SPEECH AND LINGUISTIC ANALYSIS
Properties of fronted direct object in Italian
Sandra AUGENDRE
UMR 5263 CLLE-ERSS(ERSSâB), Département de Sciences du Langage,
Université Michel de Montaigne Bordeaux 3, 33607 Pessac Cedex
augendre.sandra@wanadoo.fr

Abstract
The work presented in this paper focuses on a comparison of various occurrences of the same syntactic sequence in Italian: Object-Verb (OV). In this kind of utterance, the object occupies a “non canonical” position (preverbal position) and assumes the syntactic function of an object (no clitic is present). Classified among the so called “marked” (non canonical) structures in Italian grammars (cf. Grande Grammatica Italiana di Consultazione, 1988), OV order receives various names and descriptions from linguists. Based on a corpus of spontaneous productions, my study aims at reevaluating the properties attributed to OV order in Italian, for instance, the equivalence established between OV order, cleft sentence and narrow focus, the range of context possibilities for this structure or its pragmatic and prosodic characteristics.

Keywords: Italian; fronted direct object; syntax; pragmatics; prosody.

1. Introduction
The work presented in this article is based on a corpus constituted to study different object constructions in Italian, and focuses on a comparison of various occurrences of the same syntactic sequence in Italian: Object-Verb (OV). In this kind of utterance, the object occupies a “non canonical” position (preverbal position) and fully assumes the syntactic function of an object (no object clitic is present):

Example 1:

IL DOLCE ha mangiato.
THE CAKE he ate

'(It is) THE CAKE (that) he ate / He ate THE CAKE.'

Unlike a dislocated object, here the preverbal SN is strongly connected to predication: it assumes the function of an object and there is no coreferent expression in the utterance.

In this paper, we first will give an overview of most of the previous studies that have been carried on OV order in Italian. Then, we will describe the data we have worked on and our methodology. Finally, we will present our analysis and results.

2. OV order’s description
Classified among the so called “marked” (non canonical) structures in Italian grammars (cf. Grande Grammatica Italiana di Consultazione, 1988), OV order has not attracted much attention (cf. Berretta, 1998 and Brunetti, 2009 for two works based on corpora) and receives various names and descriptions from linguists.

In relation to the object initial position and the communicative status of the argument, the structure is often called rhematic (Stammerjohann, 1986) or contrastive (GGIC, 1988; Graffi, 1994; Ferrari, 2003) topicalization, left rhematisation (Berretta, 1998), focus-background structure (Brunetti, 2009), or more simply NP preposing (Abeillé, Godard & Sabio, 2008).

Retained as relatively infrequent in Italian by these linguists, OV order is described as limited to spoken dimension (Berretta, 1998; Brunetti, 2009), associated with a specific prosodic structure (peak of intensity on the object and fall of F0 after this argument, cf. Tamburini, 1998) and at a communicative level, the object is described as assuming a contrastive focus function (Sornicola, 1981).

This work aims at evaluating these correlations on occurrences present in spontaneous data.

3. Data and methodology
3.1 Corpora
Our corpora has been constituted in Sardinia and initially aimed at describing subject and direct object constituents in Italian utterances.

It is composed of spoken and written productions and divided in four parts: chat, e-mail, informal speech (spontaneous conversations) and formal speech (university lessons).

The entire corpora gathers 3000 utterances that contain a subject (realised by an independent element or in verb’s ending), eventually associated with a direct object (640 cases).

3.2 Data collected
In our corpus, we listed only 11 cases of fronted direct object, the result that confirms the very low rate of use previously attributed by the linguists to OV utterances in Italian. The general properties of our OV occurrences are the following:

- Only 3 of the 11 OV utterances come from the written corpus and 8 appear in spoken dimension. This repartition shows that this order is particularly related to prosody, that facilitates OV utterances’ interpretation even if it is also available in writing.
- All written OV utterances appear in chat, not in e-mails and all spoken OV utterances (except
one) appear in informal speech: these data indicates a close link between OV order, conversation and informality.
- Concerning the type of OV utterances, we have two (oral) interrogative structures and then exclamative ones.
- In 10 of the 11 OV utterances, the subject is not realised (utterance limited to O+V) and for the remaining case, the subject is postverbal (O+V+S).
- Finally, concerning the fronted objects, they are all directly followed by the verb (or are separated from it by clitics) and are short phrases (two words or less, except one case). Types of objects, divided in two classes, are the following:

A. NP (6 cases): l’ora/the hour, la finalità di parole/the finality of words, una torre/a tower, alcune parole/some words, un po’/a little and a proper name.
B. Proforms (5 cases): qualcosa/something, questo/this (three cases), qualcosa/something.

