
Times	<i>perché qui è il '59 //</i> [because it was 1959 //]
Events	<i>che / era la volta che tu cantavi / ...</i> [that / it was the time that you were singing / ...]

Modern linguistics has given considerable attention to the pseudo-cleft phenomena in syntax (Higgins 1979; Moro 1997; den Dikken 2001), but it has not noticed that such uses properly belong to the identity uses type, and more precisely to the class of the identity of events (Panunzi 2009). The utterance in (11), which contains a pseudo-cleft:

- (11) *quello che non accetti / è che ti giudichi un deficiente //*
[what you don't accept / is that an idiot judges you]

is, in fact, equivalent to the following periphrasis:

- (11') *l'evento (non identificato) che non è possibile accettare equivale a l'evento (identificato) di essere giudicato da un deficiente*
[the (unidentified) event that is not possible to accept is equal to the (identified) event that an idiot judges you]

as foreseen by the identity interpretation of verb *essere*.

For what regards the predicative-locative uses, characterized by the assignment of a thematic role to the complement; the corpus-driven analysis confirms the cognitive tendency towards the extension of the thematic structures starting from the locative basis. The locative relation is extended to the whole variation of *essere* predicative uses, on the same cognitive spaces foreseen by the Gruber-Jackendoff's hypothesis (spatial location, temporal location, possessive relation). The whole variation is shown in table 6.

Also in this case, it has been possible to find in the corpus types of use that which was not foreseen through hypotheses. In fact, the "benefactive" and the "causative" relations fall within the predicative-locative variation of *essere*, since they represent the cases in which the *theta* roles assigned to the complement is, respectively, "benefactor" and "cause".

Table 6. Variation within the predicative-locative uses

Spatial location	<i>qui / siamo alla foce del fiume //</i> [here / we are at the mouth of the river]
- metaphorical	<i>poi / le persone / che erano già in graduatoria permanente / hanno fatto anche la domanda / ai presidi //</i> [and then /the people/ who already were in the permanent list / also addressed the application / to the head teachers]
- “co-localization”	<i>eravamo sempre con Adriana / e il suo bambino //</i> [we always were with Adriana and her child]
Temporal location	<i>sì / la [/] la fiera / l' è il secondo martedì di luglio //</i> [yes / the fair is on the second Tuesday of July]
Possessive relation	<i>questi / erano della mamma //</i> [these ones / belonged to my mother]
Benefactive relation	<i>è per la gente che si deve divertire //</i> [it is for the people who want to have fun //]
Causative relation	<i>non so se è per questione di soldi / di interessi / per questione di culture forse diverse //</i> [I don't know if it is because of money / of profit / because of different cultures //]

The following table summarizes the verbal use typologies of *essere* found in the corpus, related to the different verb values and with the involved line of variation. The consistency of the lines of variation found within the copular, identity and predicative-locative uses constitutes a semantic argument that build up the independence of each sector of variation.

Table 7. Internal lines of variation in *essere* types of use

Type of use	Semantic values	Line of variation
Copulative	<i>Property attribution</i> (extensional relation)	Type of the property
Identity	<i>Identity relation between different referential expressions</i> (intensional relation)	Ontological class of the entity to which the expressions refer
Predicative-locative	<i>Locative relation</i> (thematic relation)	Cognitive domain within which the relation between the theme and the object of reference is set up

Each concept is then applied in linguistic use with a regular, predictable and independent variation. The hypothesis of a polysemy in the meaning of the verb *essere* receives a confirmation by the finding of different types of variation with respect to each *sema*.

3.3 Conclusions

In this study it has been showed that is possible to identify in the corpus the main semantic interpretations of verb *essere* on the basis of positive linguistic evidences. The adopted taxonomical criteria, derived from an extensive survey on the logical and linguistic tradition, turned out well-grounded with respect to the data, and permitted an effective distinction of the corpus instances among copular, identity and predicative-locative uses, with a very low rate of underdeterminacy.

In short, we can claim that the classes of the semantic variation correlate with the types of use of *essere* and determine their differential properties:

- the type of *essere* that varies with respect to the “type of assigned property” represented by its complement correlates with the interpretation of an extensional relation (inclusion in a class) and with a lack of referentiality of the complement;
- the type of *essere* that varies with respect to the “ontological type” of the complement correlates with the interpretation of an intensional identity relation, with the referentiality of the complement and its absence of thematic role, and with the lack of state representation;
- the type of *essere* that varies with respect to the “cognitive domain” to which the complement refers correlates with the referentiality of the complement, associated to the assignment of a thematic role, and with a state representation.

The assumption of a classification criterion that is only based on the complement properties assures us that the identification on corpus of the types is assumed on the theoretical level. The proposed taxonomy of such a basis is then objectified by the productivity of these classes, by the quantitative consistency of the classification and by the observation of the internal variation projected by each taxonomical type.

In particular, this latter aspect constitutes an independent criterion for the validation of the taxonomy. The concept of variation, in fact, presupposes the existence of a super-ordinate level, characterized by general properties that are able to identify a constant (mostly at the semantic level) within which it is possible to observe the variation itself. In other words, whereas there is a variation, a higher level category comes to be objectified, and it constitutes the stable frame within which the variations take place. Such a category is the lines of variation, which on the one hand defines the variation domain, and on the other hand constitutes an *a posteriori* validation for the adopted taxonomical criteria.

Notes

¹ In confirmation of this, Aristotle claims that every assertoric proposition can be translated in a copular sentence (De Interpretatione 12.21b, 9; the example concerns the proposition *a man walks*, that should be equivalent to *a man is walking*). This should demonstrate that every *rhema* is analyzable into its basic components: the “tense”, expressible through the verb *to be*, and the “predicate”, expressible through a nominal element.

² A millenary tradition, in effect, if we consider that is the one still prevailing within the historical linguistics, at least till the middle of 20th century (see Meillet 1921; Benveniste 1960).

³ In his description of the linguistic meaning components, in fact, Frege (1892a) claimed that the *concept* corresponds to the *reference* (*Bedeutung*) of a predicate, whereas the object corresponds to the *reference* of an singular term.

⁴ Den Dikken (2001) reports the following terms oppositions, all bound to the semantic distinctions of copular sentences: *classifying/identifying* (Kruisinga / Erades, 1953); *intensive/extensive* (Halliday 1967, Huddleston 1971); *non-equational/equational* (Bolinger 1972); *ascriptive/equative* (Lyons 1977); *attributive/identificational* (Gundel 1977); *predicational/specificational* (Akmajian 1979, terms used also in Higgins 1979 and Declerck 1988). All the same, it has to be underlined that it is not possible to trace back all these term pairs to the same type of duality.

⁵ In this work, as well as in the greater part of the following studies, the *predicational* and *specificational* classes are in relation with the possible interpretations of pseudo-cleft sentences.

⁶ See also den Hertog (1903).

⁷ According to the authors, such an interpretation would be applicable also to the classical identity statement: *the morning star is the evening star*. Blom & Daalder claim that the constituent on the left side would be described in a narrower way than the other, and that, all in all, such a proposition would affirm that the property of “being the evening star” is assigned to the element “the morning star”, as in a standard *predicational* sentence.

⁸ The notion of *raising verb* foresees the presence of two arguments: a clausal one (a subordinate clause, with predicative function) and a nominal one (as the subject). The peculiarity of these constructions is that the nominal argument in the position of subject of the *raising verb* is actually an argument of the verb of the subordinate clause. For example, in the sentence *he seemed to do something*, the constituent in the subject position of the verb *to seem* is rather the subject of the verb *to do* of the bound clause (explicitly, *it seemed that he did something*).

⁹ A *small clause* (Chomsky 1981) is defined as a minimal predicative structure, that contains a predicate (mostly, a nominal one) and its arguments, with no further specification of [mode], [tense] and [aspect] features. The following examples contain, in square brackets, some cases of *small clauses*: (a) I consider [*you intelligent*]; (b) Mark painted the [*bike red*]; (c) She saw [*Henry drink a glass of whiskey*].

¹⁰ In the exemplification we will adopt the terminology proposed by Mikkelsen (2005).

¹¹ For other proposals, see Heggie (1988a, 1988b; *predicate topicalization*), Heycock and Kroch (1999; *subject raising*) and Rothstein (2001; *transitive structure*); see Mikkelsen (2005) for a detailed state of the art of various interpretation given for copular clauses in the generative grammar framework.

¹² Gruber refers to *identificational pattern e positional pattern*, and identifies, within each *pattern*, some parallelism between *Motional* and *nonMotional* verbs. His proposal defines two distinct primitives within which a variation of the “movement/change” semantic feature operates. The verb *to be* is then present in both patterns, as a verb that carries the *nonMotional* feature. This way, the Gruber’s hypothesis differs from Jackendoff’s one since it proposes two distinct interpretations for the verb *to be*, and it refuses the possibility to retrace the categorization processes to the spatial semantic matrix.

¹³ The extension of the verb variation turns out evident if we consider that the “complement” of the structure can be constituted by: (a) an adjective; (b) a noun phrase (both definite and indefinite); (c) a nominal form of a verb (participle in adjectival function, infinitive); (d) a pronoun (personal, indefinite, demonstrative); (e) a prepositional phrase; (f) an adverb; (g) a numeral.

¹⁴ What role is then performed by the identity complement? Given the nature of the involved relation, it is possible to hypothesize that its role is restricted to provide, as referential term, the properties that constitute the sense that is put in intensional relation with the one expressed by the subject.

¹⁵ This parallelism is included in the original Akmajan’s proposal, that, referring to the Fregean distinction, suggests to use the symbol [=] as the logical transcription of the *specificational* relation, and [is] for the *predicational* one.

¹⁶ Such an utterance would be treated as the analogue one *his brother’s car is red*, in which the subject is constituted by an identified element. On the other hand, a constituent as [*the thing that I prefer*] can occur also in sentences with a completely different type of predicate, maintaining the same subject function: *the thing that I prefer is falling on the floor*; *the thing that I prefer broke*. The interpretation of the constituent in subject position as a predicative element has to give account of such a distribution: on the contrary, it would be no explanation for the fact that the same constituent in the same position should have the value of a predicate in (5) and of a normal subject in the just mentioned sentence.

References

- Akmajan, A. 1979. *Aspects of the grammar of focus in English*. New York: Garland Press.
- Benveniste, É. 1960. Être at avoir dans leurs fonctions linguistiques. *Bulletin de la Société de Linguistique* 55: 113-34.
- Blom, A. and S. Daalder. 1977. *Syntaktische theorie en taalbeschrijving*. Muiderberg: Coutinho.

- Bolinger, D. 1972. A look at equations and cleft sentences. In E.S. Firchow, K. Grimstad, N. Hasselmo and W. A. O'Neil (eds), *Studies for Einar Haugen*. The Hague: Mouton, 96-114.
- Cresti, E. 2000. *Corpus di italiano parlato*, voll. I-II, CD-Rom. Firenze: Accademia della Crusca.
- Cresti, E. and M. Moneglia (eds). 2005. *C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages*, vol. + DVD. Amsterdam: Benjamins.
- Declerck, R.. 1988. *Studies on copular sentences, clefts and pseudoclefts*. Dordrecht: Leuven University Press and Floris.
- den Dikken, M. 2001. Specificational copular sentences and pseudoclefts. A case study. In *The blackwell companion to syntax*, 5 voll., M. Everaert and H. van Riemsdijk, vol. IV, cap. 61. Oxford: Blackwell.
- den Hertog, C. H. 1903. *Nederlandse spraakkunst*. Amsterdam: W. Versluys.
- Donnellan, K.S. 1966. Reference and definite descriptions. *Philosophical review* 77: 281-304.
- Fillmore, Ch. J. 1968. The case for case. In E.W. Bach and R.T. Harms (eds), *Universals in Linguistic Theory*. New York: Holt, Rinehart & Winston, 1-88.
- Frege, G. 1982a. Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik* 100: 25-50.
- Frege, G. 1982b. Über Begriff und Gegenstand. *Vierteljahrsschrift für wissenschaftliche Philosophie* 16: 192-205.
- Gruber, Jeffrey S. 1976. *Lexical structures in syntax and semantics*. Amsterdam: North Holland.
- Gundel, J.K. 1977. Where do cleft-sentences come from? *Language* 53: 543-559.
- Halliday, M.A.K. 1967. Notes on transitivity and theme in English (Part 2). *Journal of Linguistics* 3, 2: 199-244.
- Heggie, L.A. 1988a. The Syntax of Copular Structures. PhD diss., USC.
- Heggie, L.A. 1988b. A unified approach to copular sentences. In H. Borer (ed.), *Proceedings of WCCFL 7*. Stanford: Stanford Linguistics Association, 129-142.
- Heycock, C. and A. Kroch. 1999. Pseudocleft connectedness: Implications for the LF interface level. *Linguistic Inquiry* 30, 3: 365-397.
- Higgins, R.F. 1979. *The Pseudo-cleft Construction in English*. New York: Garland.
- Huddleston, R. 1971. *The sentence in written English*. Cambridge: Cambridge University Press.
- Jackendoff, R. 1972. *Semantic interpretation in generative grammar*. Cambridge (MA): The MIT Press.
- Jackendoff, R. 1983. *Semantics and cognition*. Cambridge (MA): The MIT Press.
- Kruisinga, E. and P.A. Erades. 1953. *An English grammar, Vol. I: Accidence and syntax, first part*. Groningen: Noordhoff.
- Lyons, J. 1977. *Semantics*. Cambridge: Cambridge University Press.
- Meillet, A. 1921. *Linguistique historique et linguistique générale*. Paris: Champion et Klincksieck.
- Mikkelsen, L. 2005. *Copular clauses. Specification, predication and equation*. Amsterdam: Benjamins.
- Moro, A. 1988. Per una teoria unificata delle frasi copulari. *Rivista di Grammatica Generativa* 13: 81-110.

- Moro, A. 1997. *The raising of predicates: predicative nouns phrases and the Theory of clause structure*. Cambridge: Cambridge University Press.
- Panunzi, A. 2005. "Essere" e "esserci" nella lingua italiana d'uso. Indagine su un corpus di parlato spontaneo e primi confronti interlinguistici nelle lingue romanze. In I. Korzen (ed.), *Lingua, cultura e intercultura: l'italiano e le altre lingue* (Copenhagen Studies in Language 31). Copenhagen: Samfundslitteratur press, 255-266.
- Panunzi, A. 2006. L'analisi corpus-driven delle strutture ESSERE + Preposizione in italiano. Costrutti grammaticali e variazione marcata del predicato. In E. Corino, C. Marellò and C. Onesti, *Proceedings of 12th EURALEX international congress*. Alessandria: Edizioni dell'Orso, 1021-1028.
- Panunzi, A. 2007. Il verbo *essere* nella lingua italiana: analisi della variazione d'uso in un corpus di parlato spontaneo. PhD diss., University of Turin.
- Panunzi, A. 2008. Strutture copolari dell'italiano parlato. In M. Pettorino, A. Giannini, M. Vallone and R. Savy (eds), *La comunicazione parlata. Atti del congresso internazionale*. Napoli: Liguori, 626-644.
- Panunzi, A. 2009. Strutture scisse e pseudoscisse: valori d'uso del verbo *essere* e articolazione dell'informazione nell'italiano parlato. In A. Ferrari (ed.), *Sintassi storica e sincronica dell'italiano. Subordinazione, coordinazione, giustapposizione* (Atti del X Congresso della SILFI, 2008). Firenze: Cesati, 1121-1137.
- Russell, B. 1905. On denoting. *Mind* 14: 479-493.
- Russell, B. 1919. *The philosophy of mathematics*. London: Allen & Unwin.
- Salvi, G. 1991. Le frasi copulative. In L. Renzi and G. Salvi (eds), *Grande grammatica italiana di consultazione. Vol. II, I sintagmi verbale, aggettivale, avverbiale*. Bologna: Il Mulino, 163-189.
- Stowell, T. 1978. What was there before there was there? In D. Farkas, W. M. Jacobsen and K. W. Todrys (eds), *Papers from the Fourteenth Regional Meeting of the Chicago Linguistic Society*. Chicago: Chicago Linguistic Society, 458-471.
- Strawson, P.F. 1950. On referring. *Mind* 59: 320-344.
- Tognini Bonelli, E. 2001. *Corpus linguistics at work*. Amsterdam: Benjamins.
- Van Peteghem, M. 1991. *Les phrases copulatives dans les langues romanes*. Wilhelmsfeld: Gottfried Egert Verlag.
- Verheugd, E. 1990. *Subject arguments and predicate nominals: a study of French copular sentences with two NPs*. Amsterdam: Rodopi.

CHIEDE

A SPONTANEOUS CHILD LANGUAGE CORPUS OF SPANISH¹

Marta Garrote Salazar, Antonio Moreno Sandoval

Autonomous University of Madrid

1. CHIEDE

The spontaneous child language corpus, CHIEDE, is made up of around 60.000 words. About a third of the whole corpus is comprised of child language and the rest of adult speech. The main feature of CHIEDE is the spontaneity of interactions: texts are recordings of communicative situations in their natural context. The resource is presented in different formats: an orthographic transcription, an automatic phonological transcription, an XML tagged version and the text-sound alignment. We also provide results obtained through statistical methods, of data from the annotated texts.

CHIEDE fulfills all the requirements of a modern spoken language corpus. It is in an electronic format, allowing the storage and manipulation of data and the interchange with other interested researchers. Its proportioned design and diversity – sex, age and communicative situation variables – guarantees that it is linguistically representative. Its presentation on a web site makes it freely available (<http://www.llf.uam.es/chiede>). Finally, its classification structure allows it to be properly utilized.

1.1 Corpus design

To record the corpus CHIEDE, we entered into a collaboration with a Spanish public school, which allowed us to record children from Infant School groups (from 3 to 6 years old). CHIEDE is a transversal corpus, made up of three groups of individuals, divided by ages (from 3 to 4 years old, from 4 to 5 and from 5 to 6).

Our corpus represents two kinds of interactions: *spontaneous collective conversations*, recorded at a daily activity in classroom, and *personal interviews* in which an adult talks to a single child.

CHIEDE consists of 58.163 words, in 30 texts, with 7 hours and 53 minutes of recordings and 59 child participants. Each recording is aligned with its corresponding orthographic transcription, including a header with metadata or sociolinguistic and contextual information. Apart from the audio and the text files, two other kinds of files are included: those in which an automatic phonological transcription has been carried out and those where the text appears in XML format with the morphosyntactic annotation. The files are identified with a name where the age of the child participant is specified.

Table 1. CHIEDE measurements

CHIEDE	MINUTES	URNS	UTTERANCES	WORDS
	473'11''	10.042	15.444	58.163

A goal of the project was to obtain a balanced corpus between participants and communicative situations. For this purpose, the same number of children of each age group (three, four and five years old) were recorded and placed into three subcorpora; each subgroup contained the same number of boys and girls. The number of recording hours and words are similar for the collective interventions and the individual ones. The total number of participants in the interviews is 24, in three groups of 8 individuals having the same age, further divided by sex. For the collective conversations, the number of participants is not as balanced, as it depended on the number of pupils at classroom. Thus, the three-year-old group was made up of 21 pupils, and the four-year-old group, 21 as well, while the five-years-old had only 17 pupils. The three groups make 59 participants.

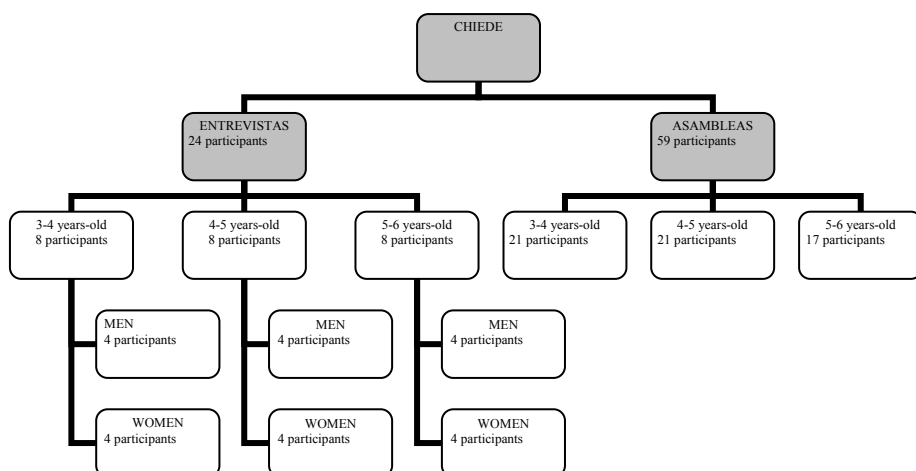


Figure 1. CHIEDE design

1.2 Format

We based our annotation criteria on the annotation system developed for the C-ORAL-ROM project, which is based upon the CHAT format. We also made consistent use of its XML schema, thus guaranteeing textual standardization throughout the corpus.

Each file contains metadata and a transcription. The metadata includes sociolinguistic information. The transcription is orthographic and includes tags that mark disfluencies, noises, overlaps and prosodic units; it is structured by turns introduced by a three-letter code which identifies the participant. Morphosyntactic tagging is provided at a separate level. Figure 2 shows a text fragment.

```
@Title: Jorge y Marta
@File: JOR4
@Participants: JOR, child, (man, 5:0, 1, Ciudad Real)
TEA, adult, (woman, B, 3, Madrid)
@Birth_Date: 20/02/2001
@Date: 22/02/2006
@Place: Ciudad Real
@Situation: conversation in an empty classroom at school.
@Topic: daily matters
@Source: CHIEDE
@Class: informal, family/private, dialogue (child not-known adult)
@Length: 13 '39"
@Words: 2097
@Acoustic_quality: A
@Transcriber: Marta
@Revisor: Ana and Marta
@Comments: JOR (middle class; birth order: 1st)

*JOR: aquí ///
*TEA: a ver si puedes /// ¿ cuántos años tienes Jorge ?
*JOR: &eh tengo -> / cuatro ///
*TEA: cuatro /// que fue tu cumple el otro día /// ¿ a que sí ?
*JOR: cinco sí ///
*TEA: ¡ah! ¿ cinco ? ¿ o cuatro ?
*JOR: bueno / hoy &cum [/] mañana cumplí cinco // pero ahora / tengo cuatro ///
```

Figure 2. CHIEDE fragment

2. The computational tool

The orthographic transcription is not sufficient in itself. At LLI-UAM, various computational tools have been developed to transform the plain format of a

transcription text into a proper tagged format to match the metadata with the lexical elements and calculate the corresponding statistics. In this section we will explain this process, and later we will see how the data extraction process has been carried out.

2.1 Results of CHIEDE annotation

Once sampling, transcription, and revision have been completed, the corpus is annotated, including phonological, morphological and part of speech tags. The following statistics are gathered:

- Mean Length of Utterance (MLU): in syllables and phonemes
- Frequency of use of lemmas and word forms
- Type/token ratio: lexical diversity
- Most frequent words
- Most frequent categories

All these data allow us to describe language use and establish linguistic behavior patterns.

2.1.1 Data extracted from the morphosyntactic tagger

The morphosyntactic tagger deals with three linguistic levels: the morphological, the syntactic and the lexical ones. We used the morphological information included in the lexicon entries to obtain the different lemmas that appear in the corpus, plus the part of speech information added to each word or multiword.

Table 2 presents data related to the whole corpus, including adult language, sorted by age group:

Table 2. Word forms and Lemmas by ages

	Total words	Different word forms	Different lemmas	Word form/lemma ratio
Adults	36.905	2.910	1.804	1,61
Group 3	5.713	985	718	1,37
Group 4	6.374	1.155	839	1,38
Group 5	8.993	1.450	1.056	1,37

In this table, the first column includes the total number of words for each group, the second column, the number of different words that appear and, the third one, the number of different lemmas. The total number of words for the child subcorpus is 21,080. The last column shows the lexical diversity, that is, the ratio of word forms for each lemma. This ratio scarcely changes for the three child groups, while it does change for the adult subcorpus, as word inflection increases in adult language.

Differences can be seen between the three child groups regarding the increase of word forms and lemmas. For the three-to-four period, there is an increase of 170 word forms and 121 lemmas; for the four-to-five one, this number goes up to 295 word forms and 217 lemmas. This shows that for the first of these periods – from three to four years old – learning is slower than for the second one.

We also find that, although the number of recording hours is similar for the three groups, there is a difference of 3,219 words between the three-year-old group and the five-year-old one. Five-year-old children are already able to have a pseudo-adult conversation. This can be more clearly appreciated by inspecting the MLU in syllables and phonemes. As discussed in the next section, the MLU in phonemes is 10.29 for the three-years-old group, while for the five-years-old one it increases 3.82 points.

Table 3. Word frequency by age groups

Word frequency (10 first words)					
Age group: 3		Age group: 4		Age group: 5	
Word	Frequency	Word	Frequency	Word	Frequency
y	322	y	341	y	502
no	271	que	213	a	289
sí	204	el	206	no	274
el	162	sí	195	que	259
yo	157	no	179	el	251
un	148	a	165	sí	248
la	141	la	153	la	217
a	128	de	148	me	189
me	119	en	136	de	172
se	106	mi	131	se	171

Regarding word frequency (Table 3), it is worth pointing out the abundant use of the copulative conjunction “y”. This is due to two facts: firstly, this is the first coordination strategy that children learn; secondly, “y” have a different use apart from being a copula, and it is usually used as a discourse marker, not only relational, but also a subjectivity one: it is used by the speaker as a turn-taking strategy. The following example clearly shows this last function of “y”:

- *DAI: y [/] y yo sé +
- *TEA: ¡ qué bonito ! y a ver María ///
- *DAI: y mi &pa +
- *TEA: espera / espera a María /// a ver María ///
- *MRI: mi papá no le regala / nada a mi mamá ///
- %alt: (6) na

*TEA: pero bueno +

*MRI: ¬ y mi mamá sí ///

Apart from that, it is also worth noting the high frequency of negative and affirmative adverbs (“sí” and “no”) and the use of pronouns and determiners. In short, in any spoken language corpus the most frequent elements are grammatical categories; the lexical ones appear after the first twenty positions in CHIEDE. Possibly most striking is the lack of discourse markers, as “bueno”, “¿sabes?” or “¿no?”, so frequent in adult spoken language.

If we do not take into account the grammatical categories, the acquisition of lexical categories as verbs, nouns or adjectives is quicker from four to five years old than from three to four (Table 4). In the second period (four to five years old), the number of new lemmas more than doubles. Thus we see that there is a great increase in lexical category acquisition during the fourth and fifth-year period.

Table 4. Lexical categories by age

		Verbs	Nouns	Adjectives
Age group: 3	Word forms	906	836	157
	Lemmas	138	313	50
Age group: 4	Word forms	937	998	141
	Lemmas	143	357	57
Age group: 5	Word forms	1324	1340	161
	Lemmas	190	439	72

The two following graphs illustrate the growth of word forms (Figure 3) and lemmas (Figure 4) for the three age groups.

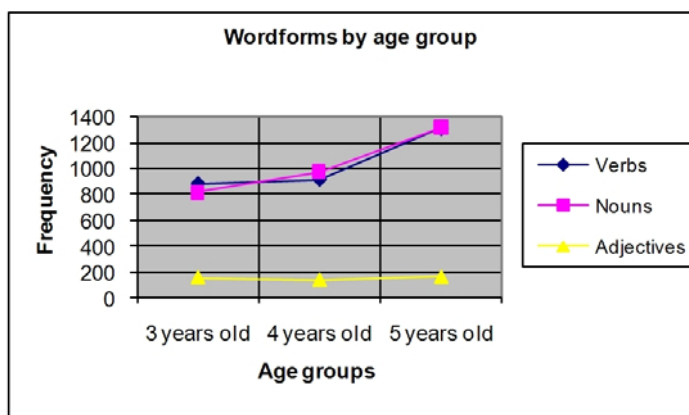


Figure 3. Word forms by age group

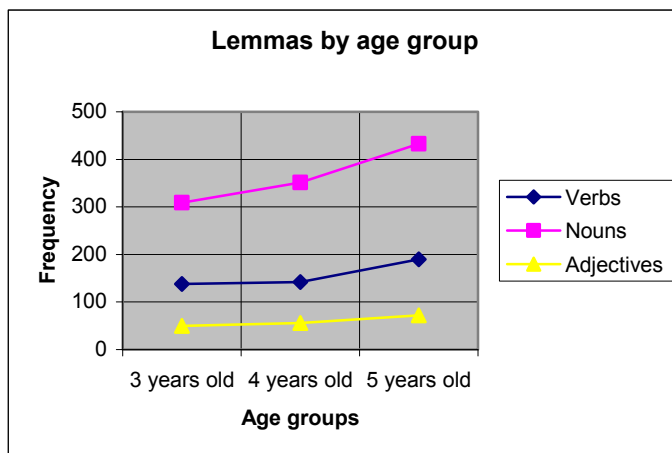


Figure 4. Lemmas by age group

The last thing we are going to deal with in this section is the analogy errors typical of child language. This phenomenon takes place during the first years of acquisition, when the child has developed adequate skills in morphology and inflection. Once the general morphological rules are learnt, the child tends to apply them without discriminating, for instance, between regular and irregular forms. These analogy errors have been always one of the main arguments of those who reject behaviorist theories of learning. They argue that if the child simply reproduces what he listens, this kind of errors would not be possible; on the contrary, it is easier to find an explanation for the analogy errors if we consider that the child is able to learn the morphological rules of his/her language and apply them in a creative way, innovating and not imitating. In fact, it is interesting how, after the adult correction, the child generally does not rectify and goes on keeping his proposal.

In CHIEDE, which totals 21.080 words, there are 31 analogy errors, that is, 0.15% of the whole². If we reduce the list of word forms to lemmas, we can see that there is a total of 17 specially problematic: “acordar”, “cerrar”, “conducir”, “decir”, “escribir”, “hacer”, “ir”, “leer”, “mentir”, “morder”, “poder”, “poner”, “querer”, “ser”, “soltar”, “tener”, “traer”. According to our data, this phenomenon is more frequent in four-year-old children, being reduced in five-year-old ones. So it seems to be that it is at the age of five when children start to understand the irregular inflection of Spanish verbs. In our corpus, the most persistent irregular verb form is “hacío/hizo”.

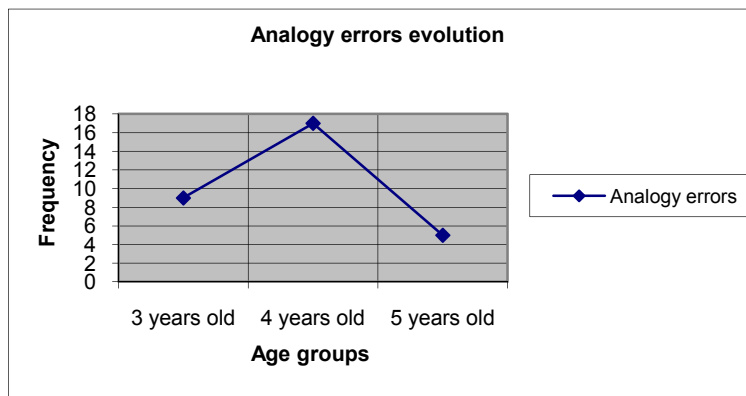


Figure 5. Analogy errors evolution

2.1.2 Data extracted from the phonological transcription

Psycholinguistics, in the attempt to explain the acquisition process of a first language, usually resorts to different measurement patterns. Childhood development stages are calculated, measuring the number of words that comprise the lexicon of a particular age, the number of phonemes they handle, or the most frequent syllables.

Most experts place the phonological system acquisition period from nine months to four years old. However, most research on child language does not consider individuals older than 36 months, and the child language description stops when the child reaches the age of three years old, though it is held that the acquisition process continues until the beginning of puberty. That is the reason why we consider our corpus of interest for the scientific community, as we provide data belonging to children from three to six years old.

In Table 5, we present the phonological data of the three age groups of our corpus. The total number of phonemes is 75,535, and the MLU in phonemes is 12.72 per turn. It is striking in Table 5 the fact that at the age of three, the child has already acquired the complete Spanish phonological system.

The table we present does not show the phonemes' order of appearance, since they have been already acquired by the children that participate in the corpus, but rather their frequency of use. It is known that children tend to use the phonemes they better know, while they avoid those which are harder. In the frequency table, we can see how the phonemes in the latest rows are the least frequent in Spanish, and therefore it is normal that their frequency of use is lower than that of the most usual phonemes.

However, the numbers increase as children ages do. This argues that the linguistic acquisition process is still active from three to five years old, and that research on first language acquisition must not stop at the age of 36 months.

Apart from the phonological data extracted from CHIEDE, we have also added a fourth column that includes the same information from C-ORAL-ROM (Moreno et

al. 2006). Although the data are similar for both child and adult language, we can see, especially in the first positions of the table, a higher similarity between the five-year-old group and the adult one than between the last one and the three and four-year-old group.

To appreciate more clearly the child phonological system – at least the one presented by the individuals that participated in our corpus – we present five independent graphs. Table 5, reported in Appendix to this text, includes the vocals frequency and their evolution from one group to another.

The following graphs show the consonants frequency according to their manner of articulation: plosive, nasal, liquids and fricatives (and the affricate tʃ).

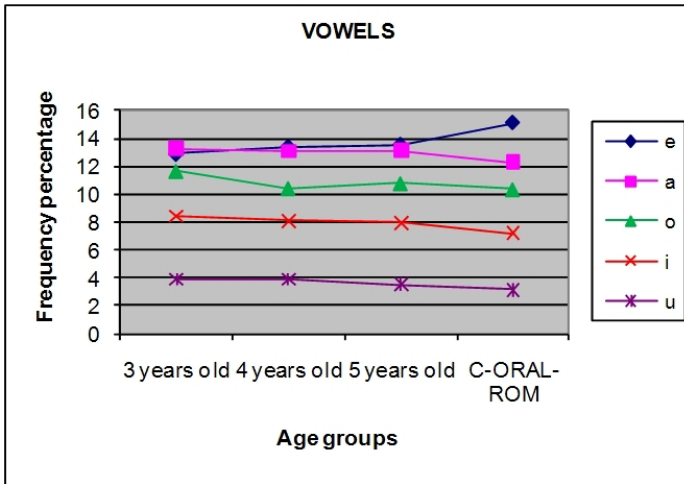


Figure 6. Vowels by age group

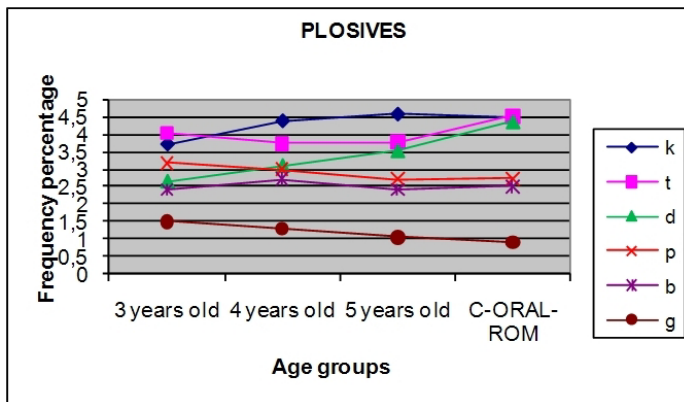


Figure 7. Plosives by age group

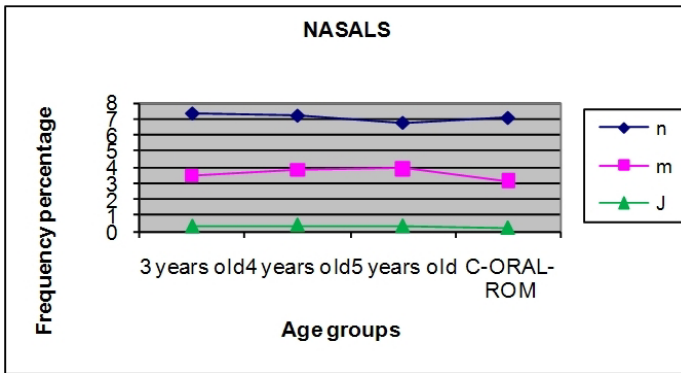


Figure 8. Nasals by age group

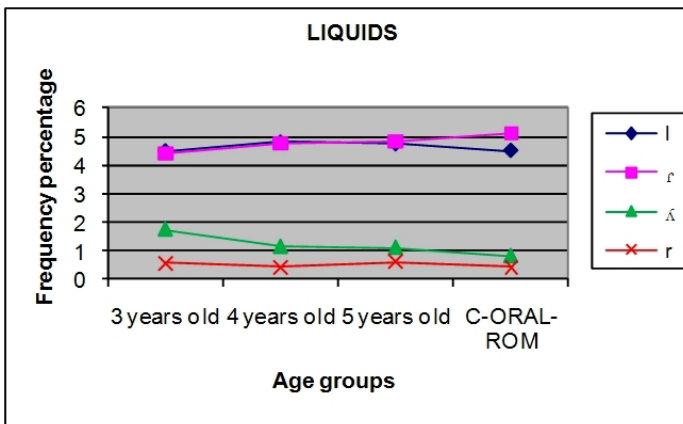


Figure 9. Liquids by age group

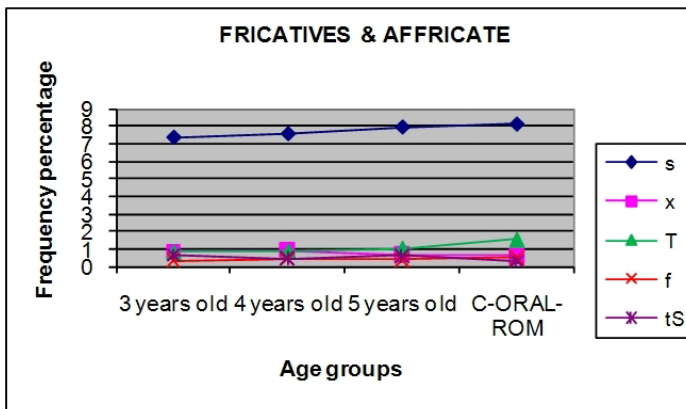


Figure 10. Fricatives and the affricate tʃ by age group

Another possibility that offers the automatic phonological transcriber is the segmentation of words into syllables. In this way, we can quickly and reliably know the total number of syllables that comprise our corpus, which ones are those syllables and which is their frequency of use. The total number of syllables is 35,086 and the MLU, 5.91. With these data, we can easily calculate the Mean Length of Utterance in phonemes and syllables for each age group. In the following table, we present the exact figures and the increase percentage from three to five years old. Figure 11 shows this MLU increase.

Table 6. Mean Length of Utterance

	Mean Length of Utterance	
	Phonemes	Syllables
Age group: 3	10.29	4.88
Age group: 4	13.57	6.26
Age group: 5	14.11	6.49

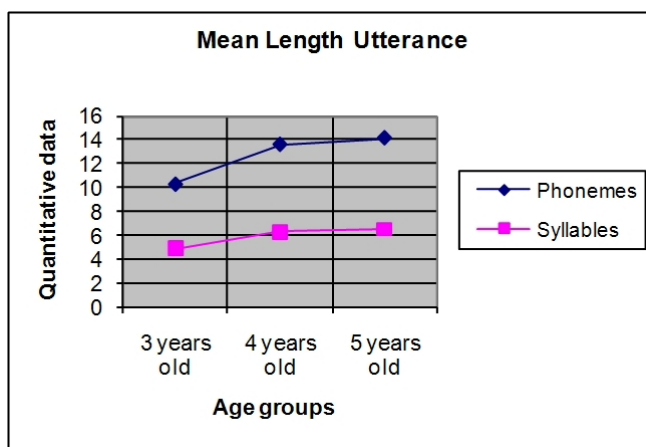


Figure 11. MLU increase

3. Using xml-tagged corpus for relating meta-data linguistic features

The original annotation of CHIEDE has been designed to take into account a wide range of phenomena, including the acoustic ones (prosodic marks, noises, etc.) that can be used by the speech technology community.

Our aim in this experiment was searching for relevant lexical units in two subcorpora: one of adult language and another of child language. The first step consists on the segmentation of each speech turn in utterances to prevent wrong-formed word groups. This task is similar to tokenization in written language corpora.

The utterance segmentation is also necessary to delimit the context which the morphosyntactic tagger uses to disambiguate.

A computational program generates a new tagged corpus with a single tag: UNIT (utterance), with attributes for *speaker*, *startTime* and *endTime*. In Figure 12, we can see the result of this process of XML conversion. The numbers stand for the sound alignment times, expressed in milliseconds. In this way, each utterance is limited, identifying its corresponding speaker.

The next step is the morphosyntactic annotation from the XML file. The morphosyntactic tagger procedure is the following (Guirao et al. 2006):

- Unknown word detection.
- Lexical processing: the program splits the fused words (amalgams and verbs with clitics).
- Multi-word recognition: through a lexicon.
- Single word recognition.
- Unknown word recognition.
- Disambiguation phase 1: a feature-based Constraint Grammar resolves some of the ambiguities.
- Disambiguation phase 2: a statistical tagger (TnT tagger, Brants 2000) resolves the remaining ambiguous and unknown words.

```
<UNIT speaker="JOR" startTime="0" endTime="4.482">aquí </UNIT>
<UNIT speaker="TEA" startTime="4.482" endTime="7.655">a ver si puedes </UNIT>
<UNIT speaker="TEA" startTime="7.655" endTime="9.246">¿ cuántos años tienes Jorge ?
</UNIT>
<UNIT speaker="JOR" startTime="9.246" endTime="12.459">&eh tengo -> / cuatro </UNIT>
<UNIT speaker="TEA" startTime="12.459" endTime="13.131">cuatro </UNIT>
<UNIT speaker="TEA" startTime="13.131" endTime="14.267"> que fue tu cumple el otro día
</UNIT>
<UNIT speaker="TEA" startTime="14.267" endTime="14.817">¿ a que sí ? </UNIT>
<UNIT speaker="JOR" startTime="14.817" endTime="15.667">cinco </UNIT>
<UNIT speaker="TEA" startTime="15.667" endTime="16.411">¡ah! </UNIT>
<UNIT speaker="TEA" startTime="16.411" endTime="17.09">¿ cinco ? </UNIT>
<UNIT speaker="TEA" startTime="17.09" endTime="17.755">¿ o cuatro ? </UNIT>
<UNIT speaker="JOR" startTime="17.755" endTime="23.601">bueno / hoy &cum [/] mañana
cumplí cinco // pero ahora / tengo cuatro </UNIT>
```

Figure 12. XML Conversion

The final result after the tagger revision is a XML file where the text is morphosyntactically analyzed:

```
<Text>
<p>
<f h="JOR" st="0.0" et="4.482" id="1">
```

```

<sf t="enu" id="1-1">
<w cat="P" lem="aquí" id="1-1-1"> aquí </w>
</p>
<p>
<f h="TEA" st="4.482" et="7.655" id="2">
<sf t="enu" id="2-1">
<w cat="MD" lem="a ver" id="2-1-1"> a ver </w>
<w cat="C" lem="si" id="2-1-2"> si </w>
<w cat="V" lem="poder" tie="pres_ind" num="sing" per="2" id="2-1-3"> puedes </w>
</p>
<p>
<f h="TEA" st="7.655" et="9.246" id="3">
<sf t="int" id="3-1">
<w cat="PUNCT" lem="¿" id="3-1-1"> ¿ </w>
<w cat="P" lem="cuántos" gen="masc" id="3-1-2"> cuántos </w>
<w cat="N" lem="año" gen="masc" num="plu" id="3-1-3"> años </w>
<w cat="V" lem="tener" tie="pres_ind" num="sing" per="2" id="3-1-4"> tienes </w>
<w cat="NPR" lem="Jorge" id="3-1-5"> Jorge </w>
<w cat="PUNCT" lem="?" id="3-1-6"> ? </w>

```

Thus, each word in the corpus can be related to the speaker. The file keeps in the header all the socio-contextual information, being possible to create as many subcorpora as different features appear in the header – an adult language subcorpus, a child language subcorpus, etc. After the division into subcorpora, it is possible to calculate the occurrences (tokens) for each lexical unit (types). The procedure can be applied to any type of linguistic information that had been annotated in the corpus.

3.1 Extracting word clusters

If we calculate the statistics on each unit directly, the result would not be correct, as the pluri-verbal lexical elements (that is, idioms) would not be included in the count. Frequent discourse markers like “por ejemplo”, “o sea” or “es decir” would not appear if we work on lexical units made up of a single word. To solve this problem, it has been created an idioms list by categories, including nominal compounds (“fin de semana”). Each idiom is considered a lexical unit, equivalent to a single word.

3.2 Applying the statistics of surprise

To identify distinctive words, lemmas, or categories of a given subcorpus we have used the log-likelihood ratio test proposed by Dunning (1993). This method does not assume normal statistical distributions of units in a corpus. Instead, the log-likelihood ratio λ assumes a binomial distribution more appropriate for rare but

distinctive words. In addition, this test does not need balanced subcorpora for comparison.

This method has been successfully applied for finding collocations (Dunning 1993) and terms (Daille 1994). In order to test the method to find distinctive units in specified domains, we can work on two hypotheses:

Two registers (or subcorpora) show no difference in distinctive units (*Null hypothesis*).

- i. For a given subcorpus, we can find out distinctive units (*Alternative hypothesis*).
- ii. We applied the test to two well-defined subcorpora: adult and child language. Results are shown in Tables 7 and 8.

Table 7. Distinctive word forms in adult language

WORDS	ADULTS (36.905)	CHILDREN (21.080)	DUNNING
qué	1.123	108	510.29
te	743	59	373.43
a ver	371	23	207.58
bien	304	14	189.00
ah	270	18	146.32
claro	231	15	126.53
tú	264	27	113.88
has	184	9	112.02
tu	197	14	103.64
cómo	249	27	103.26

Table 8. Distinctive word forms in child language

WORDS	CHILDREN (21.080)	ADULTS (36.905)	DUNNING
mi	334	24	524.66
yo	417	166	300.54
sí	647	428	255.77
me	423	248	198.53
tengo	130	28	141.16
Candi	67	9	88.55
porque	150	86	71.99
un	431	424	71.25
padre	60	13	64.86
he	79	27	64.09

Results confirm the alternative hypothesis and the suitability of the Dunning test for the task. Most of the top 10 lemmas in both domains have a low occurrence, but all are typical terms in their domain.

3.3 Preliminary results

Our aim was to show a range of possibilities for applying this method to information extraction from a corpus. By the moment, we present incomplete data; currently, there exists a disproportion of social and register features regarding the linguistic ones. Our intention is to enlarge the corpus later.

In this paper, the linguistic phenomena taken into account are words and idioms, phonemes and categories.

Below, we present the Dunning test results for the two subcorpora: adult and child language. The first one is made up of 36,905 words and the second, 21,080. Tables 9 and 10 show the distinctive categories in each subcorpus.

Table 9. Distinctive categories in adult language

CATEGORIES	ADULTS (36.905)	CHILDREN (21.080)	DUNNING
MD	1.731	449	264.71
P	6.564	2.739	234.8
INTJ	524	81	162.52
V	6.450	3.167	59.07
AUX	1.278	522	44.88

Table 10. Distinctive categories in child language

CATEGORIES	CHILDREN (21.080)	ADULTS (36.905)	DUNNING
POSS	453	360	127.55
N	3.174	4.419	110.27
ADV	1.861	2.428	96.97
Q	1.739	2.497	42.94
NPR	910	1.242	33.33
C	2.184	3.403	19.83
PREP	1.773	2.786	13.63
ART	1.338	2.068	13.29

These results make us interpret the following:

- In adult language, contrary to what happens in child language, elements like discourse markers (DM) or interjections (INTJ) are highly frequent. Both elements belong to the pragmatic level and require higher linguistic skills.
- While in adult language verbs (V) are the element that guides the speech, children from three to six years old base their speech on nouns (N).

- Possessive pronouns (POSS) are the most distinctive element in child language. According to J. Piaget (1965), until seven years old, child language is characterized by being egocentric, that is, it is a simple accompaniment of the action and the child does not have any other perspective than his/hers. If we have a look back at Table 8, we can see that some of the most common words in child language are “mi”, “yo” or “me”.
- In child language, categories like conjunction (C), preposition (P) and article (ART) are distinctive. In particular, the high occurrence of conjunctions in this subcorpus is caused by the frequent use by children of the copulative conjunction “y”, explained in section 2.1.1.
- Finally, in Table 10 the category proper noun (NPR) appears as distinctive of child language. This is a consequence of the context of situation, that is, a school context where children constantly demand the teacher’s attention, calling his/her name.

Apart from categories, the Dunning test has been also applied to the phonological level. The orthographic transliterations obtained from the audio recordings are automatically transcribed in IPA. In this way, the texts can undergo the same process explained in section 2.1 to extract the phonological information.

Table 11. Distinctive phonemes in adult language

PHONEMES	ADULTS (136.721)	CHILDREN (77.240)	DUNNING
t	6.408	2.949	90.86
b	4.197	1.914	63.6
e	19.938	10.342	58.27
k	6.851	3.352	49.62
s	11.382	5.924	28.72

Table 12. Distinctive phonemes in child language

PHONEMES	CHILDREN (77.240)	ADULTS (136.721)	DUNNING
ʌ	1.024	1.108	128.15
i	6.277	9.635	82.59
p	2.265	3.176	72.57
m	2.928	4.348	55.2
o	8.490	14.247	16.88
u	2.872	4.667	13.39
r	407	571	12.71
g	957	1.461	12.66
x	618	940	8.54

The most interesting thing about the results is the distinctive character of phonemes λ and r in child language. The first one (λ) can be the result of the high frequency of the personal pronoun “yo” in the egocentric language of children. Both phonemes are liquid, curiously the last phonemes that children acquire in the learning process—together with fricatives— due to the difficulty that involves their place of articulation (Anula 1998).

4. Conclusions and future work

In this paper we have presented a spontaneous child language corpus, CHIEDE, made up of 60,000 words, of which a third part correspond to child language. The main contribution of this work is the creation of a linguistic resource that is still in short supply. The research on language acquisition must be based upon the direct observation of reality. With CHIEDE, we provide a wide sample of child language in both, audio and text format. Moreover, texts are enriched by phonological and morphosyntactic annotation, from which information relating to these linguistic levels can be automatically extracted.

Future work will address the following issues:

- Increase in the size of the corpus, not only in number of words, but also in the number of participants and communicative situations.
- Carrying out qualitative researches from the quantitative data. The different phonological and morphosyntactic phenomena can be an object of study for future researches.

We have also proved the significance of the Dunning test as a method for the validation of psycholinguistic hypothesis in spoken language, as well as for determining a register typology. This test correlates linguistic to socio-contextual data applying the Statistics of Surprise. For this task, it is necessary to have an annotated corpus and the use of XML. The preliminary results are promising and had not been shown for Spanish before. However, it is rather premature to extract conclusions and interpretations for these data, as the corpus size is clearly insufficient.

Notes

¹ This research has been supported by the Madrid Regional Government under the contract MAVIR (S-0505/TIC/0267) and the Spanish Government under the project BRAVO-RL (TIN2007-67407-C03-02).

² We have taken into account only verb forms, as, because of their high irregularity, they are one of the most problematic issues in learning Spanish.

References

- Anula, A. 1998. *El abecé de la psicolingüística*. Madrid: Arco-Libros, D.L.
- Biber, D. 1988. *Variation across speech and writing*. Cambridge: CUP.
- Biber, D. 1995. *Dimensions of register variation*. Cambridge: CUP.
- Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finnegan. 1999. *Longman grammar of spoken and written English*. London: Longman.
- Cresti, E., F. Bacelar do Nascimento, A. Moreno, J. Veronis, Ph. Martin, K. Choukri 2002. The C-ORAL-ROM project. New methods for spoken language archives in a multilingual romance corpus. In M. González Rodríguez and C. Paz Suárez Araujo (eds), *Proceedings of LREC 2002*. Paris: ELRA, 2-9.
- Daille, B. 1994. Combined approach for terminology extraction: lexical statistics and linguistic filtering. PhD diss., Paris 7.
- Dunning, T. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19, 1: 61-74.
- Garrote, M. 2008. *CHIEDE. Corpus de habla infantil espontánea del Español*. PhD diss., Universidad Autónoma de Madrid.
- Guirao, J.M., A. Moreno Sandoval, A. González Ledesma, G. De La Madrid and M. Alcántara. 2006. Relating linguistics units to socio-contextual information in a spontaneous speech corpus of Spanish. In A. Wilson, D. Archer and P. Rayson (eds), *Corpus linguistics around the world*. Amsterdam: Rodopi, 101-113.
- Labov, W. 1966. *The social stratification of English in New York City*. Washington: Center for Applied Linguistics.
- Miller, J. and R. Weinert 1999. *Spontaneous spoken language*. Oxford: Clarendon.
- Moreno, A., D. T. Toledano, N. Curto and R. de la Torre. 2006. Inventario de frecuencias fonémicas y silábicas del castellano espontáneo y escrito. In L. Buera, E. Lleida, A. Miguel and A. Ortega (eds), *Actas de las IV Jornadas de Tecnologías del Habla*. Zaragoza: Universidad de Zaragoza, 77-80.
- Moreno, A. 2002. La evolución de los corpus de habla espontánea: la experiencia del LLI-UAM. In A. Rubio Ayuso (ed.), *Actas de las II Jornadas en Tecnologías del Habla*. Granada: Universidad de Granada.