3.3 Structure of the analysis
Our analysis of OV utterances relies on three dimensions: syntactic (one specific syntactic structure: O+V), pragmatic (relation between OV and information structure) and prosodic (properties of OV utterances).

The analysis of OV utterances present in our corpus aims at showing if OV order in Italian has a specific domain of use or a given pragmatic value, more precisely, in which dimension(s) (spoken/written Italian) OV is represented, which communicative need(s) this structure responds to and which kind of prosodic structure it is associated with.

4. Analysis of OV utterances
The number of OV utterances available in the corpus confirms the weak degree of use of this order and the distribution of the occurrences proves that there is a close link between OV, conversation (8/11 occurrences appear in spoken dimension) and informality (10/11 occurrences are present in spontaneous data).

By analyzing OV utterances, we aim at defining the domain of use of this structure, its information structure (focus domain, type of focus...) and also at distinguishing different prosodic structures according to each OV utterance properties (object’s part of speech, type of referent, information structure, contextual data...).

4.1 Anaphoric vs non anaphoric fronted objects
Our analysis began with the classification of OV utterances according to the status of fronted objects’ referents, in order to verify the distinction established by Benincà (1988) and resumed by Berretta (1998) between left rhematisation and anaphoric anteposition. The fronted object can be anaphoric or not:

- In the first case, it is a coreferent expression related to an element present in linguistic or extralinguistic contexts (simple anaphora) or a global resolution of a part of the previous speech (recaptitulative anaphora). This type of OV utterance is analyzed by both linguists as cases of anaphoric anteposition because object’s referent is contextually given and because OV order is here motivated by the will to leave postverbal/focal position available for another element, which is often the subject. Among the 11 OV utterances present in the corpora, 5 objects are anaphoric expressions, like in the following example:

Example (2):
A: C’è anche questo che non ho capito
There is also this that I don’t understand
B: Questo non hai capito?
You don’t understand this [this (acc.) you don’t understand]?

- In the second case, the object is the element marked as the most prominent of an all focus utterance (emphasized object) or the element that constitutes the informative contribution of the utterance, that can be contrastive (contrastive focalisation) or not (completive focalisation). In this category, we find 6 of our 11 fronted objects, like in the following example, that represents a case of emphasized object in an all focus utterance:

Example (3):
Hanno fatto anche il lavoro di trascriptione // naturalmente non su tutto perché // un po’ facevano anche in classe // guidati dagli insegnanti
They also did the transcription work // naturally not on all because // they did a little in class [a little (acc.) they did in class] // helped by the teachers

4.2 Substitution test by a cleft or by a presentational sentence
For all OV utterances, we also put in relation object referent status and information structure of the utterance. We thus tried to replace OV sequences by a cleft sentence (è X che / it is X that/who) and by a presentational sentence (c’è X che / there is X that/who), in order to verify the presupposed status (substitution by a cleft sentence acceptable) or non presupposed status (substitution by a presentational sentence acceptable) of the object and of what follows it in the utterance.

The results of this test are presented in the tables below.
Anaphoric objects | Cleft/Presentational Test
--- | ---
*Questo non hai capito*<br>This you don’t understand | // cleft sentence
*Questo non riesco a capire*<br>This I don’t manage to understand | // presentational sentence
*L’ora non so*<br>The time I don’t know | // presentational sentence
*Questo vorrebbe dire*<br>This maybe it should mean | // cleft sentence
*Qualcosa mi ricordo*<br>Something I remember | // presentational sentence