- Moreno, A. and J.M. Guirao 2003. Tagging a spontaneous speech corpus of Spanish. In N. Nicolov, R. Mitkov, G. Angelova and K. Boncheva (eds), *Proceedings of Recent Advances in NLP (RANLP-2003)*. Amsterdam: John Benjamins, 292-296.
- Piaget, J. 1965. *El lenguaje y el pensamiento en el niño*. Buenos Aires: Paidós.
- Uchimoto, K. 2002. Morphological analysis of the spontaneous speech corpus. In Shu-Chuan Tseng (ed.), *Proceedings of Conference of Computational Linguistics (COLING 2002)*. Taipei, Taiwan, 1298-1302.

Appendix

Table 5. Phonemes frequency by age groups.

Phoneme	Age group: 3		Age group: 4		Age group: 5		C-ORAL-ROM (Adults)	
	Absolute	Relative	Absolute	Relative	Absolute	Relative	Phoneme	Frequency Absolute Relative
a	2616	13,29	3190	13,44	4360	13,58	e	188196 15,12
e	2539	12,90	3108	13,09	4215	13,13	a	152664 12,27
o	2300	11,68	2484	10,46	3472	10,81	o	129208 10,38
i	1656	8,41	1922	8,10	2551	7,94	s	100881 8,11
n	1447	7,35	1792	7,55	2550	7,94	i	89799 7,22
s	1444	7,34	1709	7,20	2159	6,72	n	87775 7,05
l	884	4,49	1139	4,80	1553	4,84	r	63702 5,12
r	869	4,41	1130	4,76	1523	4,74	t	56287 4,52
t	793	4,03	1041	4,39	1472	4,58	l	56107 4,51
u	765	3,89	927	3,91	1258	3,92	k	55863 4,49
k	732	3,72	914	3,85	1214	3,78	d	54284 4,36
m	681	3,46	885	3,73	1134	3,53	m	39278 3,15
p	625	3,18	739	3,11	1134	3,53	u	39146 3,14
d	524	2,66	709	2,99	871	2,71	p	34135 2,74
b	477	2,42	644	2,71	777	2,42	b	31126 2,50
ʌ	344	1,75	308	1,30	361	1,12	θ	18940 1,52
g	296	1,50	274	1,15	342	1,06	g	11359 0,91
x	166	0,84	221	0,93	330	1,03	ʎ	10356 0,83
θ	165	0,84	210	0,88	214	0,67	x	7681 0,62
tʃ	125	0,64	106	0,45	200	0,62	f	6217 0,50
r	111	0,56	101	0,43	199	0,62	r	5236 0,42
f	68	0,35	95	0,40	126	0,39	tʃ	3744 0,30
ɲ	58	0,29	89	0,37	98	0,31	ɲ	2427 0,19
Total	19685	100,00	23737	100,00	32113	100,00		1244411 100

FROM TEXT TO LEXICON

THE ANNOTATION OF PRE-TARGET STRUCTURES IN AN ITALIAN LEARNER CORPUS

Giuseppina Turco * Miriam Voghera °

* Max Planck Institute ° University of Salerno

1. Introduction

In the last two decades, studies on Italian as a second language (L2) have massively increased, partly due to the immigration phenomena, involving an ever-increasing number of people (Vedovelli 2002a, 2002b; Valentini 2005). Social and cultural needs, as well as scientific interests, have given rise to an extensive bibliography at a theoretical, descriptive and applied level (De Mauro et. al. 2003; Giacalone Ramat 2003).

Many projects have been devoted to the collection of learner corpora, which provide us with objectives and reliable material to study learners' interlanguages (IL) (Pravec 2002; Granger 2002). The first Italian learner corpus was the Corpus Pavia, a 100-hour corpus of spoken Italian collected at the University of Pavia (Andorno and Bernini 2003). In recent years, other corpora have been collected: LIPS, a 70-hour corpus of L2 spoken Italian consisting in 530, 000 words, collected at the University of Siena Stranieri (Parlaritaliano site); Valico, a corpus of written text of 567,437 tokens, collected at the University of Turin (Barbera et al. 2007); a corpus of L2 spoken Italian comprising approximately 28,000 words, collected at the University of Perugia Stranieri; Cocerit (*Corpus della Certificazione dell'italiano*), a corpus of approximately 11h of spoken interviews, collected at the University of Roma 3 (Ambroso and Bonvino 2008).

Second Language Acquisition (SLA) scholars agree on a general design of 'mise en grammaire' (Giacalone Ramat 2007) of L2 learners, although not all linguistic levels are equally investigated. Studies on morphology and grammatical categories, such as tense, aspect and gender, have traditionally occupied a prominent position, while a stronger interest in syntax and textuality have emerged more recently (Giacalone Ramat 2003, 2007). This also reflects on the annotation systems and tools associated to corpora, which mostly take into account the parts of speech

and the lexicon, and are not expressly syntax-oriented. In fact, the LIPS Corpus has generated a frequency lexicon, while Valico and the Corpus of L2 of the University of Perugia Stranieri can be searched by words and parts of speech (PoS)¹.

In this paper we show the first results of a project on the syntax development observed in learners of L2 Italian. Its main goal was to compare non-native and native structures. In order to produce quantitatively comparable data, we created a Treebank corpus of texts written by students of L2 Italian (Turco 2005), parallel to a Treebank corpus of spoken and written Italian collected at the University of Salerno (Voghera et al. 2004, 2005). Within this context, we adapted the tagging system developed in the Treebank project, AN.ANA.S. (Voghera et al. 2004, 2005) and shaped AN.ANA.S. L2 (Turco 2005, cfr. §5). The use of an L2 annotation system that has the same basic structure of the system we use to annotate native language allowed a straightforward comparison between non-native and native production and the minimum use of *ad hoc* categories for the description of L2 texts.

One of the most challenging tasks in learner corpora annotation is the treatment of structures that deviate from those of native speakers, which we call Pre-target Structures here (cfr. § 3). Usually the deviation from native structures is treated by error tagging systems, which are associated with learner corpora. Although they can vary widely as far as the delicacy of the analysis is concerned, they basically share a two-dimensional architecture: the deviation is assigned to different level of analysis (morphology, grammar, lexis, syntax) and to different categories or types (Grange 2003; Diaz-Negrillo and Fernandez-Dominguez 2006). These types of taggers are not easily compatible with a syntactic annotation, which segment the texts into different units (such as phrases, clauses and sentences) since there is no association between the error tag and linguistic constituent, and therefore it is not immediately possible to go back to the unit to which the deviation applies. In other words, in a Treebank annotation system it would be desirable to track down not only which type of deviation is present in the text, but also to correlate such a deviation to the syntactic unit wherein it occurs. This would permit us retrieve the frequency of Pre-target Structures per linguistic unit and to ascertain in which type of constituents Pre-target Structures are more commonly clustered. To this end, we developed an annotation system, AN.ANA.S. L2, which integrates the tagging of deviation structures and the syntactic analysis, by obligatorily assigning the deviating structure to different levels of syntactic segmentation (cfr. §5).

In the following pages, we present the first results of the application of AN.ANA.S. L2 and focus on two intertwined subjects: 1) the frequency and the type of Pre-target Structures (PtSs) across three L2 Italian proficiency levels (PL), in order to identify the developmental changes from the early stage of L2 proficiency to a more advanced stage; 2) the evaluation of AN.ANA.S. L2 annotation system for the analysis of L2 writing assessment, which we hope will provide fresh empirical evidence for the annotation of Pre-target Structures.

2. Within the perspective of Interlanguage (IL)

The present research has been developed within the perspective of interlanguage (IL, Selinker 1972, 1992) which is a linguistic intermediate system second language learners construct when trying to come to terms with their target language (L2): It is “a separate linguistic system based on the observable output which results from a learner’s attempted production of a target language norm” (1972, 214)². Since the definition was put forth, IL has generally been recognized as sequences of grammars learners develop at different stages in their L2 acquisition process. As a natural language, it constitutes an independent system of rules characterized by its own internal consistency, in continuous development toward the target language (TL). The starting point of such a continuum is represented not only by learner’s L1 but also by the whole of general knowledge that we, as speakers of an L1, already hold. Thus, the focus of investigation is not exclusively based on first language interference phenomena, but also on the learner’s acquisition strategies and his developing grammar as system (Richards J.C. 1971, 1975). In fact, since IL is to be rule-driven and characterised by internal consistency, it can be defined as systematic. Evidence of such a sistematicity might be represented by the regular occurrence of some structural patterns at a given stage of the IL. We can exemplify this in the following case (1) extracted from a descriptive text³:

- (1) Un soggiorno ammobiliato e comodo e non più grande.
 ‘A living room furnished and comfortable and not more big.’

Una cucina pratica e l’appartamento è non più caro.
 ‘A kitchen room functional and the flat is not more expensive.’

più grande [more big]	<i>instead of (?)</i>	troppo grande [too big]
più caro [more expensive]	<i>instead of (?)</i>	troppo caro [too expensive]

As seen above, the learner regularly extends the use of the adverb *più* to contexts where the adverb *troppo* would be appropriate in the TL.

IL is systematic because it is built on regular patterns which can be defined as universal. This is probably because it reflects how cognitive mechanisms control acquisition, irrespective of the personal background of learners, their mother tongue, or the setting in which they learn.

Nonetheless, systematicity should not necessarily imply the idea of a static nature. IL can take different forms and change over time (Braidi 1999), since the permeable and interchangeable nature of learner’s grammar seems to be constantly open to

different influences and types of processing (such as L1 transfer phenomenon). Also, IL grammars are often described as transitional because they shift from one stage to another. They are continuously involved in a process of reconstruction, and, as learners come in contact with new structures and usages, IL is reshaped and reorganized, taking forms different from previous stages⁴. All this gives ILs the quality of being discrete, in the sense that there are discernible differences which specifically characterize one interim stage of an IL from subsequent ones.

The properties of IL suggest a natural order of acquisition in the developmental route of language learning (Bettoni 2001; Ellis 2003). It seems that the acquisition process takes place by developmental stages; that is, the passage from one stage to another is marked by the presence of a new rule⁵. Further, there is an implicit order among developmental stages, in that the acquisition of a given rule seem to entail the acquisition of an earlier rule at a previous stage but not to imply the acquisition of a later one at a later stage (Giacalone Ramat 1993). However, a given developmental stage can be fulfilled even when a learner has not yet totally learnt how to apply a rule which had been acquired earlier in all its possible contexts, or, when the learner does not master all the functions of a form or all the formal expressions of a function. Evidence of a natural order of acquisition derives from many corpus-based studies, which register that in contexts of spontaneous acquisition, the first linguistic structures to be acquired are isolated (chunks of) words or formulas (Ellis 2003). A system of rules building relations between constituents will be developed to some extent only afterwards. This trend seems to be shared by all learners, regardless of mother tongue. Within such a perspective, all utterances which seem to deviate from the target should not necessarily be thought of as determined solely by the process of L1 transfer but rather by the process of learning strategies. Some of these strategies may be common to all learners while giving the IL the property of being a creative system.

This somehow suggests that IL is systematic within the variability. As said earlier, the acquisition proceeds gradually by hypothesis and trials, which can vary to a large extent from the standard. For these reasons, in such a continuum, we may find common patterns of production shareable by all learners, independent from the L1, in addition to a wide range of variability; we may notice a natural order of acquisition concerning some sequences as well as many factors of individual variability, unavoidably due to personal aspects, such as the pace each single learner's development takes to make one step towards the next stage.

To sum up, the production of learners is governed by a coherent linguistic system and submitted to a process of gradual restructuring, which partly progresses in common stages. Systematicity can manifest at different levels: within the IL of one learner and within the ILs of a group of learners sharing a common feature (such as the proficiency level or the L1). In the present investigation, we are interested in the systematic linguistic features relative to learners of L2 Italian, which we shall define as Pre-target Structures (PtS).

3. From Errors to Pre-target Structures

Following Selinker (1972), the notion of IL is at the base of a second language acquisition theory which looks at the learning process as a dynamic process. IL assumes a perspective "which should be held distinct from the 'teaching perspective'" and where "claims about the internal structures and process of the learning organism" (...) "provide the *raison d'être* for viewing the second-language learning from the learning perspective" (1972, 209). This separation was necessary in light of a new theory, which called for a new type of ("psychologically-relevant") data.

Nowadays, learning and teaching perspectives are not so far apart any more, since many teaching approaches initiate from learner production (such as the communicative approaches, Nunan 1991; Brown 1994). However, a common approach to the treatment of structures which diverge from the native ones does not exist. Acquisitional-oriented research⁶ considers deviating structures within the context of IL and consequently as part of a dynamic grammar system. Pedagogical-oriented research (among the others Bitchener 2008; Lee 2008a, 2008b; Truscott and Hsu 2008) and computational automatic treatment of L2 texts primarily described and classified the deviating structures to create for instance, error taggers (Grange 2003; Diaz-Negrillo and Fernandez-Domínguez 2006). This two-fold way of looking at learner productions would come from two different interests: one focused on the learning process, the other on the learning product.

Yet the analysis of the structures deviating from native structures affords an important source of knowledge of the IL development by learners in different stages. Therefore, the analysis of deviating structures must become part of the construction of IL grammars. In this perspective, deviating structures are possible linguistic structures, which reflect approaches to the target language. To this end, our annotation system, AN.ANA.S. L2, focuses on the linguistic structures to which errors apply rather than on the error itself. The premise on which AN.ANA.S. L2 was built was to design the development of syntax proficiency as a progressive acquisition of different constructions or structures, which must map into linguistic units, such phrase, clause. etc. (Ellis 2003). Deviation from native language is thus considered a feature of a different syntactic unit. Units that lack all or some target features must be considered Pre-target structures; thus, we can have Pre-target phrase, clause, etc. Their frequency variation characterises different learning stages.

In such a way, errors are seen in a dynamic perspective, focused on learners' process and its valuation. This would go along the same direction of learner-oriented L2 teaching approaches, focused on social and individual aspects of the learning process, the reasons for its failures and success, and of corpus-based learning and teaching approaches, which have been developed exploiting computer error tagging systems (Kettemann and Marko 2002).

By PtS we mean structures perceived as agrammatical with respect to the target language. We see examples of what we would perceive as improbable structures with respect to the target language produced by beginners ((2)-(4)):

- | | | | |
|-----|--|-------------------|--|
| (2) | Ho nato 5 marzo
'(I) have born 5th March' | <i>instead of</i> | Sono nato il 5 marzo
'(I) was born in the 5th of March'
[I was born in the 5th of March] |
| (3) | Mi ha piaciuto
'To me has liked' | <i>instead of</i> | Mi è piaciuto
'To me is liked'
[I liked it] |
| (4) | Me va matto
'Me goes mad' | <i>instead of</i> | Vado matto
'(I) go mad'
[I go mad] |

Even expressions that match with the grammar of the target language but do not fit in well with the context and/or do not convey the intended meaning are considered PtS. We found such cases due to the inappropriate contextualization of linguistics features (communicative intentions, situation, etc.), like the examples below ((5)-(6)) produced by beginners:

- | | | | |
|-----|---|-----------------------|--|
| (5) | Sono molto buono
'(I) am very good' | <i>instead of (?)</i> | Sto molto bene
'(I) am very well'
[I am very well] |
| (6) | Vorrei alti gradi all'Università
'(I)'d like high degrees at the University' | | |
| | alti gradi
[high degrees] | <i>instead of (?)</i> | voti alti
[high marks] |

As is well known, L2 learners also have to deal with the existence of many geographical and socio-cultural varieties. As far as Italian is concerned, we must take into account the deep diatopic differences that can interfere with the learning process (Dal Negro and Molinelli 2002; Lepschy 2005). In general, we do not believe that a sole grammar of the target language exists; it is unrealistic to believe in a unique use of the language. Therefore, when a learner uses a geographically marked structure or lexical item we do not tag it as PtS, since we prefer to keep away from a tempting prescriptive approach. What follows is an example of target language variation, produced by an advanced learner:

- (7) Ci stavano delle differenze *instead of* C'erano delle differenze
 'There stayed some differences' 'There were some differences'

The use of the verb form *stare* (*to stay*), stating the same existential function as the verb form *essere* (*to be*), is typically used within the Italian variety spoken in the South of Italy.

Therefore, for a PtS identification it seems reasonable to adopt different criteria. It is important to obtain different perspectives beforehand. We have looked at clauses in context (not in isolation), assessed language features within a complete discourse, and accounted for their full significance. In doing so, it has been a straightforward process to move beyond a narrow focus on grammatical features and to integrate the evaluation of extra-clausal domains (that is, units like the sentence, the paragraph and the text).

4. The Corpus and the Participants

The texts of our corpus were collected at the Greenwich University of London and elicited by three groups of learners – beginner (BEG), intermediate (INT), advanced (ADV) – whose level respectively corresponds to the Basic User level (A2 profile), the Independent User (B2 profile) and the ADV to Proficient User (C1 profile) of the Common European Framework⁷.

In the following table, we show the total number of words and texts produced by a total of 41 participants.

Table 1. Number of texts and words of Italian L2 Corpus

CORPUS	BEG	INT	ADV	Tot
N TXT	64	47	41	152
N WRD	6014	6054	6055	18123

The subjects were undergraduate students who had different mother tongues and used English as a second language in their daily life. Their Italian course was an optional language learning course which was part of their combined honour degree. Therefore, the language course was timetabled for two hours per week and taught by a native-speaking Italian teacher. Other than those two hours, students were also expected to spend four hours a week working on their portfolio (see Common European Framework) and doing some extra independent learning activity in the language centre. Texts were administered at the beginning of the semester (which covers a time period of 4 months). Thus the time limit to complete the portfolio was extended along the duration of the course. The weighting of this form of assignment was 25% and incorporates two other forms of in-class assignments administered

during the course. As an incentive, students were told at the beginning of the course that their portfolio would be scored and credited 25% if completed to the best of their ability. They were also told that some of their writing samples would be chosen and employed as the object of investigation for a research project. Also, students' background information was collected during the course.

The type of task prompt designed for the portfolios consisted of a bare prompt; namely, a simple explanation of how to perform the task-like activities to be submitted. As previewed by the European Framework, texts belong to the three macro-categories of text types: narrative, descriptive and argumentative. These are types of texts with well-defined and identifiable communicative functions. They may represent, then, a good way to examine learners' proficiency development and to reconstruct their IL at a given stage of their acquisition.

5. The annotation and the Tag-set

AN.ANA.S. L2, developed within a Treebank project at the University of Salerno, is a syntax-oriented annotation system which allows the organization of linguistic units of a text within a hierarchical structure of syntactic levels: 1) text; 2) paragraph (or turn, if spoken dialogic texts); c) sentence; d) clause; e) noun phrase (NP), verb phrase (VP), prepositional phrase (PP); predicative noun phrase (PredP).

The analysis of the texts in examination is based on a prior segmentation and creation of trees (Figure 1) in a linear sequence. Tree diagrams allow the identification of the *type* of relation between phrases and the *levels* of syntactic relations. Each *entity* (<text>, <paragraph>, <sentence> etc.) stands for the linguistic unit represented by a node of the tree. Each entity is defined by a set of *attributes* that assume a *value* selected within a pre-determined definite set. Tree diagrams are created through the annotation in XML format. While using XML, each entity is assigned a specific label (or *tag*). Figure 1 below illustrates two shallow tree diagrams (Cutugno and Voghera 2004).

The given annotation procedure essentially entails endowing a text with 'tags' (descriptive labels) according to XML standards. A hierarchical role is assigned to each single part of the text in relation to all possible syntactic levels and, within each level, associating tags which are representative of each syntactic constituent. Each tag is defined by a certain number of attributes included in it which permits the retrieval of a series of information when required⁸.

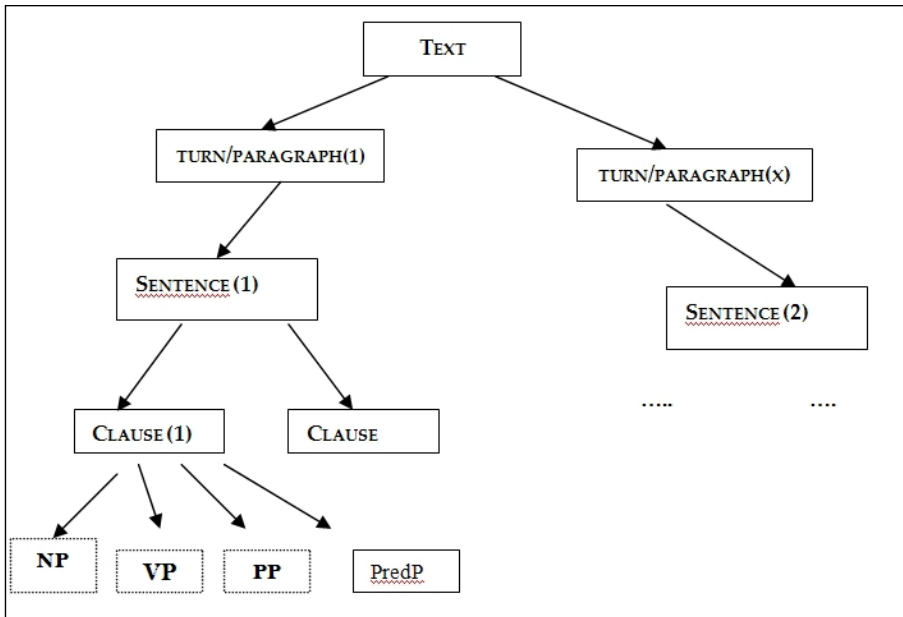


Figure 1. AN.ANA.S. level of analysis representation

PtSs are assigned to each above-mentioned level. This permits the retrieval of PtSs per level of textual encoding (<text> and <paragraph>) and per level of syntactic encoding (<sentence>, <clause>, <phrase>). Lexical deviations affecting head phrases are marked at phrase level.

After a snap identification of PtS tagged with a more comprehensive label, the second step of our analysis entailed developing a more detailed ontology of pre-target structures, thereby obtaining a more precise identification of them, thanks to the support of the annotation system, AN.ANA.S. L2 and the editor XGate (Parlaritaliano site). Such a typology represents a starting point for PtS coding, which could then evolve and improve in light of experience gained during a deeper inspection of the corpus. As said before, previous research lacked an existing ontology of such structures that could be easily translated into a Treebank annotation scheme: not many studies in L2 Italian writing assessment provide a detailed *modus operandi* on how to classify, evaluate and annotate what is being referred in the literature as errors or mistakes. Subsequently, we aimed to enrich the editor with specific tags for each level of analysis. Appendix 1 presents the list of deviations we marked and the tag set organized per syntactic hierarchical level. In fact, the annotation process goes top-down, that is, from the text up to the phrase. At each level, there are the more typical deviations clustered for that level. Some of them correspond to cases reported in prior literature on writing assessment in English as EFL (Kroll 1990).

When analyzing PtS, confusion may arise between its description (what is it?) and the explanation for its cause (what is it caused by?). The two levels must be distinguished. From a preliminary superficial description, we have noticed that, by and large, PtS may result in⁹: a) *omission* of elements, that is, when learners' production leave out a linguistic element in comparison with target production being compared as shown in the example (8), where we can see that the head of NP is missing (\emptyset); b) in *extra* elements, when learner's production contains one more element with respect to the compared native production (for instance (9) wherein we see more verbal construction, either *ho* or *vorrei*); c) and in *deviating* use of structures, like the example (10), where learners' production presents a linguistic element which is different from the one present in the comparable native production. These are characteristics applying to all linguistic levels discussed above. We may find omission of linking elements like subordinator in-between clauses, at a sentence level, as well as prepositions in-between phrases, at a clause level. The following examples demonstrate the cases discussed:

- | | | | |
|------|---|-------------------|---|
| (8) | La \emptyset più importante è che...
'The \emptyset most important is that...' | <i>instead of</i> | La cosa più importante è che...
'The thing most important is that...'
[The most important thing is that...] |
| (9) | Ho vorrei una grande casa
'(I) have (I) would like a big house' | <i>instead of</i> | Vorrei una grande casa
'(I) would like a big house'
[I would like a big house] |
| (10) | Sono andato a Grecia
'(I) have gone to Greece' | <i>instead of</i> | Sono andato in Grecia
'(I) have gone in Greece'
[I went to Greece] |

As is well known, tagging is not a straightforward process. A deviating structure or element can have different domains, i.e. it may involve different levels of codification. A deviation can produce PtS at a different level. In example (11) the head of the VP, presumably the verb *essere* ('to be'), is missing and the modal *dovere* ('must') does not present the agreement with the subject *giardino* ('garden'), but is used in the infinitive form.

- (11) Il giardino dovere \emptyset grande
'The garden must-INF be- \emptyset big'
- instead of (?)*
- Il giardino deve essere grande
'The garden must be big'

Clearly, in this case, the deviation affects both the phrase level and the clause level, so we must recognize the presence of a pre-target phrase and of a pre-target clause.

After tagging the whole corpus, it is possible to extrapolate information about the syntactic constituents examined through the implementation of a semi-structured database¹⁰. Analytic results are extracted through *queries* of the database, supported by XPath queries languages (Bird and Liberman 2001, Scott and Bird 2002) which, after performing a query, run (*top-down* and *bottom-up*) all the way through the syntactic tree codified in XML, and then select and group the elements involved in the query.

Then, as seen in Figure 1, the rating scale of PtSs regards different levels of analysis, which may be subcategorized into two superordinate categories: the inter-clausal subscale (sentence, paragraph, text) and the intra-clausal subscale (phrase, clause). At the textual level, we have considered pre-target structures which do not adhere to principles of cohesion and coherence. We have followed some basic criteria: the overall argument is well-conveyed; the description or narration of an event is clear since temporal and causal relationships are logically expressed; ideas are all relevant to the topic; the division of text into paragraphs is justifiable in terms of content relevance so that ideas are well-related to one another. Referencing and the use of linking words are appropriate so that the passage from one unit to another does not seem imperceptible. What follows is the opening paragraph of a letter (12), produced by beginners, where a PtS at a level of text is shown, since what should be at the end of a letter, turns out to be the *incipit* of the letter.

- (12) Tanti saluti da Pheonix.
[Many regards from Phoenix.]

Visitavamo Ø nostri amici per una settimana in albergo...
'(We) visited the-Ø our friends for a week in hotel...'
[We visited our friend for a week at the hotel...]

At a paragraph level, we considered PtSs run-on sentences and cases of comma splices that do not occur where they should in order to operate a clear division of ideas, like shown in the example below (13):

- (13) In la casa le camere dovere Ø grande e splendidamente decorate con i bagni.
'In the house the rooms must BE-Ø big and splendidly decorated with the bathrooms.'

Che chiamo un bellissimo e speciale casa abitare in.
'That (I) call a beautiful-MASC/SING and special house-FEM/SING to live in.'

instead of (?)

(...) Quello che io chiamo una casa bellissima e speciale in cui abitare.
 ‘That which I call a house beautiful and special in which to live.’
 [What I consider to be a beautiful and special house to live in.]

At a sentence level PtSs present deviating syntactic constructions (mirroring L1 transfer); deviating use of the *consecutio temporum*; omission, extra or deviating use of coordinating/subordinating conjunctions; omission or deviating use of referencing. The following example (14) by a BEG is a clear case of deviating usage of referencing, because we find the clitic feminine *la* instead of the masculine form *lo* to refer back to the action of *swimming*:

- (14) Non potrei nuotare e la mia mamma ha deserta insegnarla
 ‘Not (I) could swim and the my mother has deserted-FEM teach-her-PRON-COREF *of to swim*’

instead of

(...) e la mia mamma ha deserta insegnarmelo¹¹
 ‘(...) and the my mother has deserted-FEM teach-to me-it-PRON-COREF *of to swim*’

PtSs at clause level present deviating relationships between phrases: improbable phrase order; the deviating use of arguments; their omission or extra presence; the deviating agreement between phrase constituents (like subject-verb); the deviating use of prepositions; their omission or extra presence. In the following example (15) both lack of agreement between subject and verb and a deviating phrase order produced by BEG occur.

- (15) il giorni più belle è mia sorella giornata del matrimonio
 ‘the-MAS/SING most beautiful-FEM/PLUR day-MAS/PLUR is my sister day of the wedding’

instead of

il giorno più bello è...
 ‘the-MAS/SING most beautiful-MAS/SING day-MAS/SING... is...’
 the most beautiful day is...’

PtSs at phrase level present deviating relations between the head of the phrase and other elements: deviating gender/number agreement, deviating in-phrase word order

6. Frequency and distribution of PtS

In this section we present the first quantitative data of PtSs per linguistic level (LL) and proficiency level (PL). Even if we do not claim to give statistical significance to our results, we think that a quantitative evaluation of PtSs distribution can be a measure of the IL development and can reveal interesting trends in learning patterns. In fact, we think that quantitative analysis can shed light on many aspects of ILs (Granger 2003; Ellis 2003).

The number of PtS found in the corpus (Table 2) shows that the writing learning process proceeds regularly from the lower to the upper levels, since PtSs percentage decreases by nearly 50% as long as it goes from the lower PL (BEG) to the upper PL (ADV).

Table 2. PtS (absolute and relative) numbers per PL

PL	BEG	INT	ADV	TOT
N_PtS	608	301	139	1048
% PtS	58,02 %	28,7 2%	13,26 %	100%

Figure 2 shows the distribution of PtSs in relation to each LL. We can notice that learners seem to have better control at the highest levels of textual and syntactic planning, since these levels score the lowest number of PtSs. Instead, the most critical domain seems to be at the phrase level, wherein more than 40% out of all PtSs occur.

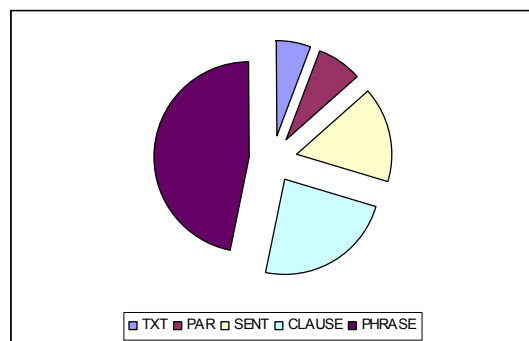


Figure 2. PtS Percentage per PL and LL calculated on the whole corpus

We find that such a frequency pattern, affecting the highest levels of the writing processing, is common to all three PLs, although PtSs distribution diverges from one PL to the other, as shown in Figure 3.

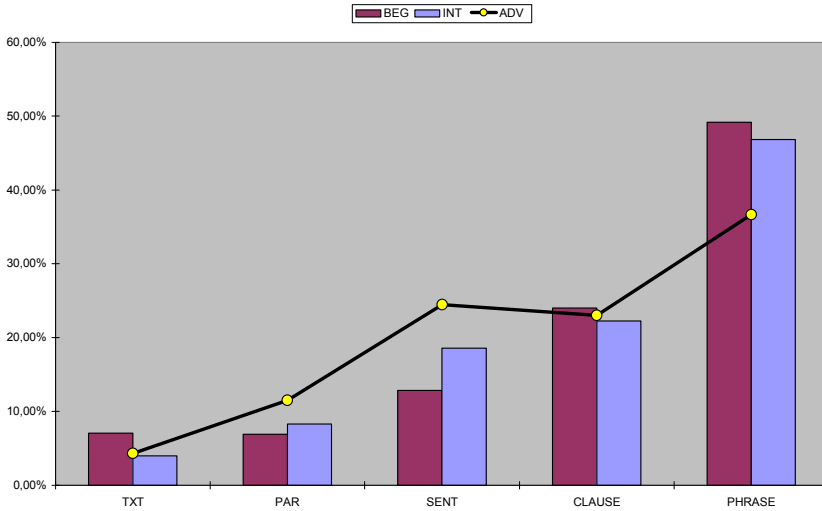


Figure 3. Distribution of PtSs in relation to LLs for each PL

The three PLs show a similar pattern of frequencies: most of the PtSs are distributed in the three syntactic levels (phrase, clause, sentence), while they decrease rapidly at a textual level (paragraph and text). Results coming from our data suggest that learners across the three PLs are more proficient when building up a textual frame and ensuring a basic coherence within the text, while they are less proficient in the shape of the single syntactic chunks. Considering the type of texts (narrative, descriptive, argumentative) and learners (undergraduate students), textual planning may represent the level learners can deal with better than the other levels. This may be due to a cultural-educational factor; in general, a good master of textual level patterns can be traced to due to learners' education and schooling backgrounds¹²: "L2 writing relates closely to native-language literacy and particular instructional contexts" (Myles 2002, 8). Also, one more reason may be due to the fact that all learners share a common textual-literacy tradition, since most of them come from European countries and have been educated in very similar contexts. Finally, a factor which can contribute to good textual planning is the relatively short size of texts included in the corpus (Turco in preparation).

The development of learner's syntax seems to be more internally complex and not at all linear. In such a respect, two points have clearly emerged from our first findings: firstly, the frequency of PtSs show that the phrase level is a critical point across all the PLs; secondly, syntactic proficiency does not seem to start from the bottom levels (phrase) and go up to the higher levels (sentence level). Data clearly show that the master of clause internal rules does not necessarily imply the master phrase internal rule. In fact, the number of pre-target structures at different levels do

not vary proportionally: Pre-target structures are always more numerous at phrase level, regardless of the PL.

The highest number of Pre-target phrases derives from the fact that the phrase is the syntactic domain of many grammatical choices about categories values in Italian. In NP learners must express the gender, number and definiteness of the head noun, eventually controlling the agreement of determinants and adjectives (Chini 2003). It is worthwhile to recall that Italian flexional morphology is highly redundant and that both determinants and adjectives have different endings according to the gender and the number of the head nouns. The complexity of multiple agreements in Italian unavoidably leads to the production of a higher number of deviations: in fact, most of the Pre-target NPs present lack of agreement, as shown in the example (21), where the learner miss the agreement between the head noun *spiaggia* (FEM) and the target of agreement, the adjective *bello* (MAS):

- (21) *la spiaggia più bello del mondo*
 ‘the-FEM/SING most beautiful-MAS/SING beach-FEM/SING of the world’

instead of

la spiaggia più bella del mondo
 ‘the-FEM/SING most beautiful-FEM/SING beach-FEM/SING of the world’
 [the most wonderful beach in the world]

If we compare the syntactic roles within the Pre-target NP, we notice that Pre-target NP are more frequent as object-role than as subject-role or as circumstantial: half of the Pre-target NPs are indeed in the syntactic role of object. This probably depends on the assumption that subject position is generally more preserved in ILs because of its salience for the learners: syntactically speaking, the subject represents the element which pilots the agreement inside the clause; semantically, subjects represent the actant which has a major control on the action; finally, pragmatically, subjects express the topic of the sentence and/or of discourse (Keenan 1976; Giacalone Ramat 2003).

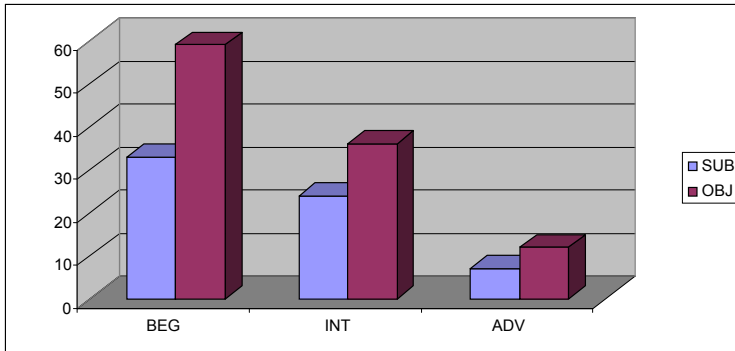


Figure 4. PtS NPs syntactic roles

Pre-target VPs mainly show two features: lack of number agreement between the auxiliary and the past participle in the compound tenses and deviating mood choice. The lack of number on past participles is probably due to different causes. Firstly, the agreement between the subject and the past participle can vary in Italian depending on different factors: argumental structure of verb, auxiliaries and *essere* ('to be') or *avere* ('to have') (Salvi 1991). The following examples show very well that learners do not master the rule of participle agreement, according to which the past participle must agree with the subject when the verb is intransitive and has the auxiliary *essere* like in (22), while the subject and the past participle must not agree with transitive and intransitive (unergative) verbs with the auxiliary *avere*, like in (23):

- (22) ...ci siamo svegliato molto presto
 '(we) have-PLUR woken-SING up very early'

instead of

...ci siamo svegliati
 '... we have-PLUR woken-PLUR up'
 [... we have woken up]

- (23) Abbiamo parlati tutto il giorno e la notte
 '(We) have-PLUR spoken-PLUR all day and night long'

instead of

Abbiamo parlato
 '(We) have-PLUR spoken-SING'
 [We have spoken]

Moreover, when the past participle occurs as part of a compound form, in Italian the controller of number agreement (the subject) can be either absent, because of its pro-drop property (see (22) and (23)) or relatively distant. In this latter case, distance seems to play an important role in number value assignment, since it has been noted that the more the distance from the controller, the harder the task for learners (Giacalone Ramat 2003).

Secondly, past participle occurs frequently in native language as an invariable form to express perfective actions, such as in *finito* ('finished', 'gone'), *capito* ('understood'). These uses probably determine the uses by L2 learners of invariable past participle as an expression of perfect in basic IL (Banfi and Bernini 2003). The difficulty of marking gender and number on past participles manifests also within the NPs where the past participle has the function of noun modifier (Chini 2003). In the acquisitional sequence of gender and number assignment within the NPs, the past participle comes last after pronouns, articles and adjectives.

The deviating mood choice within the VPs confirms acquisitional sequences presented in other studies (Banfi and Bernini 2003). It is well known that mood is the last notional category, after aspect and tense, manifested by the verb in Italian ILs. This explains why in our data inappropriateness in mood choice persists even at the advanced PL.

- (24) Che peccato tu non potrebbe venire!
'What pity you not would come-CONDITIONAL'

instead of

...tu non possa venire
'...you not come-SUBJUNCTIVE'
[... you could not come]

- (25) Vorrei uno studio in modo da potrei studiare
'(I) would like a studio so to-PREP (I) could study-COND'

instead of

... in modo da poter studiare
'...so to-PREP to study-INF'
[... so to be able to study]

As far as verb endings are concerned, relatively few Pre-target VPs present deviations: the Pre-target VP which present deviating verb endings are 23% of the PtSs in the beginner PL and 15% of the PtSs in the advanced PL. According to Giacalone Ramat (2000; 2003) this is due the high degree of diagrammatic transparency in many Italian regular verb forms which can be easily segmented and

processed by learners. Such a property somehow eases the meaning-to-form matching, which is essential for the acquisition of the morphological markers.

While Pre-target phrases decrease regularly, passing from the beginners to the advanced, Pre-target clauses remain constant across the three levels and Pre-target sentences increase (Table 3).

Table 3. Percentages of Pt sentences and clauses

PL	Pt SENTENCES	Pt CLAUSES
BEG	12,83%	24,01%
INT	18,60%	22,26%
ADV	24,46%	23,02%

The most frequent deviation at clause level is the deviating subject-verb agreement. This feature is significantly present not only at the initial acquisitional stages, but also at the advanced level, although it may differ from one stage to another. In example (26), there is agreement deviation at the NP level, because of the lack of agreement between the masculine controller noun (*giorni*) and the adjective target which is feminine (*belle*), and at clause level, because the verb is singular rather than plural. These cases are typical of the beginner stages in which the learners apparently do not master agreement rules at any syntactic level.

- (26) *il giorni più belle ...*
 ‘the-MAS/SING day-MAS/PLUR most beautiful-FEM/PLUR ...’
 (BEG)

instead of

- il giorno più bello...*
 ‘the-MAS/SING day-MAS/SING most beautiful-MAS/SING ...’
 [the most beautiful day...]

Examples (27) and (28) show two different cases. In (27) the learner perfectly masters the agreement within the NP, but does not master the difference between the NP and the VP; in (28) the opposite happens.

- (27) *...ci sarebbero un bagno romano ...*
 ‘...there would be-PLUR a-SING/MAS bathroom-SING/MAS roman-SING/MAS’
 (BEG)

instead of

...ci sarebbe un bagno romano ...
 ‘...there would be-SING a-SING/MAS bathroom-SING/MAS roman-SING/MAS’
 [...there would be a Roman style bathroom]

- (28) Le cose che mi fanno piangere sono specialmente le cose triste
 ‘The things which to me make cry are especially the-PLUR things-PLUR sad-SING’

instead of

...le cose tristi
 ‘the-PLUR things-PLUR sad-PLUR’
 [...sad things]

Finally, we can have cases of deviating subject-verb agreement, such as in example (29) produced by an intermediate learner, showing only the agreement within the NP [*tutta la classe*] but not between the NP (singular) and the verb (plural), probably because the collective noun *classe* is not randomly interpreted as plural at a semantic level. In this case, the agreement rule is semantically driven.

- (29) Tutta la classe hanno mostrato i loro piatti
 ‘Whole the class-SING have-PLUR shown the their dishes’
 (INT)

instead of

Tutta la classe ha mostrato
 ‘Whole the class-SING has shown the their dishes’
 [The whole class has shown their dishes]

From these examples, we can suppose that subject-verb agreement and agreement among the constituents of the NP can be independently mastered. Our data bring evidence to the major relevance of subject-verb agreement for the learners, since they seem to use it before and better than the agreement between the elements inside the phrase. This depends on the fact that subject-verb agreement is necessary to guarantee cohesion and coherence to the discourse, while the disagreement among the elements of the phrase normally does not have consequences as far as the transmission of meaning is concerned.

The last syntactic constituent we have considered is the sentence. As we have seen in Table 3, at the advanced level Pre-target sentences occur twice as much at the beginner level. This is because the degree of syntactic complexity may differ to a

large extent according to the PLs. At the beginner level, sentences with two or more clauses are nearly 50% of the total, the pluriclausal sentences become 70% at the intermediate level and 75% at the advanced level (Turco in preparation). The most common deviation is lack of conjunction or other linking elements, such as relative pronouns and prepositions. The examples (30)-(32)

- (30) Il giorno più belle della mia vita è il giorno Ø ho incontrato il mio ragazzo
 ‘The most day-MAS/SING beautiful-FEM/PLUR of the my life is the day Ø (I) have met the my boyfriend’

instead of

Il giorno più bello della mia vita è il giorno in cui ho incontrato il mio ragazzo
 ‘The most day-MAS/SING beautiful-MAS/SING of the my life is the day in which (I) have met the my boyfriend’
 [The most wonderful day in my life is when I met my boyfriend]

- (31) ...anzi, le strade erano così cattivo Ø è scoppiato un pneumatico.
 ‘...rather, the streets-FEM/PLUR were so bad-MAS/SING Ø has blown out a pneumatic’

instead of

...anzi, le strade erano così cattive che è scoppiato un pneumatico.
 ‘...rather, the streets-FEM/PLUR were so bad-FEM/PLUR that Ø has blown out a pneumatic’
 [...rather streets were so bad that a pneumatic has blown out]

- (32) anche dare alla gioventù la opportunità Ø imparare un mestiere
 ‘...also to give to the youthness the chance Ø learn-INF a job’

instead of

...la opportunità di imparare un mestiere
 ‘...the chance of-SUB PREP to learn a job’
 [...the chance to learn a job]

It is interesting to note that as the degree of sentence complexity grows deeper, as at the upper levels of proficiency, the presence of co-reference deviations increases, while they are basically absent at the beginning level within the sentence.

Finally, as far as the lexicon is concerned, we have taken into account deviating lexical choices which are shown not to be pragmatically appropriate to the communicative situation. The Table 4 shows that the INT level scores the highest frequency of deviating lexical choice.

Table 4. Frequency of Lexical Deviation per Proficiency Level

PL	Lexical Deviation
BEG	5,26%
INT	6,64%
ADV	5,04%

It appears that there is no immediate correlation between lexical proficiency and PLs. Further investigations on the frequency of lexical types and tokens would be useful, so as to gain deeper insight into the highest number of lexical deviation in the INT level with respect to the BEG level. However, we think that this is in partly due to the type of tasks: written communicative situations are more guided in the case of BEG, whereas they are freer in the case of the INT level. From a first qualitative analysis, we can say that INT learners' lexicon seems to be richer than BEG. Somehow, they are more creative in the use of their lexical repertoire (partly required by freer task prompts). This may lead to a higher probability of inappropriateness in their productions (ex. (33)-(34)). Conversely, in the case of BEG, specific writing contexts and guided task prompts represent a constraint to creativity and lead learners to adhere more to the usage of formulas or more common words in their productions.

(33) Ho picchiato *una macchina parcheggiata...*

'(I) have beaten a parked car down'

instead of (?)

Ho tamponato una macchina

'(I) hit a car...'

[I hit a car]

(34) ritirano *che la pena di morte è una condanna appropriata*

'(they) withdraw that the penalty of death is a sentence appropriate'

instead of (?)

ritengono *che la pena di morte è una condanna appropriata*

'(they) believe that the penalty of death is a sentence appropriate'

[They believe that the penalty of death is an appropriate sentence]

7. Final remarks

We would like to conclude with some considerations about the efficiency of the annotation system discussed here and the syntactic development our data show across the three PLs.

Our annotation system, AN.ANA.S. L2, tries to reconcile the syntactic analysis of learner productions with a systematic tagging of deviating structures. The choice was to map deviating structures onto textual and syntactic units, to retrieve not only the single deviation, but, in a manner of speaking, the domain of the deviation. It is clear that a deviating structure can affect constituents of different size and level, giving rise to different PtS.

Since our focus was on PtS more than on deviating aspects, AN.ANA.S. L2 uses a relatively light tag set, if compared with established error tagging systems (Granger 2002, 2003, Diaz-Negrillo and Fernandez-Dominguez 2006). However, the annotation was sufficient, in our opinion, to enlighten the critical syntactic points in the learning process, and in turn to obtain the valuable property of manageability.

The description of the syntax of IL by the frequency of different PtS across the three PLs reveals how learners build their syntactic proficiency through the construction of different syntactic chunks. Additionally, looking at the internal constituency of syntactic units permits us to see how the same linguistic aspect can be differently mastered at different levels. For instance, our data seem to show that subject-verb agreement is used before the agreement among phrase elements is used.

Quantitative results suggest learners use a type of writing proficiency which goes from the highest to the lowest levels of syntactic planning: in our corpus most of the PtSs occur at the phrase level, while pre-target clauses and sentences are half as numerous. Moreover, pre-target phrases remain numerous in all three PLs: at the advanced level 36% of PtSs are phrases. This implies that learners can produce well-formed sentences, whose phrases lack many native features.

This seems to indicate that there is not a unique direction of syntactic development (Pienemann 1986, 1998); i.e., learners do not proceed from micro-units to macro-units or vice versa, but they go back and forward continuously. The development of syntax does not consist of an additive process that merely goes from simple to complex units, but of a constant process of redefinition of the relationship between forms and functions. This means that syntactic development may not always be so linear and implicational (lexical entries > lexical categories > phrasal > inter-phrasal level, etc.), but learners work simultaneously at different levels. Learners seem to start building up chunks, no matter how accurately, and then to come back to the internal features of the chunks. Syntax develops through an inter-relational process; that is, it does not start from building relations 'within' one single level but 'between' different syntactic levels. In this way learners progressively acquire the capacity to use the same linguistic resource (for instance agreement) at different levels.

The learning process does not have a unique direction of development and does not merely consist of increasing the number of structures. ILs corresponding to different stages of learning are not characterized only by the presence of new structures, but also by the capacity of using old structures in new contexts. ILs are open and transitional systems, which necessarily change as long as the learners act linguistically in increasingly complex situations. This can imply the acquisition of new structures and the mapping of new forms with known functions and known functions with new forms (Tomlin 1990; Cristofaro and Ramat 1999; Braidì 1999; Bettoni 2001; Ellis 2003).

Finally, we have considered if and to what extent our results depend on the specificity of writing. There is no doubt that students' responses are conditioned by cultural-educational reasons, since their writing proficiency and their text construction process is determined by past writing experience-writing knowledge and practice that students receive through instruction, as found for other languages in many studies (Berman 1994; Carson 1992; Cumming 1989; Silva 2005). This skill can come from both general L1 literacy training provided in school (which in our case is a common literacy because learners all come from mature literate contexts) and from L2 writing educational instructions. This can explain why we found very few PtSs at textual level: it is presumable that university students can cope with short descriptive and narrative texts. Also, we should take into account the idea that planning at a textual level calls for abilities which are partially different from micro-level forms of syntactic planning: textual processing lies very much on the building of semantic relations, such as the coherence or the overall organization of the text (i.e. introduction, body and conclusion), while syntactic organization implies the ability of building relationships of structural nature.

However, we think that the great difference in the proficiency of text mechanisms and lower syntactic ability may lead to the hypothesis that while writing, learners build up the textual frame in order to offset the deficits that eventually affect the lowest levels. In that sense, an ill-structured phrase can receive significance from a well-formed textual structure, as the example (35) shows:

- (35) Abbiamo parlati tutto il giorno e la notte, e da allora, noi amore l'altro e ci sposiamo.
 '(We) have spoken-PLUR all the day and the night, and since then, we love-NOUN the other and each other (we) marry'

instead of

Abbiamo parlato tutto il giorno e tutta la notte e da allora ci amiamo e ci siamo sposati
 '(We) have spoken-PLUR all the day and all the night, and since then, (we) each other and we get married'

[We have spoken all the day and night long, and since then we love and we get married.]

On the contrary, a well-formed phrase misses significance if placed in an ill-formed textual structure (36).

Incipit of a letter

(36) Tanti saluti da Pheonix.

[Many regards from Phoenix.]

Visitavamo Ø nostri amici per una settimana in albergo...

‘(We) visited the-Ø our friends for a week in hotel...’

[We visited our friend for a week at the hotel...]

In a written text where no disruptive intervention by other speakers can occur, like in spoken dialogues, building up a coherent net on the textual level signifies setting the first stone of successful communication.

Notes

¹ It is well known that the PoS tag-sets can vary greatly depending on the different theoretical and descriptive choices made by researchers. In any case, an intrinsic feature of any PoS tagging is the word-by-word annotation.

² The definition of IL has changed over time, since preliminary research on it have appeared. Several attributes have been used: referred as ‘approximative systems’ in Nemser 1971; ‘idiosyncratic dialect’ in Corder 1971, Selinker 1972; Richards 1972, Schumann 1974; Selinker 1994.

³ We usually compare the pre-target structures with the native structures; sometimes the corresponding native structure is just an hypothesis. In these cases, we use a question mark to precede the reconstruction.

⁴ Naturally, learners can even remain at an initial state of proficiency; i.e., IL can fossilize at a determinate stage (Selinker 1972).

⁵ For further details on L2 Italian see Giacalone Ramat (1993, 2003).

⁶ Mainly developed for the analysis of spontaneous language acquisition by immigrants.

⁷ A2: Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.”; B2: “...Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. Can

interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.”; C1: “...Can understand a wide range of demanding, longer texts, and recognise implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors and cohesive devices.” (see Common European Framework).

⁸ Since the beginning of our project we developed different versions of AN.ANA.S. In Appendix 2 we describe the basic attributes for each level; the complete version is described in Voghera *et al* 2004, 2005; Voghera and Turco 2006).

⁹ This matches error description due to phenomena like omission, overgeneralization, substitution, etc., we find in the literature (Corder 1972, Richards, Ellis 1997). However, our classification is not based on the source of the deviations, since we do not take into account the L1 as a variable nor its possible influences on learners' L2 productions.

¹⁰ For illustration of the database scheme derived from the annotation process, Cutugno and D'Anna in press.

¹¹ We do not give a translation for this example, since we are not sure about the intended meaning for 'ha deserta': *ha desiderato?* ('has wished') *ha mancato?* ('has missed').

¹² This concerns the rhetorical patterns we have investigated, like the overall organization of ideas and the use of connectives. It should be said that many studies on Contrastive Rhetoric (Kaplan 1966) in L2 writing show that students at this level may transfer macro-level rhetoric patterns from L1 to L2 such as paragraph organization, linear organization structure (Connor 1987), etc.