Table 1: Anaphoric fronted objects’ substitution test

Non anaphoric objects | Cleft/Presentational Test
--- | ---
*Qualcosa evito di chiedere*<br>Something I avoid asking for | // presentational sentence
*Alcune parole non riusciva a leggere*<br>Some words she did not manage to read | // presentational sentence
*Un po’ facevano in classe*<br>A little they did in class | // presentational sentence
*Una torre avevo fatto io*<br>A tower I had made | // cleft sentence
*La finalità di parole vorrà dire*<br>The finality of words it should mean | // cleft sentence
*Usandra mi hai detto*<br>Usandra you told me | // cleft sentence

Table 2: Non anaphoric fronted objects’ substitution test

The substitution test allows us to show, on one hand, that contextual level and utterance level are relatively independent, and on the other hand, that the equivalence often established between OV order and the cleft sentence is only relative:

- Among anaphoric and non anaphoric objects, half (respectively 3 on 5 and 3 on 6) corresponds to a presentational sentence (wide focus) and half (respectively 2 on 5 and 3 on 6) to a cleft sentence (narrow focus). It is thus not possible to establish a clear relation between the status of fronted objects’ referents to one of the two types of focalisation (wide and narrow).
- Among the 11 OV utterances of the corpus, more than half (6 cases) are equivalent to a presentational sentence (the subordinate clause is not presupposed) and only 5 to a cleft sentence (the subordinate clause is presupposed), data that reveals that in OV utterances, what follows the object is not inevitably presupposed, but especially that this configuration (fronted object narrow focus) is even less frequent than the other one (wide focus).

4.3 More detailed analysis

After the presentation of all properties of our OV utterances, we will now concentrate on four representative examples and their analysis: a non focus anaphoric object (4), a fronted object in an all focus sentence (5), a fronted object focus (6) and a contrastive fronted object (7).

4.3.1. Anaphoric fronted object (5 cases)

In this first configuration, the object’s referent is introduced in the linguistic or extralinguistic context and is then referred to by a proform in preverbal position.

Example (4):

A: C’è anche *questo* che non ho capito
There is also *this* that I don’t understand
B: *Questo* non hai capito?
This you don’t understand
‘You don’t understand *this*?’
(Is it) [**This** (acc.) (that) you don’t understand]

In the example above, B’s utterance is the identical repetition of what A says (questo + negation + capire / this + negation + to understand) but as a question. The informative content of OV utterance does not come from the elements’ newness but only from the modality of the utterance (request of confirmation).

![Figure 1: prosodic structure of the utterance “questo non hai capito?”](image)

In Figure 1, we can observe that no considerable prominence is attributed to the preverbal proform (147 Hz, 51 dB and a duration of 267 ms for QUES(to)) and only the past participle, situated at the end of the question, is realised as prominent here (229 Hz and 52 dB on (ca)PI(to)).

4.3.2. All focus OV utterances (3 cases)

In this configuration, the fronted object is contextually new and represents the anchorpoint of a completely informative utterance.
Example (5):

A: Ho fatto qualcosa?
‘Do I help in something?’
B: Sì grazie
‘Yes thanks’
C: Alcune parole non riusciva a leggere
‘She didn’t succeed to read some words’
(There are) [some words (acc.) (that) she didn’t manage to read]

OV utterance aims here at closing a conversation by calling back the event which caused it: B and C asked A to read a document and C resumes in conclusion the cause of this recourse (they needed A because B did not manage to read some words).

If the utterance informs that B did not manage to read some words, it presents the object (alcune parole) as a major indication, thanks to the initial position of the object and to F0’s fall between it and its right context. In fact, at prosodic level, the preverbal SN is marked as the utterance most prominent element, unlike what we observed previously for anaphoric objects.

In terms of F0, the curve’s highest points correspond to the tonics of the adjective alcune/some and of the noun parole/words (192 Hz on (al)CU(ne) and 220 Hz on (pa)RO(le)). Furthermore, the melodic curve falls considerably from the tonic of the object phrase’s noun (from 220 Hz on (pa)RO(le) to 151 Hz on non). At intensity level, we also observe a fall from the noun: we have three peaks on the three syllables of the noun (50 dB, 49 dB and 50 dB) and then lower values until verb’s tonic (52 dB on LEG(gere)).

4.3.3. Non constrastive fronted object (2 cases)
In the third configuration, the object constitutes the informational and prominent part of the utterance without being implicated in a paradigmatic opposition, whereas its right context is totally secondary at communicative level.

4.3.4. Contrastive fronted object (1 case, written)
In the last configuration, object’s referent is introduced as both utterance’s informational contribution and as a paradigm member. This case (fronted object narrow focus introduced in opposition to one or more other referents) corresponds to the one globally presented as prototypic by the linguists (cf. part 2). However, among our 11 OV utterances, only one of them is contrastive.

Example (7):

A: L’albero con la carta igienica, eri tu?
‘The tree with the toilet paper, it was you?’
B: Albero??? Di carta igienica????
‘Tree??? Of toilet paper????’
B: No UNA TORRE avevo fatto io.
‘No it is a tower that I had made’
(1) [A TOWER (acc.) (that) I had made]

In this last example, the contrastive value of the fronted object is undeniable: to describe the same object,
A introduced the notion of tree and B replaced it by the concept of tower, kind of contrast called replacing focus by Dik (1997: 331-332); A says that B built a tree (assertion of to make a tree (B)) and B rejects part of A assertion by replacing object’s referent by another one (negation of to make a tree (B) and assertion of to make a tower (B)). In this unique OV utterance, the only referent both contextually new and informative is the fronted object, as the fact that A built something is already presupposed in the previous discourse. What follows the object is presupposed and the utterance is equivalent to a cleft sentence (no, è una torre che... / no, it is a tower that...).

Finally, besides a focalisation of the fronted object, the utterance also contains a postverbal pronoun (OVSpr), whose presence is pragmatically motivated: the pronoun is not realized as an informational contribution but strengthens the contrastive value of the utterance by creating a second paradigm (io / I vs. someone else), connected to the first one (albero / tree vs. torre / tower), but that remains implicit. The effect obtained with the realization of the pronoun in final position is similar to the the one proposed by Blasco-Dulbecco (1995: 59) for the sequences moi je in French: " the tonic pronoun [...] seems to aim essentially the naming of an element distinguished among the others of its sort; as if it expressed a kind of contrast or of instigation. This is the case not only for the dislocation before the verb [...] but also for the dislocation after the verb ". Indeed, in our example, the subject is introduced as a contrastive topic as its presence can be interpreted in the following way: to build a tower (me) involves to build a tree (not me / someone else).

5. Results and conclusions

To conclude, we will first sum up the properties of our corpus OV utterances and then the results of their analysis at pragmatic and prosodic levels.

Concerning the number and the distribution of OV utterances, our data confirms the weak degree of use of OV order (11 cases in the corpus) and the close link between OV order, conversation and informality. Indeed, the available occurrences are mostly present in speech dimension (2/3), rather conversational and informal.

Our fronted objects have the following formal properties: in terms of part of speech, we have 5 NP and 6 proforms and in terms of length, 10 of our fronted objects are short phrases (≤ 2 words).

In terms of information, we distinguished first two types of object’s referents: the anaphoric ones (5 cases) and the non anaphoric ones (6 cases). Among anaphoric fronted objects (a NP and 5 proforms), we isolated those that resume partially the previous speech and have only a single referent. Among non anaphoric fronted objects, we distinguished those present in all focus utterances (3 cases) and those that constitute the utterance informational contribution (3 cases).

Then, we tried to verify the link often established between OV order and focus-background information structure by using two substitution tests (OV / cleft sentence and OV / presentational sentence). These tests revealed that independently of the status of object’s referent in the discourse (activated or not), the preverbal object of most of OV utterances does not constitute alone the utterance assertion (substitution by a cleft sentence impossible), in other words what follows the object does not tend to be presupposed.

Furthermore, only one of our fronted objects is clearly a contrastive focus, data that shows that OV order is neither limited to a narrow contrastive focalisation.

To conclude, OV order does not seem to be reserved to narrow focalisation at all (5 cases on 11) nor to contrastive focalisation (1 case on 11), and is more often connected to the will to mark the argument as the most prominent of a wider informational contribution (6 cases on 11).

Finally, at prosodic level, we first saw, with the three OV utterances present in written productions, that OV order, even if mostly used in spoken productions, does not inevitably need the prosodic marks to be interpreted.