References

- Ambroso, S. and E. Bonvino. 2008. Livelli diversi di competenza nella gestione dell'italiano L2. Ipotesi dall'analisi di un corpus. *Testi e linguaggi* 2: 39-67.
- Andorno, C. and G. Bernini. 2003. Premesse teoriche e metodologiche. In A. Giacalone Ramat (ed.), *Verso l'italiano*. Roma: Carocci, 27-36.
- Banfi, E. and G. Bernini. 2003. Il verbo. In A. Giacalone Ramat (ed.), *Verso l'italiano*. Roma: Carocci, 70-115.
- Barbera, M., E. Corvino and C. Onesti. 2007. *Corpora e linguistica in rete*. Perugia: Guerra edizioni.
- Berman, R. 1994. Learner's transfer of writing skills between languages. *TESL Canada Journal* 12, 1, 29-46.
- Bettoni, C. 2001. *Imparare un'altra lingua*. Roma and Bari: Laterza.
- Bird, S. and M. Liberman. 2001. A formal framework for linguistic annotation. *Speech Communication* 33: 23-60.

- Braidi, S. 1999. *The Acquisition of Second-Language Syntax*. London: Arnold.
- Brown, H.D. 1994. *Principles of Language Learning and Teaching*. Englewood Cliffs, NJ: Prentice Hall Regents.
- Carson, J.G. 1992. Becoming biliterate: First language influences. *Journal of Second Language Writing* 1, 37–60.
- Chini, M. and S. Ferraris. 2003. Morfologia del nome. In A. Giacalone Ramat (ed.), *Verso l'italiano*. Roma: Carocci, 37-69.
- Common European Framework. www.coe.int/t/dg4/linguistic/CADRE_EN.asp
- Connor, U. 1987. Argumentative patterns in student essays: Cross-cultural differences. In U. Connor and R. Kaplan (eds), *Writing across languages: analysis of L2 text*. Reading MA: Addison-Wesley, 57-71.
- Corder, S.P. 1971. Idiosyncratic Dialects and Error Analysis. *IRAL* 9, 147-160.
- Corpus Pavia. www.unipv.it/wwwling
- Corpus Valico. www.corpora.unito.it/valico
- Corpus L2 spoken Italian. www.elearning.unistrapg.it/osservatorio/corpus/frames-cqp.html
- Cristofaro, S. and P. Ramat (eds). 1999. *Introduzione alla tipologia linguistica*. Roma: Carocci.
- Cumming, A. 1989. Writing expertise and second language proficiency. *Language Learning*, 39, 81-141.
- Cutugno, F. and M. Voghera. 2004. AN.ANA.S.: Analisi sintattica e annotazione XML a contatto. In F. Albano Leoni, F. Cutugno, M. Pettorino and R. Savy (eds), *Il parlato italiano. Atti del Convegno Nazionale (CD-ROM)*. Napoli: D'Auria Editore, M03.
- Cutugno, F. and L. D'Anna. In press. Limiti e complessità del recupero delle informazioni da tree-bank sintattiche. *Atti del convegno della SLI (Vercelli, settembre, 2006)*. Roma: Bulzoni.
- Dal Negro, S. and P. Molinelli. 2002. *Comunicare nella torre di Babele. Repertori plurilingui in Italia oggi*. Roma: Carocci.
- De Mauro, T., M. Vedovelli and M. Barni and L. Miraglia. 2002. *Italiano 2000. Indagini sulle motivazioni e usi pubblici dell'italiano diffuso fra stranieri*. Roma: Ministero degli Affari Esteri.
- Diaz-Negrillo, A. and J. Fernandez-Domínguez. 2006. Error Tagging Systems for Learner Corpora. *RESLA* 19, 83-102.
- Ellis, N. 2003. Constructions, chunking, and connectionism: the emergence of second language structure". In C.J. Doughty and M.H. Long (eds), *The Handbook of Second Language Acquisition*. Malden, MA: Blackwell Publishing, 63-103.
- Ellis, R. 1997. *Second language acquisition*. Oxford: Oxford University Press.
- Giacalone Ramat, A. 1999. Italiano di stranieri. In A.A. Sobrero (ed.), *Introduzione all'italiano contemporaneo. La variazione e gli usi*. Bari. Laterza: 341-410.
- Giacalone Ramat, A. 2003. Il quadro teorico. In A. Giacalone Ramat (ed.), *Verso l'italiano*. Roma: Carocci, 17-26.
- Giacalone Ramat, A. 2007. On the road. Verso l'acquisizione dell'italiano lingua seconda. In M. Chini, P. Desideri, M. Elena Favilla and G. Pallotti (eds), *Atti del 6° Congresso Internazionale dell'Associazione Italiana di Linguistica Applicata*. Perugia: Guerra edizioni, 13-41.

- Giacalone Ramat, A. (ed.). 2003. *Verso l'italiano*. Percorsi e strategie di acquisizione. Roma: Carocci.
- Grange, Sylviane. 2003. Error-tagged Learner Corpora and CALL: A Promising Synergy. *CALICO Journal* 20: 465-480.
- Granger, S. 2002. A Bird's-eye of learner corpus research. In S. Granger, J. Hung, S. Petch-Tyson (eds), *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam: Benjamins, 4-33.
- Kaplan, R.B. 1966. Cultural thought patterns in intercultural education. *Language Learning*, 16, 1-20.
- Keenan, E.L. 1976. Towards a universal definition of subject. In C.N. Li (ed.), *Subject and topic*. New York: Academic Press, 303-333.
- Kroll, B. 1990. *Second language writing: research and insights for the classroom*. Cambridge: Cambridge University Press.
- Kettemann, B. and G. Marko (eds). 2002. *Teaching and learning by doing corpus analysis: Proceedings of the Fourth International Conference on Teaching and Language Corpora*. Amsterdam and New York: Rodopi.
- Lepschy A.L. and A.R. Tamponi (eds). 2005. *Prospettive sull'italiano come Lingua Straniera*. Perugia: Guerra.
- Lepschy, G. 2005. Lo standard. In A.L. Lepschy and A.R. Tamponi (eds), *Prospettive dell'italiano come lingua straniera*. Perugia: Guerra, 15-21.
- Myles, J. 2002. Second language writing and research: the writing process and error analysis in student texts". *TESL-EJ* 8: 1-18.
- Nemser, W. 1971. Approximative Systems of Foreign Language Learners. *IRAL* 9: 115- 123.
- Nunan, D. 1991. Communicative tasks and the language curriculum. *TESOL Quarterly* 25, 2: 279-295.
- Parlaritaliano. <http://www.parlaritaliano.it>
- Pienemann, M. 1986. L'effetto dell'insegnamento sugli orientamenti degli apprendenti nell'acquisizione di L2. In A. Giacalone Ramat (ed.), *L'apprendimento spontaneo di una seconda lingua*. Bologna: Il Mulino, 307-326.
- Pienemann, M. 1998. *Language processing and second language development: processability theory*. Amsterdam: Benjamins.
- Polio, G.Ch. 1997. Measures of linguistics accuracy in second language writing research. *Language Learning* 47: 101-143.
- Pravec, N.A. 2002. Survey of learner corpora. *Icame Journal* 26: 81-114.
- Richards, J.C. 1971. Error analysis and second language strategies. *Language Sciences* 1: 12-22.
- Richards, J.C. 1974. Error analysis and second language strategies. In J.H. Schumann and N. Stenson Rowley (eds), *New frontiers in second language learning*. Mass.: Newbury House, 32-53.
- Salvi, G. 1991. L'accordo. In L. Renzi and G. Salvi (eds), *Grande grammatica italiana di consultazione*. Bologna: Il Mulino, 227-244.
- Schumann, J.H. 1974. The implications of interlanguage, pidginization and creolization for the study of adult second language acquisition. *TESOL Quarterly* 8: 145-152.

- Scott, C. and S. Bird. 2002. An integrated framework for treebanks and multilayer annotations. In M. González Rodríguez and C. Paz Suárez Araujo (eds), *Proceedings of LREC 2002*. Paris: ELRA, 1670-1677.
- Selinker, L. D.-E. Kim and Sh. Bandi-Rao. 2004. Linguistic structure with processing in second language research: is «unified theory» possible? *Second Language Research* 20, 77-94.
- Selinker, L. 1972 Interlanguage. *IRAL* 10: 209-231.
- Selinker, L. 1992. *Rediscovering interlanguage*. New York: Longman.
- Silva, T. 2006. Second language writing. In K. Brown (ed.), *Encyclopedia of Language and Linguistics*. Oxford: Elsevier, 191-205.
- Tomlin, R.S. (ed.). 1987. *Coherence and grounding in Discourse*. Amsterdam: Benjamins.
- Turco, G. 2005. The Intraclausal syntax in texts written by L2 learners of Italian. Degree Dissertation. University of Salerno.
- Turco, G. In preparation. Complessità sintattica nell'italiano scritto L2.
- Valentini, A. 2005. Lingue e interlingue dell'immigrazione in Italia. *Linguistica e filologia* 21: 185-208. Bergamo, Università degli Studi di Bergamo.
- Vedovelli, M. 2002a. *L'italiano degli stranieri. Storia, attualità e prospettive*. Roma: Carocci.
- Vedovelli, M. 2002b. *Guida all'italiano per stranieri. La prospettiva del quadro comune europeo*. Roma: Carocci.
- Voghera, M., G. Basile, D. Cerbasi, G. Fiorentino. 2004. La sintassi della clausola nel dialogo. In F. Albano Leoni, F. Cutugno, M. Pettorino and R. Savy (eds), *Il parlato italiano. Atti del Convegno Nazionale* (CD-ROM). Napoli: D'Auria Editore, B17.
- Voghera, M., G. Basile, F. Cutugno and G. Fiorentino. 2005. Sintassi in AN.ANA.S. In F. Albano Leoni and R. Giordano (eds), *Italiano parlato. Analisi di un dialogo*. Liguori: Napoli, 189-211.
- Voghera M. and G. Turco. 2006. *Manuale utente AN.ANA.S. 3*, in www.parlaritaliano.it
- XML. <http://www.w3c.org/XML>
- XPath 1.0. <http://www.w3.org/TR/xpath>

Appendix 1: Tag-set for PtSs annotation

LEVELS	ENTITY	TAG-SET	TAG DESCRIPTION
LVL0	TEXT : in-between paragraphs	DPC	<i>Deviating coherence</i>
		DPF	<i>Deviating linking device</i>
		MLD	<i>Missing Linking device</i>
		DTT	<i>Deviating text type</i>
		DREF	<i>Deviating reference</i>
		MREF	<i>Missing reference</i>
LVL1	PARAGRAPH : in-between sentences	DSC 1	<i>Deviating syntactic constructions</i>
		DPNCT	<i>Deviating punctuation</i>
LVL2	SENTENCE : in-between clauses	DSC 2	<i>Deviating syntactic constructions</i>
		DCT	<i>Deviating consecution temporum</i>
		DX2	<i>Deviating conjunction</i>
		MX2	<i>Missing conjunction</i>
		EX2	<i>Extra conjunction</i>
		DCOREF	<i>Deviating coreference</i>
		MCOREF	<i>Missing coreference</i>
		ECOREF	<i>Extra coreference</i>
LVL3	CLAUSE: in-between phrases	MPH	<i>Missing head element</i>
		EPH	<i>Extra head element</i>
		DO3	<i>Deviating phrase order</i>
		DA3	<i>Deviating phrase agreement</i>
		MX3	<i>Deviating preposition</i>
		DX3	<i>Missing preposition</i>
		EX3	<i>Extra preposition</i>

LEVELS	ENTITY	TAG-SET	TAG DESCRIPTION
LVL4	PHRASE : in-between words	WCS	<i>Word class shift</i>
		DO4	<i>Deviating word order</i>
		DA4	<i>Deviating in-phrase agreement</i>
		EXE4	<i>Extra element in-between phrase</i>
		Dmd	<i>Deviating modifier</i>
		Mmd	<i>Missing Modifier</i>
		Emd	<i>Extra Modifier</i>
		Ddt	<i>Deviating determinant</i>
		Mdt	<i>Missing determinant</i>
		Edt	<i>Extra determinant</i>
	VERB DOMAIN	DVv	<i>Deviating Verb voice</i>
		DVp	<i>Deviating Verb particle</i>
		MVp	<i>Missing Verb particle</i>
		EVp	<i>Extra Verb particle</i>
		DVa	<i>Deviating Verb auxiliary</i>
		MVa	<i>Missing Verb auxiliary</i>
		EVa	<i>Extra Verb auxiliary</i>
		DVe	<i>Deviating Verb ending</i>
		MVe	<i>Missing Verb ending</i>
		EVe	<i>Extra Verb ending</i>
		DVt	<i>Deviating Verb tense</i>
		DVm	<i>Deviating Verb mood</i>
LVL5	LEXICON	DLC	<i>Deviating Lexical Choice</i>

Appendix 2

We show here a reduced version of the DTD (Determined Text Definition) of AN.ANA.S. L2. We report here only the main attributes for each element and, in brackets, their possible value. The more updated complete version is illustrated in Voghera and Turco 2006.

```

<!ELEMENT ANANAS_SCRITTO_L2 (text)+>
<!ELEMENT text (paragraph+ )>
<!ATTLIST text
text_identification
type of text (narrative, descriptive, explicative, argumentative, other)
production (monologue or dialogue)
well-formed (true/false)
>
<!ELEMENT paragraph (sentence+)>
<!ATTLIST paragraph
paragraph_identification
well-formed (true/false)
>
<!ELEMENT turn (sentence+)>
<!ATTLIST turn
turn_identification
turn completion (true/false)
well-formed (true/false)
>
<!ELEMENT sentence
<!ATTLIST sentence
splitted sentence (start or middle or end of the sentence)
uniclausal (true/false)
number of clauses (true/false)
well-formed (true/false)
>
<!ELEMENT clause
<!ATTLIST clause
type (main or dependent or nominal)
number of phrases
link (subordinate conjunction or subordinate preposition or null or relative)
argumentative (true/false)
well-formed (true/false)
>

```

```

<!ELEMENT NP
<!ATTLIST NP
lexeme
subject (true/false)
object (true/false)
presence of determinant (true/false)
presence of modifier (true/false)
position (pre-VP or post-VP or circumstantial)
...
well-formed (true/false)
>
<!ELEMENT VP
<!ATTLIST VP
lexeme
copula verb (true/false)
number of arguments (0 | 1 | 2 | 3)
saturation (true/false)
person (0 | 1 | 2 | 3 | 4 | 5 | 6)
position (pre-subj or post-subj ...)
...
well-formed (true/false)
>
<!ELEMENT PP
<!ATTLIST PP
preposition
lexeme
      position (pre-modifier or post-modifier or circumstantial or isolated PP)
modified phrase (NP or VP or PP or PredP)
...
well-formed (true/false)

>
<!ELEMENT PredP
<!ATTLIST PredP
lexeme
part of speech (noun, adjective or pronoun or other)
position (pre-copula verb or post-copula verb ... )
.....
well-formed (true/false)

```


PRE-PROCESSING NORMALIZATION PROCEDURES FOR NEWSGROUP CORPORA

Manuel Barbera, Simona Colombo
Università di Torino

1. Introduction

When dealing with corpora collection and computational analysis it becomes necessary to speak about “well formed” texts, i.e. ready to be analysed and processed by an automatic computer procedure.

Collecting written texts such as articles, books, poetry, newspapers, advertisements gives the chance to get, almost without particular operation, “a collection of written or spoken material in machine-readable form, assembled for the purpose of studying linguistic structures, frequencies, etc.”

C'è sempre stata una linguistica basata sullo spoglio di materiali linguistici, anche molto copiosi, ma con linguistica dei corpora, traduzione dell'inglese corpus linguistics, si intende oggi quella branca della linguistica che si occupa di elaborare i dati provenienti da larghi insiemi di testi immagazzinati su supporti leggibili dal computer. È dunque una linguistica dei corpora elettronici [...] (Marello 1996, 167)

The original texts are not always written in a form that is clear and suitable to be analysed by a computer, but this is really a matter that becomes extremely relevant with the increase of the corpus' size.

The first step for a computational approach to corpora is to know the variety of text and “non text” recorded in the collection.

It is a matter of fact that if you collect a large amount of texts that need to be verified and checked, you cannot worry about every single text, instead you just have to take care of the kind of study you have to carry out and define the best algorithm and script to make a corpus out of these texts.

If your aim is to put together a big quantity of text, it becomes actually difficult to understand if there is coherence in these texts and if this material is correct and well done. The availability of such a large number of “uncommon” texts made it possible to build a corpus that has twofold characteristics: on the one hand this kind

of registry gives us the possibility to have a rich quantity of authentic text, but on the other hand it contains a set of texts that we can define “non standard”.

To obtain a corpus you have to use a set of operations with algorithms that have to be revised though, if applied on a kind of material that generates a set problems. We are going to describe the set of operations to be done on a text to translate it in a corpus, than we will explain the troubles and the workaround used to obtain a large corpus of “non standard” texts.

2. Pre processing Standard Operation

To define a text as a corpus, also in machine-readable form, it is mandatory to apply on it the tokenization and elementary markcupping.

The tagging operation otherwise is not necessary, as we can see in the case raw corpora «in corpus-driven linguistics you do not use pre-tagged text, but you process the raw text directly and then the patterns of this uncontaminated text are able to be observed» (Sinclair 2000, 36).

2.1 Markup specification

The markcupping is the upper level of information makes it possible to underline a set of text attributes. We can define different levels of corpus annotation.

For example we can highlight emphasis forms, usually marked with typographic devices such as bold or underlined, or editing properties, such as page number, paragraph notation (in the meanwhile the punctuation is isolated at the tokenization level and not in the not in the markcupping).

Table 1. Markup example of Corpus Taurinense

versione markuppata e tokenizzata (testo CT)	
@@BrunettoLatini@@Tesoretto@@@Did	per guerra d' i vicini) ,
%001	esso Comune saggio
\$0175\$ SV	mi fece suo messaggio
[...]	a ÷ll' alto re di Spagna ,
Lo Tesoro conenza .	ch' or è re de la Magna
A ÷l tempo che Fiorenza	e la corona atende ,
froria , e fece frutto ,	se Dio no• lli ÷l contende :
si ch' ell' era de ÷l tutto	ché già sotto la luna
la donna di Toscana	non si truova persona
(ancora che lontana	che , per gentil legnaggio
ne fosse l' una parte ,	né per altro barnaggio ,
\$0180\$ rimossa in altra parte ,	tanto degno ne fosse
quella d' l ghibellini ,	com' esto re Nanfosse .
markup: @autore @@titolo	@@@genere
%capitob \$pagina	&v verso

Table 2. Weakly and strongly markup of Newsgroup Corpus (NUNC)

<pre> <HEAD> <doc-id> <id#>XXXXnnnnnnnnnnnnnnnn</id#> <mess-ID>___</mess-ID> <mess-ref>___</mess-ref> <charset>ansi;unicode</charset> <lingua>___</lingua> <aut_NA>nick,address[nnnnn@___]</aut_NA> <fornitore>bmanuel.org</fornitore> <titolo>subject</titolo> <data>(aaaa,mm;0;?,gg;0;?),0;?</data> <ora>hh:mm:ss</ora> <luogo>?</luogo> </doc-id> <set-id> <corpus>___</corpus> <fonte>NG</fonte> <f_nome>nomeNewsgroup</f_nome> <f_ed>usenet</ed> <gruppo_num>1;2;...;nMess×Thread</gruppo_num> <gruppo_nome>nomeThread</gruppo_nome> </set-id> <autore>...</autore> <testo> <tipo_forma>c-lib_var;c-lib_descr;c-lib_narr;c-lib_reg;c-lib_arg;c-art; tes;dial;ques;es-trad;dett;rias;post;email;lett; mgraf;art;rec;rom;nov;poem</tipo_forma> <tipo_stile>saggL;saggS;giorn;man;amm;legisl;acc;?</tipo_stile> <tipo_fine>divulg;spec;artist;intratt;inform;regol;celeb;?</tipo_fine> <topics>...</topics> <keyw>(____,____,____,____,____);?</keyw> <qualita>origCE</qualita> <pat>----</pat> </testo> <testo-incl_1>ripeti_testo_o_canc</testo-incl_1> <testo-incl_N>ripeti_testo_o_canc</testo-incl_N> <ref> <links>http,...</links> </ref> </HEAD> </pre>
--

The markup is a piece of information that is typical of the analysed text but not inserted in it. Considering the digital form of a text, we have to define the markup level of it to store the text.

To isolate the markup level usually a set of “graphical” forms are used, the most common are the HTML (Hyper Text Mark-Up Language) classes to identify additional information on the text.

The next level of annotation that gives more information about a text is the weakly embedded markup (Buzzetti 2002). It is then possible to mark assets of information that we can define “meta” information because they are not strictly connected to the text form but to the author, title, chapters, paragraphs, pages, lines.

This meta level of annotation is filled also with a number of strongly embedded markapped attributes setting another basket of information such as embedding of a text, text genre, poetry etc.

3. Tokenization

The process of breaking a text up into its constituent tokens is known as tokenization. Usually we use tokenization to discover on the left and on the right of a token the blank characters, isolating the atomic units useful for the automatic processing.

These tokens often do not match with the typographic word so it is evident the difference between token and word.

Table 3. Tokenization example 1 from NUNC

```
NOTE: You may want to retain every name and address sent to you,
either on a computer or hard copy and keep the notes people send
you.
This VERIFIES that you are truly providing a service. (Also, it
might
be a good idea to wrap the $1 bill in dark paper to reduce the
risk of
mail theft). So, as each post is downloaded and the directions
carefully followed, all members will be reimbursed for their
participation as a List Developer with one dollar each. Your name
will move up the list geometrically so that when your name reaches
the
#1 position you will be receiving thousands of dollars in CASH!!!
What
an opportunity for only $6.00 ( $1.00 for each of the first six
people
listed above) Send it now, add your own name to the list and
you're in
business!!!

*****DIRECTIONS FOR HOW TO POST TO NEWS GROUPS!!!!*****

STEP ONE: You do not need to re-type this entire letter to do your
own
posting. Simply put your cursor at the beginning of this letter
and
drag your
cursor to the bottom of this document, and select 'copy' from the
edit
menu. This
will copy the entire letter into the computer's! memory.
```

Tokenization is a set of tools used to isolate each token as a significant part of the text:

The isolation of word-like units from a text is called tokenization. (Grefenstette - Tapanainen 1994, 79)

token means the individual appearance of a word in a certain position in a text. For example, one can consider the wordform dogs as an instance of the word dog. And the wordform dogs that appears in, say, line 13 of page 143 as a specific token. (Grefenstette 1999, 117; cfr. anche Mikheev 2003)

Table 4. Tokenization example 2 from NUNC

```

Newsgroup:it.hobby.cucina
Subject: Re: Presentazione.
From: "gennarino"
Date: Wed, 18 Sep 2002 20:59:49 GMT
Message-ID: <9Psi9.12175$Av4.239760@twister2.libero.it>
References:<amaou2$tqk$2@lacerta.tiscalinet.it>

"Micky"<
> La Molla che al ha spinto a farmi viva è stata l'ultima ricetta
che vi ho
> copiato:
> - I FICHI CARAMELLATI DI GENNARINO -, una meraviglia!
> Non ho parole per definirli e sono già destinati a diventare il
regalo di
>Natale 2002 per tutti i1 parenti ed amici,

Ovviamente, sono lo che scrivo... sotto falso nick! ;-)
(grazie! :)))
> e le castagne al liquore dell'anno scorso.

Quale ricetta seguisti, quella che postai lo? Qualunque fCosse, e'
il
momento giusto per riproporla... non ti pare?! :-)) Magar, con le
tue Impressioni...

> Unico quesito, dureranno fino a Natale?

Spero ben di si' anche perche' senno' resto *fregata* anch'io...
(Prof, lei che ne dice?! :-))
(Rispondirispondirispondi... :)

```

4. Newsgroup non standard features

The linguistics resources available on the net are marked with some specific troubles linked to the informal style of the text and the arbitrary use of the linguistic rules.

The NUNC (Newsgroup UseNet Corpora), corpora are a set of corpora based on newsgroup texts. This kind of example is really good to explain the peculiarity of

problems and characteristics typical of “non standard” text.

The linguistics approach of the Newsgroups highlights some relevant “noise” that has to be considered and managed to build a corpus with these kind of texts.

We can shortly list the most relevant ones:

- Acronyms and abbreviations
- Several kind of spelling
- Spamming
- Repeated text
- Out of topics
- Quoting
- Non standard encoding
- Formatting mistakes
- Non textual attachments
- Html including
- Emoticons
- Web art

We are working on corpora build in Italian, Spanish, English, French and German of million of words, so it is necessary for us to implement a set of automatic tools to process the text, as the large amount of material makes it impossible to do it manually.

We can see the newsgroups registry as a non standard text. In the corpora pre processing there are a lot of studies and procedures to approach this kind of text, with the intent to mark the non standard part of the text to avoid the crash of the automatic procedure. The tools used to implement the NUNC - the IMS (Institut für Maschinelle Sprachverarbeitung of Stuttgart) tools for corpus building and query - present some features to give the possibility to isolate some part of the text that will be ignored by the CQP encoding, avoiding therefore problems and crashing.

Some important research, such as CleanEval, shared task and competitive evaluation on the topic of cleaning arbitrary web pages, to use web data as a corpus, for linguistic and language technology research and development. This project wants to detect in a web page (considering the web as a corpus resources) the “dirty” and non textual part, such as boilerplater, structural annotation, code, to isolate the “text” using only it to build the corpus.

On the contrary in the NUNC it is fundamental to detect and mark a lot of “non standard” parts or sequences in order to work and manage the corpus itself.

The fundamental steps for a text is the line numbering, then the markuppung both editing specification and meta information and the tokenization.

The line number script for this kind of text has to consider the peculiarity of the text, in which there are the quoting line (the line of the mail to which the writes is answering), the text line and the empty line.

To preserve the information associated with the writer's choice to use an empty line or to write between quoting lines we have adopted a script that mark all these different kinds of line and trace this solutions. Starting from a text such as the table 5 shows, it is possible to obtain the Table 6 markup text.

Table 5. Plain NUNC text from es.ciencia.enologia

```

Newsgroups: es.ciencia.enologia
Subject: Re: trasiego
From: Juan Ledesma <jledesmaQUITAESTO@entelchile.net>
Date: Mon, 16 Dec 2002 18:19:58 -0400
Message-ID: <3DFES18E.6090603@entelchile.net>
References: <3DFAD1B2.7040101@uva.es>

joscarr wrote:

> Hola amantes del vino
> Tengo que hacer el trasiego de unos 25 cántaros de vino que se dignó
> darme mi pequeña viña. El asunto está en que me gustaría saber cómo
> debo lavar la cuba y con qué. He oído que una vez lavada hay que
> quemar azufre dentro. Decídme si es así o no. En caso negativo, ¿qué
> hay que hacer?
> Gracias y saludos
>

Depende del material, pero generalmente se utiliza un detergente
alcalino (a base de soda caustica) y un enjuague ácido (como ácido
citríco o ácido peracético), luego un enjuague y listo. Ahora si quieres
desinfectarla el ácido peracético es una buena alternativa. El quemar
azufre libera anhídrido sulfuroso que podría ayudarte a desinfectarlas,
pero de todas maneras tendrías que enjuagarlas antes de agregar el vino,
sino este podría quedar con exceso de SO2. El vapor también es muy útil.

Saludos

```

Table 6. markup NUNC text of table 5 example

```

<head> <doc-id>
  <idN>44</idN>
  <mess-ID><3DFES18E.6090603@entelchile.net></mess-ID>
  <mess-Ref><3DFAD1B2.7040101@uva.es></mess-Ref>
  <Charset>ansi</Charset>
  <lingua>spagnolo</lingua>
  <aut_NA>Juan Ledesma ,<ADDRESS@entelchile.net></aut_NA>
  <fornitore>bmanuel.org</fornitore>
  <titolo>Re: trasiego</titolo>
  <data>2002,12,16</data>
  <ora>18:19:58</ora>
  <luogo>?</luogo>
</doc-id> <set-id>
  <corpus>NUNC-ES Gneric</corpus>
  <fonte>NG</fonte>
  <f_nome>es.ciencia.enologia</f_nome>
  <f_ed>usenet</f_ed>
  <gruppo_num></gruppo_num>
  <gruppo_nome></gruppo_nome>
</set-id> <testo>
  <testoForma>post</testoForma>
  <pat>TQTQT</pat>
</testo> </head> <body>
<tit> Re : trasiego </tit>
<eLn><eLn/>
<p1> joscarr wrote : </p1>
<eLn><eLn/> <qLn ind=1>
Hola amantes del vino
</qLn> <qLn ind=1>
Tengo que hacer el trasiego de unos 25 cántaros de vino que se dignó
</qLn> <qLn ind=1>

```

5. Formatting

The collected text are recorder using some particular softwares that match together all the messages of a thread and download all the structured text in a text format.

Making this operation it is possible that the original format of the text gets lost, so at the beginning of the encoding work of a corpus we try to rebuild the original form of a message, if is possible to infer it from the structure.

So we have developed a set of scripts to mark a single message recording is hieratical information regarding the other messages. We have adopted the <set-id> tag to mark the information about the group name and the number of the message inside the whole thread

Table 7. Formatting markup from NUNC

```

<set-id>
<corpus>NUNC IT</corpus>
<fonte>NG</fonte>
<f_nome>it.discussioni.ristoranti,it.hobby.scuba</f_nome>
<f_ed>usenet</f_ed>
<greuppo_num>1;1</gruppo_num>
<gruppo_nome>Re:Cinqueterre</gruppo_nome>
</set-id>

```

In the meantime we have tried to rebuild the original formatting of the text repairing the break lines due to editor translation.

6. Filtering

6.1 Spamming Problem

The term ‘spam’ as it is used to denote mass unsolicited mailings or netnews postings is derived from a Monty Python sketch set in a movie/tv studio cafeteria. During that sketch, the word ‘spam’ takes over each item offered on the menu until the entire dialogue consists of nothing but ‘spam spam spam spam spam spam and spam.’ This so closely resembles what happens when mass unsolicited mail and posts take over mailing lists and netnews groups that the term has been pushed into common usage in the Internet community.

When unsolicited mail is sent to a mailing list and/or news group it frequently generates more hate mail to the list or group or apparent sender by people who do not realize the true source of the message. (Hambridge and Lunde 1999).

With spam we indicate the same message sent many times in different groups, out of topics within the newsgroup subject, usually advertisement or sellers.

Even if the mail server has an anti-spam algorithm, it is really difficult and expensive the filtering of this kind of messages.

In the corpus building it is relevant to isolate and mark this kind of messages mainly for these reasons:

- The text spam is “dirty” in the context of the thread, so in the corpus linguistic approach it is important to filter this “false” textual information
- The spam messages are usually full of attachment in form of images, html, executable programs, they are messages with particular kind of text that create big troubles to the encoding software.

We have performed a set of script that put in acts a set of strategy that linked together lead to clear the corpus from a big quantity of spam.

Subject Filtering: spamming messages present the same subject across the different newsgroups, so we check the subject of all the messages and store them in a database of subjects, then we count each subject and cut all the messages that have a subject that occur more than the frequency. This value is parametric and is influenced by the corpus typology and dimension.

Table 8. Subject Filtering example from NUNC

```

=====
Newsgroups: es.charla.gastronomia
Subject: Meet someone right now! - free phone sex dating sex:
From: ijntrt@dffdqiuhkoerw.com
Date: 24 Jun 2003 21:00:33 GMT
Message-ID: <3ef5bbf1$0$60397$c3e8da3@news.astraweb.com)

Live Phone Dating -

Call Now: 1-800-418-CHIC

Omgvfniesygrut

POST
From: wbbjsg@dffdqiuhkoerw.com Newsgroups: humanityquest.addict.ion
;Subject: () () () () () FKEE PHONESEX - girls call this sex:

Live Phone Dating -

Call Now: 1-800-418-CHIC

tbgodiwolpnkbtffowfi

```

Post length: the spam message with announcements or advertisement are usually quite shorts, so we have filtered the messages that appear to be shorter than a fixed number of lines. With this kind of solution we lose a set of messages that are short

but not spam, we have considered that the textual information brought by a short message is less relevant than the problems brought about by a spam text.

Table 9. Message length filtering example from NUNC

```

Newsgroups: es.charla.gastronomia
Subject: Re: Papas vitelotte ;)
From: "Afrodita" <afrodita@allandalus.com>
Date: Thu, 29 Jan 2004 19:29:07 +0100
Message-ID: <bubjdi$g2upm$1@ID-189574.news.uni-berlin.de>
References: bvbqr5\$gmb67\$1@ID-189574.news.uni-berlin.de

http://www.chez.com/mairiedeboree/caviar2.html

Son preciosas =D

Afrodita o Galamera en la red

```

Message ID Filtering: spamming messages if sent all together from an automatic sender have the same message id across the newsgroup. Sometimes the spamming algorithm is so sophisticated that the subject is changed but not the identification number of the message. So we use a script based on the same logical structure of the subject filtering.

Cross posting: if in the destination groups of a message there are too many newsgroups (we have adopted a configurable over limit of 15) the message could be a cross posting spam message, i.e. a message sent together to many different newsgroup.

Table 10. Cross Posting

```

Newsgroups: free.uk.politics.parties.green, free.uk.politics.parties.lib-dea,
free.uk.politics.parties.new-labour, free.uk.scanner.radio,
free.uk.Scotland.dating-singles, free.uk.Scotland.football-hearts, free.uk.Scotland.football-bibs, free.uk.Scotland.gays, free.uk, ...
Subject: This is as azasisg way to sake xosey, and finally one that isn't a con!
From: "mrs J price" <3p018b11748blueyonder.co.uk>
Date: Sat, 07 Aug 2004 16:19:15 GMT
Message-id: <D29Sc.9520582$.1e1s98fel.news.blueyosder.co.uk>

```

6.2 “Dirty” characters

In the messages there could be a lot of reasons that introduce “dirty” characters. It is important to find them and mark them as “extra testo” to avoid the crashing of the encoding program. These characters have been detected following different approaches.

First of all the different kinds of editor give the possibility to save the texts in format different from the ASCII standard. So the re-encoding of the different ASCII format, considering unicode different translation clear the text and let the possibility to recognize different type of characters.

Table 11. Unicode translation error

We will pay you =A3100

The newsgroup posts often have some non-textual attachments, such as images, html, executable programs, and we have designed perl script to isolate and mark these pieces of post.

We have adopted the criteria of marking this “dirty” text and not to delete it because we wish to preserve the original form of the text, marking the parts that could create trouble to the encoding algorithm but leaving then inside the text in their original form and position.

Table 12. Non textual attachment

<pre>R01GODlh<ú^7APcAAP///+rp@puSp@G2rDO3úUc62n53a?JMdbSvvVtXh2xre8b?lx8cú4yLprOy 21<»r12Ku25ux315xKp^K2pp4Wy460vm22qftvLÁ2176+283N59fR34u1sçKSoNvb7c703L29 yqO3rtTO4crKlNvb5erq90/v+O7u99PT23bG2«P36vLy993Y3Pv7/vb2+fn5**/v8Kqr0oKKu)ib 55SV@92K28vX2p6ür7S1vqqtV52frOP18CQv«qur203p7pan30zu830PzK8T6?1/xo6V@h9p2C52 3£WC31aS402xr4uw6LXK8i21cüXaa55q97PH2/YirlrbI»51VTh3032cvX5?v9/>/w8Qr8€0€h€2Wk</pre>

6.3 Repeated text

The newsgroup text format itself is based on the quoting phenomenon that creates the repetition of the text. If all the repeated text is used in the quoted part of a message it will be simple to isolate this text and not to consider it in the frequency count.

But the variety of writers and approaches to this kind of communication gives the possibility to refer part of messages yet written without using the quoting strategy. So it becomes really relevant to mark the repeated text in order to avoid “false” counting information.

The detection of this text via usual script algorithm based on regular expression has two main problems, the first one is the slowness of the procedure, the second one is the “theoretical definition” of the length of the text that is marked as repeated. On this definition is based the regular expression criteria for the algorithm. It is really different to find a sequence on n-word, a sequence of lines, and a sequence of paragraph.

Therefore we have started adopting a workaround to be sure that this phenomenon is under control even if we have “lost” some part of “good” text.

All the messages are recorded in order implement a script that looks into a thread for the longest message, using only one message for thread, including the quoting part.

In this way we try to preserve the most quantity of text, preserving also the textual information given from question and answer mechanism.

Implementing the algorithm some particular attention has been used to filter the thread that are recorded on the newsgroup server starting from the answer instead of from the first post.

A parallel approach to this kind of problems is to look for the repeated text not using pattern recognition, but using a statistical approach to the whole text based on the cluster analysis or on the numerical translation of the messages.

We are working on this kind of solution, our workaround in the meanwhile gives us the possibility to have linguistic query and information not counterfeited from a wrong frequency results.

6.4 Emoticon

An emoticon is a symbol or combination of symbols used to convey emotional content in written or message form. The word is a portmanteau of the English words emotion (or emote) and icon. In web forums, instant messengers and online games, text emoticons are often automatically replaced with small corresponding images, which came to be called emoticons as well. An example of a well known emoticon is a smiley face :-) (Wikipedia - <http://en.wikipedia.org/wiki/Emoticon>).

In the newsgroup text there is a large use of “emoticons” used inside the message or at the end as signature.

It is really important to recognize this set of characters and mark them in the right way and not as a sequence of punctuation.

Is necessary to implement an algorithm that mark the difference between a simple “...” at the end of a text from ☺.

The original intent of the parser was to build an emoticon library that grows with the growth of the corpus in which we store all the emoticons of the texts. Implementing this solutions we realized that it is really difficult to isolate a set of emoticons because there are a lot of graphical symbols and a lot of variations on the same.

For example ☺ the simplest one written as : -) can be also : -)) or : -))) and so on, so we have implemented a pattern recognition, using the regular expression, to find them

Table 13. Pattern sequence for emoticon recognition

```

my $templateEmot3chr =
'|(?:[\#080\(\)]|\&lt;|&gt;)?' # cappello (opzionale)
.'(?:[1 B\$\:\#\;\>8\|\?]\|\&lt;|&gt;)' # occhi (obbligar.) .
.'[-\*\^\=\~o]\' # naso
.'(?: \' #una o più ripetizioni di ciò che segue (da non mem.)
.'[\(\pP\|X\|D\?I\|co\$\$5\*90\+rJ\@vizj\#vxCBfSl\}T]\' #bocl.
.'|\&lt;|&gt;\' # boc2 o barba es. :-)<
.'|\&lt;|&amp;|\' # boc3 o barba3

```

7. Newsgroup tokenizer

We have now discussed about the relevance of tokenization in a corpus encoding process. We wish to underline some peculiar features of tokenization in this particular form of the text. It is important to notice that even the standard pattern marked in a tokenization script are not that simple in a context in which all the text shows some “strange” behaviour that raises some more difficulties.

For example the recognition of numeric expression and the tokenization of them in respect of the number that they represent, that is a work typical of each tokenizer, becomes a little bit difficult if in the plain text there are tokens or parts of tokens made of a sequence of numbers as a part of some other segment, such as a URL address.

So our tokenizer, which is developed with a common part and a segment language dependent, marks and consequently tokenizes some usual patterns such as calendar date, phone numbers, abbreviations, numerical expressions, and some non standard patterns, which are typical of these kind of texts - such as genitive and auxiliary, e-mail address, URL and news addresses, emoticons...

We have mentioned the genitive and auxiliary because even in non English corpora we often find a large set of English expression, so we have decided to tokenize them in the right way with respect of the grammatical English rules.

We have developed a script that works in an interactive way, marking and isolating, step by step the different level of words or of segment of words, leaving the ambiguous set of words to be tokenized until the next step of the script in which the global tokenization structure could have disambiguated the segment.

The input is a line to parse, step by step, the line is split isolating the different rules of tokenization, starting with the one on which is easy to define how separate the token, ending with the more complex pattern.

Table 14. R1 R2 R3 Lines to tokenize

R1	On Thu. 06 May 2004 Pamela Sachs, president of the Air Canada flight attendants local of the
R2	Canadian Union of Public Employees, said the airline's unions agreed to concessions that will
R3	save the airline \$1.1 billion a year for 6 years.

Table 15a. R1 line Tokenization

1	On	8	the
2	Thu	9	Air
2	,	10	Canada
3	<date>06 May 2004</date>	11	flight
4	Pamela	12	attendants
5	Sachs	13	local
2	,	14	of
6	president	15	the
7	of		

Table 15b. R2 R3 lines tokenization

R2			
1	Canadian	9	<gen_aux>'s</ gen_aux >
2	Union	10	unions
3	of	11	agreed
4	Public	12	to
5	Employees	13	concessions
2	,	14	that
6	said	15	will
7	the		
8	airline		
R3			
1	save	6	year
2	the	7	for
3	airline	8	<num>6</num>
4	<money>\$1.1 billion</money>	9	years
5	a	10	.

Each language has its peculiarities and tokenization difficulties. We have implemented a specific module for Italian, Spanish and English.

The Italian tokenization has for example the peculiarity of the apostrophe that requires a specific pattern. As a matter of fact it is not so easy to distinguish from accent, quote, English interference, considering also a lot of variant in the writing form, for example I'm or I' m or I 'm. So, with an interactive script we first tokenize the accent word written with the apostrophe mark for editing problems, then we look for the most frequent English form, we look for an apostrophe mark in front of a word in the context of the token to discover a quotation mark, and eventually we consider it as apostrophe.

Table 16. Italian tokenization pattern from it.cultura.linguistica

```

Newsgroups: it.cultura.linguistica
Subject: =?iso-8859-
1?Q?Re=3A Nome_della_22v=22_era_su icli_Re=3A?=?
=?iso-8859-1?Q?_Storia_della_Tivv=F9?=?
From: Giovanni Drogo <drogo@rn.bastiani.tt>
Date: Fri, 27 Sep 2002 11:34:11 +0200
Message-ID: <Pine.OSF.4.30.0209271124350.20440-
100000@poseidon.mi.iasf.cnr.it>

On Thu, 26 Sep 2002, Riccardo Venturi wrote:
> Per altro va da se' che il tedesco, nelle sue fasi più antiche,
non
> ha mai avuto una grafia standardizzata.
>
> > Date per il passaggio f-v ?
> Scusa, qui non capisco cosa vuoi dire. Puoi spiegarti meglio?
Forse e' una domanda senza risposta in base a quanto sopra.
Ma la/e domanda/e ere/no : quando certe parole hanno preso a
essere
scritte come f e quando come v, ed idem per un eventuale cambio di
pronuncia (oggi un tedesco o un olandese anziano ha una certa
tendenza a
pronunciare "f" anche una parola straniera scritta con "v").
In parte e' il mio vecchio discorso su quali lettere dell'alfabeto
sono
"non ambigue" nella maggioranza delle lingue (*) e quali invece
tutti si
aspettano di pronunciarla allo stesso modo.
P.eg. la "c" e' un chiaro esempio di lettera ambigua, chi se la
aspetta
come k, chi come c dolce, chi come zeta. Ma la k o la t (tanto per
dire)
tutti se le aspettano uguali (o circa). Ovviamente le convenzioni
utilizzanti digrammi sono escluse.

```

For Spanish corpora the difference from Italian ones is the different use of punctuation, also at the start of a text, the use of \tilde{c} , the different list of abbreviation the accented characters within a word and not only at the end.

Table 17. Spanish tokenization pattern from es.charla.gastronomia

```

=====
Newsgroups: es.charla.gastronomia
Subject: RE: Jimenez Drambuie;': Re: ¿Es adecuado hablar sobre
vino en este Grupo?
From: "Hans" <vladivostok25QUITAESTO@hotmail.com>
Date: Tue, 29 Oct 2002 15:28:36 +0100
Message-ID: <apm5iv$2pn8o$1@ID-102817.news.dfncis.de>
References: <anvbbi$11646$1@ID-102817.news.dfncis.de>
<AjFq9.33320$0B5.2659634@newsread2.prod.itd.earthlink.net>

No, Pang, gracias, pero no; el Drambuie no es si no licor de
whisky (hay
algún otro similar, pero de la marca White Horse cuyo nombre he
olvidado) No
sabía que existía una versión SECA pero la buscaré para catarla
por pura
curiosidad (¿tiene algún sentido una versión seca de una versión
dulce de
algo seco, es decir, del whisky?). A la combinación scotch +
drambuie se le
llama, s.e. ú o., "Clavo oxidado".

```

As we have been trying to show the use of non standard text gives the possibility to test these suites of tools to verify their validity and attendance by a computational linguistic approach.

Planning the corpus we knew that the manipulation of these kinds of text would be hard and dangerous but scaling the problem step by step we have isolated the warning area of the pre processing process.

Then we have implemented some tools to solve, at least in part, some problems, as we have done for example for the spamming problem, or to isolate and control some phenomena, even if at the moment we are implementing a solution, such as for the repeated text.

References

- Amstrong, S. 1994. *Using large corpora*. Cambridge: CUP.
- Barbera, M. 2007. I NUNC-ES: strumenti nuovi per la linguistica dei corpora in spagnolo. *Cuadernos de filología italiana* XIV: 11-32.
- Barbera, M., E. Corino and C. Onesti. 2007. *Corpora e Linguistica in rete*. Perugia: Guerra.
- Baroni, M., F. Chantree, A. Kilgarriff and S. Sharoff. 2008. *CleanEval: A competition for cleaning Webpages*. Proceedings of LREC 2008.
- Bolasco, S., B. Bisceglia and F. Baiocchi. 2004. Estrazione automatica di informazione dai Testi. *Mondo Digitale* III, 1: 27-43.

- Buzzetti, D. 2002. Digital representation and the text model. *New Literary History* 33, 1: 61-88.
- Casavecchia, S. 2005. Progettazione ed implementazione di corpora di lingua inglese basati sui newsgroups. MA thesis, Facoltà di lingue e letterature straniere Università di Torino.
- Hambridge, S. and A. Lunde. 1999. A Set of guidelines for mass unsolicited mailings and postings (spam*). IETF RUN Network Working Group, RFC 2635.
- Kilgarriff, A. 1996. Which words are particularly characteristic of a text? A survey of statistical approaches. In L.J. Evett, T.G. Rose (eds), *Proceedings of the AISB Workshop on Language Engineering for Document Analysis and Recognition*. Brighton: AISB, 33-40.
- Kilgarriff, A. and M. Baroni. 2006. *Proceedings of the 2nd International Workshop on the Web as Corpus*. East Stroudsburg (PA): ACL.
- Krippendorf, K. 1983. *Analisi del contenuto. Introduzione metodologica*. Torino: ERI.
- Evert, S. 2004. A simple LNRE model for random character sequences. In G. Purnelle, C. Fairon and A. Disester (eds), *Proceedings of the 7èmes Journées Internationales d'Analyse Statistique des Données Textuelles*. Belgium, Louvain-la-Neuve: UCL Presses, 411-422.
- Manning C. and H. Schütze. 1999. *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.
- Marengo, C. 1996. *Le parole dell'italiano. Lessico e dizionari*. Bologna: Zanichelli.
- NUNC Corpora. www.corpora.unito.it
- Sinclair, J.M. 2000. Current issues in corpus linguistics. In R.R. Favretti (ed.), *Linguistica ed informatica*. Roma: Bulzoni, 2-38.

THE C-ORAL-BRASIL CORPUS

Tommaso Raso, Heliana Mello

Universidade Federal de Minas Gerais (Brazil)

1. The project

In this paper we present the C-ORAL-BRASIL corpus for the first time¹. This is a spontaneous speech Brazilian Portuguese (BP) corpus, which is being compiled within a research project coordinated by Tommaso Raso, with the collaboration of Heliana Mello, at the Federal University of Minas Gerais, Brazil. Along this paper it will become clear that not all the project phases have the same degree of definition, but all are clearly drafted. The project is financed by the Research Support Foundation of the State of Minas Gerais (FAPEMIG), the National Council for Technological and Scientific Development - Brazil (CNPq), the Federal University of Minas Gerais (UFMG) and the Santander Bank². The corpus was designed from its inception to study the informational structure of Brazilian Portuguese (BP) and its illocutions based on the Informational Patterning Theory (Cresti 2000)³. It comprises the fifth branch of the C-ORAL-ROM (Cresti and Moneglia 2005), the reference corpora of the four major Romance languages of Europe⁴.

The Project was officially launched in January 2007, but even before that, through the development of two master theses as a pilot project⁵, two texts – one informal-dialogic and the other tending to be formal-monologic – comprising 5,395 words, were submitted to all the methodological steps foreseen in the project. That is, recording, transcription, segmentation and informational tagging. Besides that, in the two texts, the major speech measurements and the informational units of Topic and Appendix were studied⁶. Along the following two years (2007-2008) other objectives were added to the project, notably, the study of modality in BP, within the LABLITA group paradigm⁷, and the comparison between BP and European Portuguese, with a view to identifying BP features that might enlighten language contact studies.

Today the project has a staff made up of three doctoral students, two master's degree students, four undergraduate research assistants (three have been awarded scholarships) and a statistician, besides its coordinator and a senior researcher. Several papers have been already published by the group⁸ and the corpus

compilation is at an advanced stage. The expectation is that up until the end of 2009 the informal part of the corpus will have been completed – this being the most relevant half for spontaneous speech studies to be carried out. The other half, made up of formal speech, will start immediately after the accomplishment of the first half.

2. The corpus: general features

The corpus should be comprised of at least 30 hours of recordings, those being divided into 15 hours of informal speech and 15 hours of formal speech. The formal part of the corpus has not been entirely defined yet. In principle, the same architecture adopted by the C-ORAL-ROM⁹ will be followed, but it is likely that some adaptations to the Brazilian sociolinguistic context will be needed. The defining characteristics of the informal half of the corpus, which have been entirely decided upon and are in an advanced developmental stage, will be the focus of this paper.

The architecture of the informal part of the corpus dictates a minimum of 15 hours of recorded speech, distributed in a minimum of 100 texts made up of an average of 1,500 words each. A small percentage of texts (not more than 25) can be made up of either longer texts (average of 4,500 words), or shorter ones, as long as they present a consistent textual autonomy. From the total number of texts, 20% will represent public contexts and 80% family/private ones. In each category, i.e. public and family/private, a third of the texts will be monologic and two thirds dialogic texts or conversations (i.e., dialogs carried by more than two participants). In principle, the two latter categories will have a balanced proportion.

As it was the case in the C-ORAL-ROM, a single diatopy is represented. In our case, the diatopy of the State of Minas Gerais, particularly the urban area of its capital city (Belo Horizonte), is represented. Therefore, at least 50% of the speakers in the corpus are *Mineiros*, but it is likely that an even higher percentage will ensue in the end. Traditionally the *Mineiro* speech has been divided in three major diatopic varieties¹⁰. All of them will be represented in the corpus, but the balance of this representation is not one of the project goals; thus the metropolitan area of Belo Horizonte should have a greater share of the data compiled.

The corpus attempts to represent the diastratic variation to a certain extent, which in the Brazilian context is especially important due to the social-history of BP. However, no statistical balance will be sought in this regard. Furthermore, in all the different constituent parts of the corpus, interactions among speakers from several socio-cultural strata will be included, both upon interacting with speakers from the same stratum they belong to and with speakers from different strata. As for the level of schooling classification followed by the C-ORAL-ROM, some

adaptations needed to be done so that they would better represent the Brazilian society. It is needless to mention that there are very meaningful differences between the Brazilian and the European societies. The following chart illustrates the different criteria followed by the C-ORAL ROM and the C-ORAL-BRASIL:

C-ORAL-ROM	C-ORAL-BRASIL
1 (no schooling or primary schooling)	1 (no schooling degree: up to incomplete primary schooling)
2 (middle school)	2 (up to college degree, as long as this degree is not needed for current professional activity)
3 (college student or college degree)	3 (current professional activity requires college degree or higher degree)
X (unknown)	X (unknown)

The differences in criteria specified above reflect the differences between the European and Brazilian contexts as far as the relationship between schooling and linguistic patterns are concerned. Evidently, no schematic classification can be entirely satisfactory, but we believe our choices properly represent the need to amplify the range of the intermediary stratum in BP vis-à-vis its corresponding one in the European project. In fact, in the C-ORAL-BRASIL the extension of the intermediary stratum (2) is achieved by including a share of what is inserted within strata 1 and 3 in the C-ORAL-ROM. This choice attempts to reflect greater profile variability within Brazilian educational institutions when compared to European ones.

The level of variation which is focused upon in the corpus and which is statistically significant pertains to the diaphasic level, since this really has an impact on speech structural variation. The diaphasic variation is represented in the corpus according to the following distribution: formal versus informal speech; within informal speech, public versus family/private contexts; within each context, three interactional typologies, i.e. monologue, dialogue and conversation (more than two participants); within each interactional typology, maximal variation of communicative situations.