In terms of realisation, we observed no net break between fronted objects and their right context but distinguished different prosodic structures according to OV utterance properties: object’s part of speech and referential autonomy (proforms are less prominent than NPs), referent’s type (anaphoric referents are perfectly integrated to the predication and are prosodically less prominent than non anaphoric ones), information structure (objects narrow focus are more prominent than objects that are part of a bigger focus unit)... At least, we have a small decline of F0 curve after the object and at most we have a net break between the object (focus) and its right linguistic context (background information). Fronted object’s prominence is quite particularly marked at prosodic level when the object is the utterance focus: in these cases, prosodic structure clearly distinguishes the focus from the background, as all prominence marks are attributed to the first part of the utterance while the second part is pronounced as a sequence neither prominent nor informational (less audible, flat F0 curve and low values at F0, intensity and duration levels).

To conclude, our study allowed us to confirm the weak degree of productivity of OV order, but also to widen the use of OV order to written dimension or to observe some regularities concerning fronted objects’ formal properties (part of speech, length...). At pragmatic level, our data and its analysis led us to reconsider the equivalence established between OV order, cleft sentence and narrow focus, which is only relative according to our data and at the same time, to widen the range of contextual possibilities for the structure by distinguishing different information and prosodic structures that can be associated to OV order in Italian.

6. References


Song lyrics and speech: similarities, differences and multi-dimension analysis of song lyrics from 1940 to 2009

Patrícia BÉRTOLI-DUTRA
UFMG
Av. Antonio Carlos, 6667
Belo Horizonte – MG cep: 31.270.901
patbertolid@gmail.com

Abstract

This paper shows the results of a research aiming at finding convergence of song lyrics speech and colloquial speech (general English) in order to highlight its relevance as a source for linguistic investigation. The second research goal was to find the dimensions of linguistic variation present in Anglo-American popular music lyrics. The study was theoretically based on Corpus Linguistics and the language views supported by it. Convergence was found by contrasting individual words and tri-grams (a sequence of three words) from a study corpus of over one million song lyrics to the British National Corpus and the American National Corpus. The most frequent 500 words occur in the three corpora and only three out of the 500 most frequent trigrams in the study corpus do not occur in the other corpora – such specific sequences of words reflect musical repetitions. After that, by following Douglas Biber’s framework for a Multi-dimension analysis, we were able to find six linguistic dimensions and observe how those lyrics are close or different from each other according to their linguistic elements (parts of speech and semantics).

Keywords: Convergence; Corpus Linguistics; Multi-dimension Analysis; Song Lyrics.

1. Introduction

Seeing songs as a constant presence in people’s everyday lives we have to consider the fact that the words people sing are also markedly relevant to the way people speak. In that sense we should consider song lyrics relevance as a source for linguistic investigation. Therefore, the first goal of the research presented here was to detect convergence points between Anglo-American song lyrics speech and colloquial speech. In other words, by considering song lyrics as a form of speech, linguistic characteristics present in song lyrics were contrasted to general English in order to highlight their similarities.

The second goal was to follow Douglas Biber’s model for a Multi-dimension analysis (1988) aiming at finding dimensions of variations of Anglo-American popular song lyrics and how they could compare to the original dimensions found by Biber.

2. Research areas

Three different research fields comprise the theoretical framework of this study: 1) Studies about popular music and lyrics (Frith, 1993; Moore, 2003; Straw, 2003; Hall, 2006; Middleton, 2000; Starr & Waterman, 2007; Bértoli-Dutra, 2002); 2) Corpus Linguistics (Berber Sardinha, 2004a, 2004b; Halliday, 1991); and 3) Multidimensional Analysis (Berber Sardinha, 2004a, 2004b; Biber 1988; Kauffmann, 2005).

EFL teachers have long been using song lyrics mainly in order to either improve their learners listening skills or as a motivational asset for their classes. In fact, popular music is one of the few tools learners have to keep contact with English outside the classroom. Besides that, music also conveys social aspects as well as other aspects of the culture from where it was conceived. According to Frith (1993), music is connected to the identity of a people, “it isn’t a way of expressing ideas; it is a way of living them.” Thus, in a world that is getting more and more globalized exchanging music experiences is sharing identities (Hall, 2006), for music is the cultural means that best enables us to cross borders, to go where music can take us (Frith, 1993).

It is noticeable therefore that music, and most specifically its lyrics, should be used in the classroom in a more systematic way with all their linguistic information, their parts of speech and semantic aspects fully exploited. Hence, it shouldn’t be considered only for its poetical or pronunciation aspects. In fact, we argue here that lyrics are not poetry with music but closer to actual conversation.

We have to highlight that for this study we considered popular music in a very comprehensive way, as the one highly disseminated by the media, sharing the view proposed by Starr and Waterman (2007): “we use the term ‘popular music’ broadly, to indicate music that is mass-reproduced and disseminated via the mass media (...) and that typically draws upon a variety of preexisting musical traditions (...) in which various styles, audiences, and institutions interact in complex ways.”

Another important point taken into consideration for this study was the media categorization of music styles or genres. Even though we were looking at songs for their linguistics characteristics apart from their sound, it was expected that songs classified in a specific musical genre would also share the same linguistic characteristics. Among the most common musical genres present in popular music literature (Shuker, 1994; Brackett, 2000; Frith, Goodwin & Grossberg, 2003; Starr & Waterman, 2007) the following ones were present in our corpus: country (traditional country, country soul); pop (rhythm and blues); pop rock (rock pop; pop, alternative); rock (hard rock, rock, grunge, post-grunge, English rock, punk rock, heavy metal, blues rock, emo progressive); rock and roll; vocal pop (traditional pop music).

The theoretical touchstone of the whole research is Corpus Linguistics. It is an area that is based on...
collecting and exploiting corpora, or a set of textual linguistic data carefully collected, in order to serve as source for the study of a language or linguistic variety (Berber Sardinha, 2004a: 3). The main concept underpinning Corpus Linguistics is viewing language as a probabilistic system (Halliday, 1991; Sinclair, 1991), that is, although there are a number of possible choices and lexical combinations they do not occur the same way or with the same frequency, not even randomly. In fact, each language follows certain patterns of lexical combinations, which represent each particular genre; thus the more words are considered for an analysis the bigger the chances of finding low frequency words and combinations. (Berber Sardinha, 2004a).

Finally, Multi-dimension analysis was used because we aimed at finding dimensions of variations of song lyrics according to Douglas Biber’s model (1988), which presented a set of variation of General English. Biber’s study assumes the probabilistic and functional characteristics of language (Halliday, 1991) and that linguistic variation occurs according to the context (Berber Sardinha, 2004a; Halliday; Hasan, 1989; Halliday & Webster; 2002; Sinclair, 1991). It also predicts that texts should be analyzed not only taking into account one but several linguistic features so as to determine their variation across linguistic functions. In other words, Biber states that “textual relations’ among different kinds of texts” cannot “be defined unidimensionally” (1988: 20). The idea behind this methodology is to precisely quantify the frequency of each linguistic characteristic present in each text and compare every text to each other grouping them by the salience of characteristics.

In order to accomplish his goal, Biber used a corpus of 960 thousand words (mainly from the LOB-Corpus). The texts were tagged according to their parts of speech (POS). Each POS frequency was automatically calculated, normalized and submitted to statistical procedures of factorial analysis. Factorial procedure groups the most salient frequencies showing their medium, maximum, minimum and standard deviation scales. After that, the texts presenting the characteristics in each factor were checked for their relevance. It is important to highlight here that all the texts are present in all the dimensions, what makes them different in each dimension is the salience of the specific characteristics in each dimension.

Biber’s analysis found six different dimensions of variation of the English Language: 1) Involved versus Informational Production; 2) Narrative versus Non-Narrative Discourse; 3) Situation-Dependent versus Elaborated Reference; 4) Overt Expression of Argumentation; 5) Non-abstract versus Abstract Style; and 6) On-Line Informational Elaboration Marking Stance.

Next section of this paper depicts the steps followed by each part of the study.