Recordings are transcribed according to the CHILDES-CLAN system (MacWhinney 2000), implemented through the prosodic annotation criteria created by Moneglia and Cresti (1997).

3. The diaphasic variation

Each one of the interactional types profits from the maximal situational variation achieved through our recordings. The monologic type encompasses topic variation (life narratives, interviews, work monologues, jokes, assorted narratives, etc.),

variation as far as the recorded individual's profile is concerned (family members, friends, clients, workmates, children, etc.), as well as variation in places where the recordings were carried (workplace, friends' and families' homes, restaurants, etc.). On the other hand, dialogues and conversations besides having the same above mentioned variations for individuals recorded and places where recordings were carried out, also present a broad variation as far as the activity being executed during the interactions is concerned, e.g., two or more people cooking together, two or more people working together at a computer, an individual explaining to others how a technological instrument or a computer program work, client(s) interacting with attendant in a store, a mason and an engineer checking a construction work, two people grocery shopping, two waiters waiting at a party, two or more teachers talking about work, two or more people talking in a car, two or more people chitchatting at home, two or more people eating out at a restaurant, two or more people doing a budget balance together, two or more people studying together for an exam, a private lesson, two or more people visiting an apartment to be rented, etc. Therefore, the situational variation, within each recording type, is given by the combination of the following variables, in a scale of relevance:

- Type of activity accomplished during the interaction. As the activity changes, so does the communicational situation;
- Number and individual typology of activity participants. As the number and/or the participant typology change, so does the situation;
- Place of interaction. As the place of interaction changes, so does the type of situation;
- Topic of interaction. The change of topic contributes to the change in the situation, even though, by itself, it is not sufficient to differentiate situations.

An issue that presented unforeseen complexity was how to characterize and differentiate public versus family/private contexts. The decisions taken as far as this differentiation is concerned in the C-ORAL-ROM seemed to vary among the groups who contributed to the project. Naturally more than one view could be adopted there. But for our project it is very relevant to have our criteria clearly established. The characterization of public contexts versus private contexts should be understood based on the role played by the speaker at the time of the interaction. A given speaker might interact acting upon her/his role as an individual, as usually happens in interactions among friends or family members, or she/he might act taking into account a given social role, be it due to work or power/social relationships in a given situation. That is to say, what really conditions the classification of the situation is not its actual occurrence in a public place, but the role that the interactants choose to have at that given situation: if the role chosen is that of a sister/brother, friend, or individual, the relationship is private; on the other hand, if the role is that of a

professional, a citizen, a representative of a given social-institutional entity, then the relationship is public.

Therefore, it is not the presence or absence of strangers in a given interactional context being recorded that necessarily determines our choice for the public or private characterization of an interaction. Thus, a conversation between two close friends at a restaurant crowded with people cannot be automatically labeled as public, as much as a professional interaction between an advisor and an advisee in a closed door office cannot be considered private, even though the two are all alone in the office. It should be clear that some factors may favor the public or private characterization of a situation, but no individual factor on its own can account for such characterization. What is defining for this characterization is the speaker's behavior – if she/he behaves based on her/his social role (client, professional, employee, citizen, etc.) then the interaction tends to be public, on the other hand, if her/his behavior is based on the individual's own identity, the interaction tends to be private. A couple of examples should foster a better understanding of the decision taking criteria just mentioned: in one occasion we recorded an interview with a restaurant owner carried by her sister. The interview was recorded at the restaurant and the questions dealt with aspects related to the business of running a restaurant. Our initial expectation was that this would be a family/private interaction, given the close relationship between the two interactants. However, upon listening to the recording, we realized that the person interviewed had kept a behavior absolutely distinct from her usual one with her sister along the whole interview. She behaved as a professional - in this case, a restaurant owner who illustrates, albeit informally, the characteristics of her business. In this example, the place and topic of the recording were more relevant than the personal relationship between the interactants to the shaping of the speaker's behavior. The same kind of phenomenon occurred when a sales representative was interviewed by his daughter, at home, about his profession. In this case, the topic of the interview alone was enough to determine that the speaker's behavior was not familiar/private. On the other hand, we have a recording in which a group of college students talk beside an elevator at the university, in which place there was ample traffic of students, professors, staff, and strangers at large. Nevertheless, the content and tone of the conversation were clearly private.

4. The header

The metadata in the headers follow the same criteria as those adopted by the C-ORAL-ROM, with the only exception of the schooling strata classification, already mentioned in this paper, and the acoustic quality, as explained in the next section. Therefore an example of a typical header in the corpus would be like the following:

@Title: Daughter
 @File: ifammn06
 @Participants: CAR, Carmosina (woman, C, 1, house-keeper, narrator, Alpercata (MG))
 MAR, Maryualê (woman, B, 3, professor, intervenient, Florianópolis)
 @Date: 12/04/2008
 @Place: Belo Horizonte
 @Situation: narration about how CAR adopted her youngest daughter, CAR's kitchen,
 CAR makes lunch, not hidden, researcher participant (CAR works as housekeeper at the
 researcher's house)
 @Topic: daughter's adoption
 @Source: C-ORAL-BRASIL
 @Class: informal, familiar/ private, monologue
 @Length: 9'51"
 @Words: 1508
 @Acoustic_quality: AB
 @Transcriber: Maryualê M. Mittmann
 @Revisor: Heloisa P. Vale
 @Comments: text collected and recorded by Maryualê M. Mittmann. CAR pronounces
 "dócia" and "vivendos" when it should be "dócil" and "vivendo". Sometimes CAR calls
 the researcher Mara and not Mary.

Through the examination of the metadata in the headers it will be possible to provide statistical scores pertaining to gender, age, schooling, occupation and communicative role played by the speaker, as well as provide information about the activities carried on along the recording, as well as place and topic of the recording.

5. The recording equipment and the acoustic quality

With a very few exceptions, the recordings were done in .wav format with the following equipment:

- PDD60 Marantz digital recorders, with 2 gigabytes Compact Flash memory card;
- Sennheiser Evolution EW100 G2 wireless kits (receiver, transmitter, clip microphone), with 2 battery/recharger kits adapted for the receiver, or with its own battery;
- omnidirectional Sennheiser MD 421 microphones, with Hunter PMP 103 support, RCL303569 6 meters cables or a wireless system.

The clip microphones are used in monologic and dialogic situations and the omnidirectional microphone in conversational situations. However, we have obtained some good recordings done with two clip microphones in which three or

four people interacted in a spatial configuration which facilitated the accuracy of the recording. In order to minimize the risk of having low quality conversation recordings, we intend to start employing a mixer with up to six clip microphones, which will guarantee that each one of the participants in a conversation have their own monodirectional microphone.

This equipment supports recordings done in natural environments with a sound quality that allows at least the identification of the F0 curve which is crucial for the intonation analyses carried by the WinPitch software; however, in most cases, the quality is so good that it is also appropriate for segmental analyses to be done.

In most cases when the recordings are executed with clip microphones they are kept in stereo mode. This allows for the use of more resources in speech analysis; for example, it will make it possible to better identify individual speakers in a recording in which there is speech overlap. Overall it also guarantees a much better sound quality.

The few cases in which this equipment was not used are those undertaken in acoustic cabins or in .mp3 format.

The recordings are only done with informants that willingly sign the permission from authorized by the Ethics Committee at UFMG¹¹.

The acoustic quality is indicated in the header of each transcription. The criteria followed to indicate acoustic quality differ from those found in the C-ORAL-ROM due to the technological changes that have occurred since the recordings for that corpus were undertaken (some of the recordings date back to the 1980's) and the inception of recordings for the C-ORAL-BRASIL in 2007. The C-ORAL-ROM employed the letters A, B and C to indicate, respectively, excellent, good and acceptable acoustic quality. In the collection of data the letter D would indicate that the sound quality of a given file was so poor as not to be appropriate for inclusion in the corpus. In the C-ORAL-BRASIL the letter classification has been maintained; however, two differences must be mentioned:

1. in principle, the same letter will indicate a better quality file in the C-ORAL-BRASIL than in the C-ORAL-ROM due to the technological advances already mentioned;
2. within the C-ORAL-BRASIL we have adopted a more complex quality code in which besides the single letters (A,B,C), double letters are used to indicate intermediary quality (AB, BC, CD).

The indication of acoustic quality criteria surmise the observation of the following parameters: the possibility of hearing turns, spectrum clarity, presence or lack of noise, reverberation, gain, overlappings, and recording environment. An A is assigned to recordings that will allow, in principle, many kinds of phonetic study; on the other hand, an assigned D recording will only be usable for morpho-syntactic and lexical studies due to the fact that in the majority of cases the F0 is either

unreadable or unreliable. For the intermediary scorings the F0 curve is good for at least 60% of the sound files, the other 40% present difficulties in a variable number of ways. The CD score indicates that in principle the recording could be included in the corpus, but it might as well not be used for a host of reasons. Naturally, the acoustic quality indication bears upon a variable degree of rigidity dependent on the interaction type and the actual interactional situation. The criteria are stricter in what regards monologic interactions, less rigid for dialogic interactions, and even less rigid for conversational interactions. It should be clear that the sheer fact that conversations imply a larger number of interacts prevents the achievement of optimal quality. Analogously, it cannot be expected that a dialogue between two people recorded in a supermarket will have the same acoustic quality of a recording taken at a silent place. Therefore, the conclusion is that the score letter indicates the acoustic quality in itself, but it is possible, for example, that a conversation scoring A, if evaluated in sole regard to its acoustic quality, might correspond to a monologue scoring AB.

6. The transcriptions

The transcriptions are done in accordance with the CHILDES-CLAN system (MacWhinney, 2000) implemented for prosodic tagging as developed by Moneglia and Cresti (1997).

The transcriptions are done on an orthographic basis; however several adaptations are being introduced to render better some speech phenomena found in BP which we consider especially important to be noticed. The transcription criteria will not be entirely discussed below; however some relevant examples will be provided. The transcriptions attempt to capture phenomena that might reflect lexicalizations and grammaticalizations in progress. At the same time, the transcribed text cannot withhold such difficulties as to generate comprehension problems for the reader and excessive difficulties for the transcribers, especially in those cases in which the perceptibility of the phenomenon to be transcribed is such as to render improbable a high level of agreement among transcribers. This creates the necessity for our balancing the rendering of linguistic phenomena and the readability of the text or the feasibility of the transcription. Some of the phenomena that the transcriptions attempt to capture are the following:

- lack of copular verb *ser* in focus structures when the verb is not actually uttered: *Maria que faz* VS. *Maria é que faz*; *que que cê acha disse* VS. *que é que cê acha disse*; *quando que ele vem* VS. *quando é que ele vem*; etc;

- lack of plural marking in nouns and adjectives. All item plural marking in NPs in BP is frequently omitted when it is already signaled in the determiner (or leftward most element) of an NP: *os meninos bonitos* > *os menino bonito*;
- lack of person marking in several verb forms (*nós diz*; *es faz*; etc.);
- subject pronoun cliticization in second and third person forms: *você(s)* > *cê(s)*; *ele* > *e'*; *ela* > *ea*; *eles* > *es*; *elas* > *eas*;
- preverbal negation potential cliticization: *não* > *nũ*;
- articulated preposition contraction: *para* + def. art. > *pro*, *pra*, *pros*, *pras*, *prum*; *de* + indef. art.: *dum*, *duns*; *com* + def. art.: *co*, *ca*, *cos*, *cas*; etc.;
- apheresis: in verbs, such as *estar* (*tô*, *ta*, *tando*, *tar*, etc.) and others (*agüentar* > *güentar*; *espera* > *pera*; etc.) or in some other parts of speech, such as *obrigado* > *brigado*;
- apocope in masculine diminutive forms: *sozinho* > *sozim*;
- some formulaic expressions: *Nossa*, *No'*; *Vixe'* (*Virgem Maria*); etc.;
- and others...

7. The segmentation

The transcriptions are segmented into utterances and tonal units, according to the same criteria followed by the C-ORAL-ROM, with a very few and slight changes. The symbols adopted are discussed below:

- the double bar (//) indicates intonation break with terminal value, that is, the end of an utterance. An utterance is defined in the Informational Patterning Theory as the smallest pragmatically autonomous unit, holding an intonation profile perceived as terminal (t'Hart, Cohen and Collier 1990);
- the single bar (/) indicates an intonation break perceived as non-terminal, therefore indicating the border of a tone unit (with an informational value, in principle) within an utterance;
- the sign (+) indicates an interrupted utterance. It has a terminal value but indicates that the utterance was not completed for whatever reason;
- the sign ([/n]) indicates retracting. The *n* at the side of the bar indicates the number of words involved in the retracting.

Some examples of how breaks are assigned in speech fragments from our corpus are presented below.

Example 1

These are some cases in which a sequence of simple utterances appears in the same turn – each utterance carries a terminal break only.

1A

*PAU: bom // Rogério //

[*PAU: well // Roger //



1B

*FLA: é // me falaram // que ele é muito <bom> //

[*FLA: yes // I've been told // that it is very <good> //



1C

*FLA: hhh o nosso tá longe // tá em outra cidade //

[*FLA: hhh our is far away // it's in another town //



1D

*REN: ham ham // <é> // tá // tá certo // obrigada //

[*REN: hum hum // <yeah> // ok // it's ok // thanks //



Example 2

This is a complex utterance, with non-terminal breaks within it. In the example, the comment unit, that is, the only unit which carries pragmatic autonomy, is identified.

*FBA: Tudo que é de bom / pra gente / que a gente tá se sentindo que realmente tá fazendo / né / e [/] e que tá dando retorno / a gente continua //==

[*FBA: Everything which is good / for us / that we are feeling that we are really doing / right / and [/] and that is giving feedback / we continue //==]



Example 3

This is an example of an interrupted utterance, followed by the interruption uttered by the other speaker in the interaction recorded.

*PAU: aí / por exemplo +

[*PAU: then / for example +]


*ROG: aqui já tá dando [/4] aqui já tá dando a altura //

[*ROG: here it's already reaching [/4] here it's already reaching the height //



Example 4

An example of retracting (cf. e.g. 3)

*ROG: aqui já tá dando [4] aqui já tá dando a altura // 
 [*ROG: here it's already reaching [4] here it's already reaching the height //]

Example 5


A transcribed and segmented speech excerpt.

*PAU: bom // Rogério // 
 [*PAU: well // Roger //]
 *ROG: hum //
 [*ROG: yes //]

*PAU: cê sabe aqui como é que [3] como é que tem que fazer esse muro aqui / né //
 por que que cê não tá trabalhando com linha aqui / o' //
 [*PAU: you know here how it should be [3] how this wall should be built / right //
 why aren't you working with the line here / look //]

*ROG: ah / então eu vou [2] eu vou &f +
 [*ROG: ah / then I will [2] I will &d +]

*PAU: hein //
 [*PAU: what //]

*ROG: eu vou &coloc [3] eu vou suspende' mais um pouquinho aqui / e vou pegar
 a linha e colocar por cima // 
 [*ROG: I'm going to &plac [3] I'm going to raise a little more here / and I'm going
 to get the line and place it on top //]

*PAU: ah / porque se não + aqui / o' // aí / por exemplo +
 [*PAU: ah / otherwise + here / look // there / for example +]

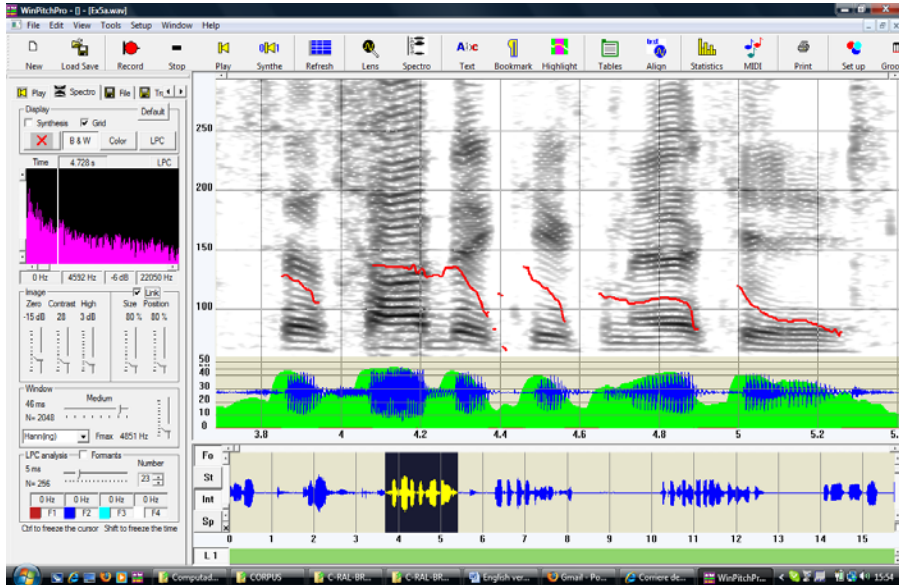
*ROG: aqui já tá dando [4] aqui já tá dando a altura //
 [*ROG: here it's already reaching [4] here it's already reaching the height //]

*PAU: o' / aqui + não // tá dando a altura daquele que a Isa /
 [*PAU: look / here + not // it is reaching the height of that one that Isa /]

*ROG: é //
 [*ROG: yeah //]
 *PAU: / marcou lá / <né> //
 [*PAU: / marked there / <right> //]

*ROG: <que a dona> Isa marcou ali // que a dona <Isa marcou> //
 [*ROG: <that Mrs.> Isa marked there // that Mrs. <Isa marked> //]

There is a spectrographic representation (using the WinPitch software by Philippe Martin)¹² of a piece of the recording transcribed above, exemplified below. It is a recording carried in an open space, and it features a mason and an engineer talking during the building of a wall. This recording reached an acoustic quality AB.



8. The transcription and segmentation processes: validation

The transcription and segmentation processes are utterly complex¹³. A long training period was undertaken in several stages that include three workshops and a graduate level course. All the members of the project staff extensively practiced segmenting and transcribing texts. It is relevant to mention that the segmentation process is done together with the transcription process, as both are based on acoustic perception. After the training period, a group of ten transcribers was selected to undergo some tests. The detailed results of the tests along with a full description of the training process will be reported elsewhere. In this paper we will only refer to the major aspects of those processes.

After the training period the ten potential transcribers were sorted in three groups, two of which had three members and one had four members. The sorting in groups was done according to the following criteria: Group #1 was established with

three students who had manifested outstanding aptitude in segmenting during the training period and at the same time demonstrated full engagement in the project. Those students had the following profile at the time: one was a doctoral student, the other a masters' student and the last a senior undergraduate research assistant who was about to enter the masters' program. Group #2 was established with three students who ranked second in the same aspects mentioned above for Group #1. The students' profile at the time was: a doctoral student, a candidate to the doctoral program, and an undergraduate research assistant. Group #3 was made up of students who were rookies in the project or had demonstrated less availability to undertake chores in the project. Their profile was: three undergraduate research students (two of which held scholarships) and a masters' degree student. The latter was not officially a member of the project staff, but had undergone the training process and had shown a high degree of enthusiasm to be integrated in our work. The three undergraduate students had less time experience in segmenting and therefore also had better chances of improving their skills with the experience to be acquired.

Each group underwent a series of tests after which, in collective meetings, the inconsistencies in break markings in the transcriptions were discussed. Groups #2 and #3 were reorganized after the initial tests, the reason being the following: we did not need three transcriber groups but only two. It was fully predictable that Group #1 would confirm its high degree of reliability, and Groups #2 and #3 could in principle show different results from those obtained at our first trials. These tendencies were confirmed and Group #1 was maintained as it had previously been and Groups #2 and #3 were merged. In this merger, only one member of Group #2 was kept and three members of Group #3 were upgraded. In the end, we had two groups: Group #1 as already described and Group #2 with four members. Group #3 was eliminated. Presently Group #2 has reached a Kappa test at four result of 0.82 (0.85 if we consider only the best three members), but they needed almost two months more than group #1 to achieve this result. Below a summarized description of the path followed by Group #1 will be reported. A relevant piece of information referring to the transcriptions used in the tests is that they were only provisory and did not have all the features that our project requires. Mostly there were transcription errors, some of which induced different solutions on the part of the segmenters. The consequence was that the disagreements artificially diminished the statistic agreement rate. The results for the tests undertaken by Group #1 are provided below:

1. Segmentation of a dialogic text comprised of around 800 words and a monologic text comprised of 800 words. Kappa test results: 0.820 for the dialogic text and 0.750 for the monologic test.
2. Segmentation of a dialogic text comprised of 1,500 words. Kappa test result: 0.839.

3. Segmentation of a monologic text comprised of around 1,500 words. Kappa test result: 0.839.

These results require a qualitative discussion in order to be better understood. Firstly, it was our objective not to have any disagreements as far as terminal breaks are concerned. In all the tests a few disagreements occurred (seven in the worst attempt and two in the best one). We checked all the disagreement cases and it became immediately clear that a few of them were simply due to a lack of attention in the process of segmentation, openly recognized by the people who had done them upon being questioned about their decisions. The other disagreements occurred due to issues yet not resolved involving the transcriptions. A typical example of the latter comprises turns in which there are only laughs and no words uttered. In some cases the person segmenting the texts decided that a laugh turn should not be segmented; in another, however, the decision was that there was a terminal break at the end of the laugh turn. The described problem generated the better part of extreme disagreements between the insertion of a terminal break and a lack of break insertion. Excluding the clear cases of distraction and ambiguous laugh turns, there was no disagreement in the segmentation of terminal breaks and lack of break insertions.

Secondly, the checking up of all the tests generated an immediate reduction in disagreements in the discussion phase. In most cases, whoever represented the disagreement from the majority's decision (without being aware that hers/his was the disagreeing decision), either did not acknowledge that she/he had made that choice or changed hers/his decision immediately. These two factors generated an improvement in the test results, which had already been originally excellent from a statistical point of view, as the Kappa score shows.

After the phase reported above we assumed that the results already guaranteed a satisfactory basis for our work, but in order to achieve results that demonstrated excellence we decided to pursue more tests which differentiated terminal breaks from non-terminal breaks. We proposed a first test that focused on terminal breaks. The result was 0.901, and it would have been even higher if we had ignored the few disagreements accounted for by expected ambiguities due to the poor quality of the transcription. The second test dealt with non-terminal breaks. The result was 0.660. The average for this result and the one for the terminal breaks would be superior to 0.800, already to be considered excellent, but we wanted to further investigate the reasons for the disagreements in decision making. We got the distraction effect, which was promptly corrected at the checking sessions, and it also became clear that one of the people segmenting had a tendency not to perceive weak breaks. This factor has little effect on the perception of terminal breaks, however it does clearly impact non-terminal breaks. At this point we decided that the best way to proceed to the beginning of the transcription work was to leave the revisions to the two transcribers who exhibited a higher degree of agreement in their decisions. In spite

of the continuation of the discussion and test processes, we decided to initiate the transcriptions. All the transcriptions to be done during the period in which discussions and tests will be still taking place shall undergo a further revision after the conclusion of the tests.

In sum, even working with approximate transcriptions that have not been revised and which therefore generate ambiguity, we can count with a Kappa agreement clearly superior to 0.800; that is, a result considered statistically excellent. This Kappa result will automatically be improved with definite transcriptions and revisions. After the revision process is concluded, new tests shall confirm a higher agreement score.

The first tests applied to Group #2 in its different compositions averaged between 0.790 and 0.750 for dialogic texts and between 0.770 and 0.680 for monologic texts, and, after two more months of training, we obtained, as we have already mentioned, 0,82 (0,85 considering only the three best trained individuals).

9. The setting up of a mini-corpus for study

In order to guarantee the availability of material so that informational studies can be started, it was decided that our work was to be carried in two distinct fronts. On the one hand, a balanced mini-corpus, made up of 20 texts, 3,000 utterances and 30,000 words, was to be developed. This mini-corpus would be fully transcribed, segmented, aligned and informationally tagged. On the other hand, the transcription, segmentation and alignment of the other texts would proceed.

The setting up of the mini-corpus has in view the pursuit of maximum quality criteria. The mini-corpus, in its various branching, only encompasses texts that show the best possible combination of the following parameters:

- branching representativity. The text should be a good prototype for the kind of variation it illustrates (public versus family/private; monologic versus dialogic versus conversational);
- greatest possible variation in activities undertaken. Two texts should never represent the same communicative situation;
- high acoustic quality. This is given by the following factors: spectrographic quality; little or no background noise; little or no signal reverberation; voice clarity; gain; reliable F0 calculation; low overlapping percentage;
- speaker diversity. The same speaker cannot be in more than one recording file, except in the case she/he is the protagonist of a monologue or a dialogue and also takes part in a conversation.

Approximate parity in representing male and female voices, as well as age groups;

- diastratic neutralization. Texts representative of a middle ground as far as diastratic variation is concerned are sought. Avoidance of extreme high or low diastratic exemplars;
- attractiveness of content. This is important for two reasons: first a highly interesting text captures the transcriber's attention, therefore guaranteeing better transcription and segmentation results; secondly it adds up to the degree of informational content because it guarantees a more spontaneous speech.

In order to guarantee the maximum quality level for the mini-corpus, the chosen texts are being transcribed exclusively by Group #1 members, and are being revised by its two members with the higher agreement scores. Before setting up a group of taggers, it will be necessary to verify the reliability and the degree of agreement reached in the tagging process in the same way it was carried for the segmentation process.

10. Present stage of the work and next steps to be taken

At the present stage we have gathered 180 recordings which cover the whole spectrum of the informal part of the corpus. Because several recordings are actually too long, in some cases over several hours, it is likely that several texts can be extracted from some of these long recordings. The number of texts already available is no doubt sufficient for the goal of the corpus. Nevertheless, the recording process will never be considered as concluded at this point, as it is always desirable that more texts are available, so that choices can be made in order to better improve the corpus up to the moment it is considered as finished. Besides, even after the publication of the corpus within the parameters proposed in its projection, its enlargement will always be desirable because this guarantees the possibility for several other studies to be done (similarly to what has been done to the Italian LABLITA corpus, which is much larger than that published in the C-ORAL-ROM)¹⁴.

As far as transcriptions are concerned, the three expert transcribers (Group #1) who reached an adequate statistical agreement score are about to finish the transcription as well as the revision of the twenty texts selected to integrate the mini-corpus discussed in session 9 above. Group #2 will transcribe and segment the remaining texts that will make up the corpus.