3. Convergence study

The initial part of the study followed the principles of Corpus Linguistics (Berber Sardinha, 2004a; Bértoli Dutra, 2002; Hunston & Francis, 1999; Sinclair, 1991) first by describing the frequency of the words in the study corpus, then by describing the lexical-grammar patterns in the study corpus and finally by contrasting the patterns found in the study corpus with lexical-grammar patterns present in general English. A corpus of 1,078,882 words of song lyrics recorded originally in English by 30 different artists (American, British and Canadian) from different periods of time (from 1940’s with Frank Sinatra to 2009’s teen movies soundtracks, such as High School Musical and Hannah Montana).

After collecting the corpus, word lists were extracted and contrasted with word lists from the reference corpora BNC and the ANC1 (single words and trigrams). Single words were analyzed aiming at verifying how the most frequent words in each corpus would match. After normalizing their frequency in the three corpora (so that they would be comparable), a sample of the 5002 most frequent words in the study corpus was taken and manually contrasted to the other corpora.

Trigrams were analyzed considering they represent the best combination of words in use. According to Lafferty (Lafferty, Sleator & Templer, 1992), “a usage of a word is determined by the manner in which the word is linked to the right and to the left in a sentence”. The authors also point out that trigrams work so well for linguistic analysis “because they are firmly based on data” and because they “they reflect simultaneously syntax, semantics, and pragmatics of the domain question.”

As a result of the contrastive analysis we found that the most frequent single words in the study corpus are also relevantly frequent in the general English corpora, as we can see at Table 1 below presenting the 15 most frequent words in the study corpus and their frequency in the reference corpora.

After analyzing single words we were able to conclude that song lyrics present high frequency of personal pronouns such as “I” and “YOU” which suggests interpersonal discourse. Besides that, we also noticed an overuse of the following words, when contrasted to the reference corpora: “baby”; “one”; “love”; “no”; “like”; “do”; “can”; “got”; “if”; “up”; “time”; “never” and “see”.

A similar procedure was taken afterwards in order to analyze the trigrams. That is, from the 129.117 different trigrams extracted from the study corpus, 5,431,734 from BNC, 1,453,050 from ANC-spoken and

---

1 It was used the BNC World Edition, with 100 million words available online at http://www.natcorp.ox.ac.uk/corpus/ and the online version of the ANC, available at http://www.americannationalcorpus.org/ with 22 million words.

2 Bearing in mind the amount of data we considered the most frequent 500 singles words and 500 trigrams as a representative sample.
The factor extraction resulted three factors that were accounted for their grammatical and semantic aspects. Grammatically they show the following oppositions: (1) infinitive, gerund and modals versus nouns; (2) personal pronouns and possessives versus qualifiers; (3) verbs in the past versus verbs in the present. Semantically the factors show the predominance of (1) movement/time/speech/people/object; (2) markers of emotion and social acts; (3) markers of music manifestation. From the interpretation of the factors emerged the following dimensions: (a) argumentative versus informative; (b) interactive versus descriptive; (c) past narratives versus immediate context; (d) personal acts; (e) emotion and society; and (f) musical manifestation.

The investigation of song lyrics on the dimensional scale showed how singers and bands, musical styles and the decade of the recordings are closer or more distant to each other in linguistic terms. The most representative style, artist and period of time for each of the dimensions, grammar and semantics, are as follows: (a) Punk Pop, Simple Pahm, 2000’s; (b) Rock’n’roll, Madonna, 1940’s; (c) Country, Johnny Cash, 1970’s; (d) Surf Rock, Beach Boys, 1960’s; (e) Heavy Metal, Metallica, 1940’s; and (f) pop Vocal, Frank Sinatra, 1940’s.

5. Considerations

This study showed how close ordinary spoken and written English are to song lyrics speech. It also validated Biber’s model for the research of contrast of linguistic features in functional terms. However, the Multi-dimension Analysis methodology cannot be considered as the only possible means for linguistic analysis of song lyrics or any other form of speech. We were able to observe how songs are close or distant, similar or different according to their linguistic elements and not only according to their rhythm and musical style generally imposed by the media.

6. Acknowledgements

The author would like to thank Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil for the financial support.

7. References


Acknowledgements

The author would like to thank Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil for the financial support.