As soon as the mini-corpus texts are entirely transcribed and revised, the three expert transcribers will start the alignment process to then immediately pursue informational tagging. The alignment stage is also the last time in which the revision

of segmentation, with special attention to terminal breaks, is done. After that point, the mini-corpus will be ready for various studies based on it to be carried by the C-ORAL-BRASIL members.

Meanwhile, Group #2 will be working on the transcription, segmentation and alignment of the other texts that will integrate the corpus. These, however, will not be informationally tagged. The whole corpus will be lexico-morpho-syntactically tagged through appropriate software, trained and rule-based taking into account the transcription choices and segmentation criteria decided upon for the C-ORAL-BRASIL. The same macro and Excel sheet used for the C-ORAL-ROM will provide the major speech measures together with the POS-tagging software, that is, for each interactional type (monologue, dialogue and conversation) there will be a calculation of turn, utterance and tonal unit numbers in relation to time and word number; the number of interrupted utterances and retracting phenomena; number of utterances carrying verbs and of those which are verbless. These data will be correlated with the characterization of utterances as simple or complex; frequency of occurrence of negation and conjunctions *E*, *MAS*, *PORQUE* and *QUE*, and their placement (beginning of turn, beginning of utterance, beginning of tonal unit, within a tonal unit and characterization as a dedicated unit – that is, whenever the unit is made up solely by the item under investigation)¹⁵.

These same measures are already being investigated under a qualitative perspective for a group of nine texts (three dialogues, three monologues and three conversations) that will integrate the mini-corpus. A qualitative study allows for the possibility of researching phenomena that cannot be captured by a quantitative inquiry carried by computational resources. For instance, an automatic tagger can tell us whether a given utterance holds a verb or not, but cannot inform us whether a given form is classified as a verb solely under morphological criteria but does not hold this classification functionally. For example, items such as *tá* or *sei* are tagged the same way when they show up in utterances in which they have a verbal value as in *o meu amigo tá bem* “my friend **is** ok” and *eu sei o que estou dizendo* “I **know** what I am saying”, and in utterances in which they function as a signal for agreement, meaning *sim* (yes) or *ok*. Analogously a tagger does not capture the functional value of a reply to a question which bears an echo verb used either to affirm or agree, as it is frequently the case in BP¹⁶. Furthermore, a qualitative analysis will make it clear whether a verbal element fills or not a nuclear slot in a given utterance. Yet another example of the relevance of qualitative analyses pertains to the frequency of occurrence of non-verbal utterances, characterized by the sole presence of elements such as *hum hum* or *ahn ahn*, which are linguistically empty even though are conventionalized as affirmations and negations.

We conclude this paper by introducing one of the prospective studies the corpus will allow us to develop, and which we find most interesting for its large descriptive potential as much as for the possibility of approaching a basic question posed by the C-ORAL-ROM project: whatever belongs to the characterization of speech and that

which is specific of a given language/culture. We are referring to the possibility afforded by the C-ORAL-BRASIL for us to compare BP spontaneous speech to European Portuguese (EP) spontaneous speech, based not solely on segmental, morphosyntactic and lexical parameters but also on those which pertain to the prosodic, informational and illocutionary domains. This comparison will allow us to develop studies in an attempt to answer the following two important questions:

1. To which extent BP and EP, two variants of the same language, are structured the same way or differently, and in the latter case scenario to which extent the difference in structuring is connected to different cultural matrixes? The studies undertaken by the LABLITA group about Italian will be an automatic meter of comparison. The very first studies already developed taking as their bases the BP pilot project indicate significant differences in the informational structure of Italian and BP. To know whether EP is closest structurally to a distinct language which shares a similar culture (Italian), or to a variant of the same language which is culturally distinct (BP) can help us significantly understand the structure of speech. Therefore, one of the next steps we will take in the project will be the tagging of a mini-corpus of EP extracted from the C-ORAL-ROM so that we can start to develop comparative studies taking into account three languages: Italian, EP and BP.
2. The comparison of the three languages (Italian, EP and BP), taking into account the fact that EP shares cultural identity with Italian and the code matrix with BP, would allow us to isolate BP features that could be natural candidates for a language contact investigation. If in some speech aspects, mainly intonation and informational aspects, EP shows a closer resemblance to Italian than to BP, then we would have good grounds to study the same phenomena in BP searching for historical explanations which could bear on the contact of BP with languages from very distinct backgrounds.

Notes

¹ Except for oral presentations at conferences and seminars. The corpus has been presented at The Pragmatics and Prosody Seminar (Rio de Janeiro, August 27 2008), at the Faculdade de Letras-UFMG Study Journey Week (Belo Horizonte, October 20-24 2008), at the VII Brazilian Corpus Linguistics Conference (São José do Rio Preto, November 6-7 2008). None of the presentations has yet been published.

² Financial support includes resources to acquire equipment and bibliography, resources for study trips abroad and for the invitation of foreign researchers, resources for student scholarships and technical support. Five sub-projects connected to an umbrella Project have been financed so far.

³ For the theoretic frame of illocution, see Austin (1962) and Cresti (2005). Many works about informational units within this frame have already been published for Italian and are accessible at <http://lablita.dit.unifi.it/preprint/>.

⁴ There is a collaboration agreement between the Faculdade de Letras da UFMG and the Facoltà di Lettere dell'Università di Firenze coordinated by E. Cresti and T. Raso. It supports, among other activities, the collaboration between the two partner groups for the compilation of the Brazilian Portuguese Spontaneous Speech Corpus and informational structure crosslinguistic studies. Profs. E. Cresti and M. Moneglia have given two mini-courses and three workshops at UFMG in Feb. 2007 and August 2008. Prof. T. Raso was invited by the LABLITA laboratory in July 2007 and July 2008; Profs. T. Raso and H. Mello were invited to spend Feb. 2009 working at the LABLITA at the Università degli Studi di Firenze.

⁵ Ulisses (2008) and Alves de Deus (2008).

⁶ Some results were published in Raso et al. (2007); Raso and Ulisses (2008) (both accessible at the Revista de Estudos da Linguagem site); Raso and Mello (in press).

⁷ This sub-project is coordinated by H. Mello. See Cresti (2002, 2003); Tucci and Moneglia (in press); Tucci (2005, 2008a and b, 2009, in press). These articles, along with several others referred to in this paper can be accessed at LABLITA site.

⁸ Besides those indicated in footnote 5 and 6, see also Maia Rocha-Raso-Andrade (in press).

⁹ Cresti and Moneglia (2005); Cresti et al. (2004); Moneglia (2005).

¹⁰ Zágari (2005).

¹¹ COEP

¹² See the WinPitch site.

¹³ For the procedures followed in the C-ORAL-ROM compilation, see Danieli et al. (2004); Moneglia et al (2005).

¹⁴ See “Corpora” section in the LABLITA site.

¹⁵ For some studies on the languages comprised in C-ORAL-ROM, see Moneglia (2004, 2006); Cresti and Moneglia (2007).

¹⁶ For example, in the following dialogic excerpt: *XYZ: o João viajou pra Itália // *XYZ: viajou //. The second utterance has the functional value of an agreement, therefore meaning *yes*.

References

- Alves de Deus, L. 2008. O Tópico no português do Brasil. Master Dissertation, UFMG.
 Austin, J.L. 1962. *How to do things with words*. Oxford: Clarendon.
 Cresti, E. 2000. *Corpus di italiano parlato*. Firenze: Accademia della Crusca, 2 voll.

- Cresti, E. 2005. Per una nuova classificazione dell'illocuzione. In E. Burr (ed.), *Tradizione e innovazione - Atti del VI convegno SILFI*. Firenze: Cesati, 233-246.
- Cresti, E. 2002. Illocuzione e modalità. In P. Beccaria and C. Marelli (eds), *La parola al testo. Scritti per Bice Mortara-Garavelli*. Torino: Ed. dell'Orso, 133-145.
- Cresti, E. 2003. Illocution et modalit  dans le comment et le topic. In A. Scarano (ed.), *Macro-syntaxe et pragmatique. L'analyse linguistique de l'oral*. Roma: Bulzoni, 133-182.
- Cresti, E., F. Bacelar do Nascimento, A. Moreno Sandoval, J. Veronis, J., Ph Martin and K. Choukri. 2004. The C-ORAL-ROM CORPUS. A multilingual resource of spontaneous speech for romance languages. In M.T. Lino, M.F. Xavier, F. Ferreira, R. Costa and R. Silva (eds), *Proceedings of the 4th International Conference on Language Resources and Evaluation*. Paris: ELRA, 575-579.
- Cresti, E. and M. Moneglia. 2005. *C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam/Philadelphia: John Benjamins.
- Cresti, E. and M. Moneglia. 2007. C-ORAL-ROM. Comparing romance languages in spontaneous speech corpora. In T.C. Silva and H.R. Mello (ed.), *Confer ncias do V Congresso Internacional da Associa o Brasileira de Lingu stica*. Belo Horizonte: UFMG, 211-228.
- Danieli, M., J.M. Garrigo, M. Moneglia, A. Panizza, S. Quazza and M. Swerts. 2004. Evaluation of Consensus on the Annotation of Prosodic Breaks in the Romance Corpus of Spontaneous Speech C-ORAL-ROM. In M.T. Lino, M.F. Xavier, F. Ferreira, R. Costa and R. Silva (eds), *Proceedings of the 4th International Conference on Language Resources and Evaluation*. Paris: ELRA, 1513-16.
- t'Hart, J., A. Cohen and R. Collier. 1990. *A perceptual study on intonation: an experimental approach to speech melody*. Cambridge: Cambridge University Press.
- LABLITA. <http://lablita.dit.unifi.it/>
- MacWhinney, B.J. 2000. *The CHILDES project. Tools for analyzing talk*, 2 voll. Mahwah, NJ: Lawrence Erlbaum.
- Maia Rocha, B., T. Raso, .M.I. Andraide. In press. Alguns aux lios dial gicos em italiano, portugu s do Brasil e em italianos cultos em contato prolongado com o portugu s do Brasil. *Fragments*.
- Moneglia, M. 2000. Specifications on the C-ORAL-ROM Corpus. <http://lablita.dit.unifi.it/coralrom/papers/Specifications-CORALROM.pdf>.
- Moneglia, M. 2004. Measurements of Spoken Language Variability in a Multilingual Corpus. Predictable Aspects. In M.T. Lino, M.F. Xavier, F. Ferreira, R. Costa and R. Silva (eds), *Proceedings of the 4th International Conference on Language Resources and Evaluation*. Paris: ELRA, 1419-22.
- Moneglia, M. 2005. C -ORAL-ROM. Un corpus di riferimento del parlato spontaneo per l'italiano e le lingue romanze. In J. Korzen (ed.), *Lingua, cultura e intercultura. L'italiano e le altre lingue. Atti del VIII convegno SILFI (Copenhagen 22-26 July 2004)*. Copenhagen: Samfunzliteratur Press, 229-42.
- Moneglia, M. 2006. Units of Analysis of Spontaneous Speech and Speech Variation in a Cross-linguistic Perspective. In Y. Kawaguchi, S. Zaima and T. Takagaki (eds), *Spoken Language Corpus and Linguistics Informatics*. Amsterdam and Philadelphia: John Benjamins, 153-79.

- Moneglia, M. and E. Cresti. 1997. L'intonazione e i criteri di trascrizione del parlato adulto e infantile. In U. Bortolini and E. Pizzuto (eds), *Il Progetto CHILDES Italia*. Pisa: Del Cerro, 57-90.
- Moneglia, M., A. Scarano and M. Spinu. 2005. Validation by expert transcribers of the C-ORAL-ROM prosodic tagging criteria on Italian, Spanish and Portuguese corpora of spontaneous speech. In C. Nicolas and M. Moneglia (eds), *La gestione unitaria dell'eredità culturale multilingue europea e la sua diffusione in rete*. Firenze: Firenze University Press, 107-135.
- Raso, T., H. Mello, L. de Deus and A. Jesus. 2007. Uma aplicação da Teoria da Língua em Ato ao português do Brasil. *Revista de Estudos da Linguagem* 15, 2: 147-166.
- Raso, T. and A. Ulisses. 2008. Tópico e Apêndice no português do Brasil: algumas considerações. *Revista de estudos da linguagem* 16, 1: 247-262.
- Raso, T. and H. Mello. In press. As especificidades da unidade de tópico em PB e possíveis efeitos do contato lingüístico. In E. Saraiva and E. Chaves Marinho (eds), *Estudos da língua em uso: da gramática ao texto*.
Revista de Estudos da Linguagem. <http://relin.letras.ufmg.br/revista/>
- Tucci, I. 2005. L'espressione della modalit  nel parlato: i verbi modali nei corpora italiano e spagnolo C-ORAL-ROM. In I. Korzen (ed.), *Lingua, cultura e intercultura. Atti del VIII convegno internazionale della SILFI*. Copenhagen: Samfundslitteratur Press, 295-308.
- Tucci, I. 2008a. La modalizzazione lessicale nel parlato spontaneo. Dati dal corpus C-ORAL-ROM italiano. In E. Cresti (ed.), *Prospettive nello studio del lessico Italiano. Atti del IX Congresso internazionale della Societ  di Linguistica e Filologia Italiana*. Firenze: Firenze University Press, 377-86.
- Tucci, I. 2008b. La modalizzazione nel parlato spontaneo. Relazione tra espressioni lessicali della modalit  e unit  d'informazione. In M. Pettorino, A. Giannini, M. Vallone and R. Savy (eds), *La comunicazione parlata. Atti del Convegno Internazionale*. Napoli: Liguori, 447-464.
- Tucci, I. 2009. The scope of lexical modality and the informational structure in spoken Italian. In L. Mereu (ed.), *Information Structure and its Interfaces*. Berlin and NewYork: Mouton, 203-226.
- Tucci, I. In press. La Modalit  nel parlato spontaneo e il suo dominio di pertinenza. Una ricerca corpus-based (C-ORAL-ROM Italiano). In M. Iliescu, P. Danler and H. M. Siller (eds), *Actes du XXVe Congr s International de linguistique et de philologie romanes* (Innsbruck 3-8 septembre 2007). T bingen: Niemeyer.
- Tucci, I. and M. Moneglia. In press. Modality and illocutionary force in spoken Italian. In C. Push (ed.), *Proceedings of the 3rd Workshop "Corpora and Pragmatics"* (Freiburg 14-17 settembre 2006). T bingen: Gunter Narr Verlag.
- Ulisses, A.J. 2008. A unidade de Ap ndice no portugu s do Brasil. MA thesis, UFMG. WinPitch. <http://www.winpitch.com>
- Z gari, M.R.L. 2005. Os falares mineiros: esboço de um atlas lingüístico de Minas Gerais. In V. de Andrade Aguilera (ed.), *A Geolingüística no Brasil - trilhas seguidas, caminhos a percorrer*. Londrina: Editora da Universidade Estadual de Londrina, 45-72

STRUMENTI
PER LA DIDATTICA E LA RICERCA

1. Brunetto Chiarelli, Renzo Bigazzi, Luca Sineo (a cura di), *Alia: Antropologia di una comunità dell'entroterra siciliano*
2. Vincenzo Cavaliere, Dario Rosini, *Da amministratore a manager. Il dirigente pubblico nella gestione del personale: esperienze a confronto*
3. Carlo Biagini, *Information technology ed automazione del progetto*
4. Cosimo Chiarelli, Walter Pasini (a cura di), Paolo Mantegazza. *Medico, antropologo, viaggiatore*
5. Luca Solari, *Topics in Fluvial and Lagoon Morphodynamics*
6. Salvatore Cesario, Chiara Fredianelli, Alessandro Remorini, *Un pacchetto evidence based di tecniche cognitivo-comportamentali sui generis*
7. Marco Masseti, *Uomini e (non solo) topi. Gli animali domestici e la fauna antropocora*
8. Simone Margherini (a cura di), *BIL Bibliografia Informatizzata Leopardiana 1815-1999: manuale d'uso ver. 1.0*
9. Paolo Puma, *Disegno dell'architettura. Appunti per la didattica*
10. Antonio Calvani (a cura di), *Innovazione tecnologica e cambiamento dell'università. Verso l'università virtuale*
11. Leonardo Casini, Enrico Marone, Silvio Menghini, *La riforma della Politica Agricola Comunitaria e la filiera olivicolearia italiana*
12. Salvatore Cesario, *L'ultima a dover morire è la speranza. Tentativi di narrativa autobiografica e di "autobiografia assistita"*
13. Alessandro Bertirotti, *L'uomo, il suono e la musica*
14. Maria Antonietta Rovida, *Palazzi senesi tra '600 e '700. Modelli abitativi e architettura tra tradizione e innovazione*
15. Simone Guercini, Roberto Piovan, *Schemi di negoziato e tecniche di comunicazione per il tessile e abbigliamento*
16. Antonio Calvani, *Technological innovation and change in the university. Moving towards the Virtual University*
17. Paolo Emilio Pecorella, *Tell Barri/Kahat: la campagna del 2000. Relazione preliminare*
18. Marta Chevanne, *Appunti di Patologia Generale. Corso di laurea in Tecniche di Radiologia Medica per Immagini e Radioterapia*
19. Paolo Ventura, *Città e stazione ferroviaria*
20. Nicola Spinosi, *Critica sociale e individuazione*
21. Roberto Ventura (a cura di), *Dalla misurazione dei servizi alla customer satisfaction*
22. Dimitra Babalis (a cura di), *Ecological Design for an Effective Urban Regeneration*
23. Massimo Papini, Debora Tringali (a cura di), *Il pupazzo di garza. L'esperienza della malattia potenzialmente mortale nei bambini e negli adolescenti*
24. Manlio Marchetta, *La progettazione della città portuale. Sperimentazioni didattiche per una nuova Livorno*
25. Fabrizio F.V. Arrigoni, *Note su progetto e metropoli*
26. Leonardo Casini, Enrico Marone, Silvio Menghini, *OCM seminativi: tendenze evolutive e assetto territoriale*
27. Pecorella Paolo Emilio, Raffaella Pierobon Benoit, *Tell Barri/Kahat: la campagna del 2001. Relazione preliminare*
28. Nicola Spinosi, *Wir Kinder. La questione del potere nelle relazioni adulti/bambini*
29. Stefano Cordero di Montezemolo, *I profili finanziari delle società vinicole*
30. Luca Bagnoli, Maurizio Catalano, *Il bilancio sociale degli enti non profit: esperienze toscane*
31. Elena Rotelli, *Il capitolo della cattedrale di Firenze dalle origini al XV secolo*
32. Leonardo Trisciuzzi, Barbara Sandrucci, Tamara Zappaterra, *Il recupero del sé attraverso l'autobiografia*

33. Nicola Spinosi, *Invito alla psicologia sociale*
34. Raffaele Moschillo, *Laboratorio di disegno. Esercitazioni guidate al disegno di arredo*
35. Niccolò Bellanca, *Le emergenze umanitarie complesse. Un'introduzione*
36. Giovanni Allegretti, *Porto Alegre una biografia territoriale. Ricercando la qualità urbana a partire dal patrimonio sociale*
37. Riccardo Passeri, Leonardo Quagliotti, Christian Simoni, *Procedure concorsuali e governo dell'impresa artigiana in Toscana*
38. Nicola Spinosi, *Un soffitto viola. Psicoterapia, formazione, autobiografia*
39. Tommaso Urso, *Una biblioteca in divenire. La biblioteca della Facoltà di Lettere dalla penna all'elaboratore. Seconda edizione rivista e accresciuta*
40. Paolo Emilio Pecorella, Raffaella Pierobon Benoit, *Tell Barri/Kahat: la campagna del 2002. Relazione preliminare*
41. Antonio Pellicano, *Da Galileo Galilei a Cosimo Noferi: verso una nuova scienza. Un inedito trattato galileiano di architettura nella Firenze del 1650*
42. Aldo Burrelli (a cura di), *Il marketing della moda. Temi emergenti nel tessile-abbigliamento*
43. Curzio Cipriani, *Appunti di museologia naturalistica*
44. Fabrizio F.V. Arrigoni, *Incipit. Esercizi di composizione architettonica*
45. Roberta Gentile, Stefano Mancuso, Silvia Martelli, Simona Rizzitelli, *Il Giardino di Villa Corsini a Mezzomonte. Descrizione dello stato di fatto e proposta di restauro conservativo*
46. Arnaldo Nesti, Alba Scarpellini (a cura di), *Mondo democristiano, mondo cattolico nel secondo Novecento italiano*
47. Stefano Alessandri, *Sintesi e discussioni su temi di chimica generale*
48. Gianni Galeota (a cura di), *Traslocare, riaggregare, rifondare. Il caso della Biblioteca di Scienze Sociali dell'Università di Firenze*
49. Gianni Cavallina, *Nuove città antichi segni. Tre esperienze didattiche*
50. Bruno Zanon, *Tecnologia alimentare 1. La classe delle operazioni unitarie di disidratazione per la conservazione dei prodotti alimentari*
51. Gianfranco Martiello, *La tutela penale del capitale sociale nelle società per azioni*
52. Salvatore Cingari (a cura di), *Cultura democratica e istituzioni rappresentative. Due esempi a confronto: Italia e Romania*
53. Laura Leonardi (a cura di), *Il distretto delle donne*
54. Cristina Delogu (a cura di), *Tecnologia per il web learning. Realtà e scenari*
55. Luca Bagnoli (a cura di), *La lettura dei bilanci delle Organizzazioni di Volontariato toscane nel biennio 2004-2005*
56. Lorenzo Grifone Baglioni (a cura di), *Una generazione che cambia. Cioismo, solidarietà e nuove incertezze dei giovani della provincia di Firenze*
57. Monica Bolognesi, Laura Donati, Gabriella Granatiero, *Acque e territorio. Progetti e regole per la qualità dell'abitare*
58. Carlo Natali, Daniela Poli (a cura di), *Città e territori da vivere oggi e domani. Il contributo scientifico delle tesi di laurea*
59. Riccardo Passeri, *Valutazioni imprenditoriali per la successione nell'impresa familiare*
60. Brunetto Chiarelli, Alberto Simonetta, *Storia dei musei naturalistici fiorentini*
61. Gianfranco Bettin Lattes, Marco Bontempo (a cura di), *Generazione Erasmus? L'identità europea tra vissuto e istituzioni*
62. Paolo Emilio Pecorella, Raffaella Pierobon Benoit, *Tell Barri / Kahat. La campagna del 2003*
63. Fabrizio F.V. Arrigoni, *Il cervello delle passioni. Dieci tesi di Adolfo Natalini*
64. Saverio Pisaniello, *Esistenza minima. Stanze, spazi della mente, reliquiario*
65. Maria Antonietta Rovida (a cura di), *Fonti per la storia dell'architettura, della città, del territorio*
66. Ornella De Zordo, *Saggi di anglistica e americanistica. Temi e prospettive di ricerca*
67. Chiara Favilli, Maria Paola Monaco, *Materiali per lo studio del diritto antidiscriminatorio*
68. Paolo Emilio Pecorella, Raffaella Pierobon Benoit, *Tell Barri / Kahat. La campagna del 2004*

69. Emanuela Caldognetto Magno, Federica Cavicchio, *Aspetti emotivi e relazionali nell'e-learning*
70. Marco Masseti, *Uomini e (non solo) topi* (2ª edizione)
71. Giovanni Nerli, Marco Pierini, *Costruzione di macchine*
72. Lorenzo Viviani, *L'Europa dei partiti. Per una sociologia dei partiti politici nel processo di integrazione europea*
73. Teresa Crespellani, *Terremoto e ricerca. Un percorso scientifico condiviso per la caratterizzazione del comportamento sismico di alcuni depositi italiani*
74. Fabrizio F.V. Arrigoni, *Cava. Architettura in "ars marmoris"*
75. Ernesto Tavoletti, *Higher Education and Local Economic Development*
76. Carmelo Calabrò, *Liberalismo, democrazia, socialismo. L'itinerario di Carlo Rosselli (1917-1930)*
77. Luca Bagnoli, Massimo Cini (a cura di), *La cooperazione sociale nell'area metropolitana fiorentina. Una lettura dei bilanci d'esercizio delle cooperative sociali di Firenze, Pistoia e Prato nel quadriennio 2004-2007*
78. Lamberto Ippolito, *La villa del Novecento*
79. Cosimo Di Bari, *A passo di critica. Il modello di Media Education nell'opera di Umberto Eco*
80. Leonardo Chiesi (a cura di), *Identità sociale e territorio. Il Montalbano*
81. Piero Degl'Innocenti, *Cinquant'anni, cento chiese. L'edilizia di culto nelle diocesi di Firenze, Prato e Fiesole (1946-2000)*
82. Giancarlo Paba, Anna Lisa Pecoriello, Camilla Perrone, Francesca Rispoli, *Partecipazione in Toscana: interpretazioni e racconti*
83. Alberto Magnaghi, Sara Giacomozzi (a cura di), *Un fiume per il territorio. Indirizzi progettuali per il parco fluviale del Valdarno empolese*
84. Dino Costantini (a cura di), *Multiculturalismo alla francese?*
85. Alessandro Viviani (a cura di), *Firms and System Competitiveness in Italy*
86. Paolo Fabiani, *The Philosophy of the Imagination in Vico and Malebranche*
87. Carmelo Calabrò, *Liberalismo, democrazia, socialismo. L'itinerario di Carlo Rosselli*
88. David Fanfani (a cura di), *Pianificare tra città e campagna. Scenari, attori e progetti di nuova ruralità per il territorio di Prato*
89. Massimo Papini (a cura di), *L'ultima cura. I vissuti degli operatori in due reparti di oncologia pediatrica*
90. Raffaella Cerica, *Cultura Organizzativa e Performance economico-finanziarie*
91. Alessandra Lorini, Duccio Basosi (a cura di), *Cuba in the World, the World in Cuba*
92. Marco Goldoni, *La dottrina costituzionale di Sieyès*
93. Francesca Di Donato, *La scienza e la rete. L'uso pubblico della ragione nell'età del Web*
94. Serena Vicari Haddock, Marianna D'Ovidio, *Brand-building: the creative city. A critical look at current concepts and practices*
95. Ornella De Zordo (a cura di), *Saggi di Anglistica e Americanistica. Ricerche in corso*
96. Massimo Moneglia, Alessandro Panunzi (edited by), *Bootstrapping Information from Corpora in a Cross-Linguistic Perspective*
97. Alessandro Panunzi, *La variazione semantica del verbo essere nell'italiano parlato*

Finito di stampare presso
Grafiche Cappelli Srl - Osmannoro (FI)