7. References


For a comprehensive view of results, refer to http://www.sapientia.pucsp.br/tde_busca/arquivo.php?codArquivo=10985

Table 1: Most frequent words in the study corpus compared to BNC and ANC

<table>
<thead>
<tr>
<th>WORD</th>
<th>Study Corpus</th>
<th>BNC</th>
<th>ANC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. THE</td>
<td>4.02</td>
<td>6.02</td>
<td>5.44</td>
</tr>
<tr>
<td>2. YOU</td>
<td>3.33</td>
<td>0.58</td>
<td>0.80</td>
</tr>
<tr>
<td>3. I</td>
<td>4.07</td>
<td>0.73</td>
<td>0.85</td>
</tr>
<tr>
<td>4. TO</td>
<td>2.36</td>
<td>2.58</td>
<td>2.60</td>
</tr>
<tr>
<td>5. AND</td>
<td>2.28</td>
<td>2.61</td>
<td>2.68</td>
</tr>
<tr>
<td>6. A</td>
<td>2.14</td>
<td>2.17</td>
<td>2.21</td>
</tr>
<tr>
<td>7. ME</td>
<td>1.59</td>
<td>0.13</td>
<td>0.15</td>
</tr>
<tr>
<td>8. MY</td>
<td>1.35</td>
<td>0.14</td>
<td>0.24</td>
</tr>
<tr>
<td>9. IN</td>
<td>1.29</td>
<td>1.93</td>
<td>1.84</td>
</tr>
<tr>
<td>10. IT</td>
<td>1.21</td>
<td>0.91</td>
<td>1.15</td>
</tr>
<tr>
<td>11. OF</td>
<td>1.17</td>
<td>3.03</td>
<td>2.73</td>
</tr>
<tr>
<td>12. YOUR</td>
<td>0.99</td>
<td>0.13</td>
<td>0.11</td>
</tr>
<tr>
<td>13. ON</td>
<td>0.91</td>
<td>0.72</td>
<td>0.63</td>
</tr>
<tr>
<td>14. THAT</td>
<td>0.87</td>
<td>1.04</td>
<td>0.76</td>
</tr>
<tr>
<td>15. ALL</td>
<td>0.80</td>
<td>0.27</td>
<td>0.23</td>
</tr>
</tbody>
</table>

The factor extraction resulted three factors that were accounted for their grammatical and semantic aspects. Grammatically they show the following oppositions: (1) infinitive, gerund and modals versus nouns; (2) personal pronouns and possessives versus qualifiers; (3) verbs in the past versus verbs in the present. Semantically the factors show the predominance of (1) movement/time/speech/people/object; (2) markers of emotion and social acts; (3) markers of music manifestation. From the interpretation of the factors emerged the following dimensions: (a) argumentative versus informative; (b) interactive versus descriptive; (c) past narratives versus immediate context; (d) personal acts; (e) emotion and society; and (f) musical manifestation.

The investigation of song lyrics on the dimensional scale showed how singers and bands, musical styles and the decade of the recordings are closer or more distant to each other in linguistic terms. The most representative style, artist and period of time for each of the dimensions, grammar and semantics, are as follows: (a) Punk Pop, Simple Pahm, 2000’s; (b) Rock’n’roll, Madonna, 1940’s; (c) Country, Johnny Cash, 1970’s; (d) Surf Rock, Beach Boys, 1960’s; (e) Heavy Metal, Metallica, 1940’s; and (f) pop Vocal, Frank Sinatra, 1940’s.

For a comprehensive view of results, refer to http://www.sapientia.pucsp.br/tde_busca/arquivo.php?codArquivo=10985

Table 1: Most frequent words in the study corpus compared to BNC and ANC

<table>
<thead>
<tr>
<th>WORD</th>
<th>Study Corpus</th>
<th>BNC</th>
<th>ANC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. THE</td>
<td>4.02</td>
<td>6.02</td>
<td>5.44</td>
</tr>
<tr>
<td>2. YOU</td>
<td>3.33</td>
<td>0.58</td>
<td>0.80</td>
</tr>
<tr>
<td>3. I</td>
<td>4.07</td>
<td>0.73</td>
<td>0.85</td>
</tr>
<tr>
<td>4. TO</td>
<td>2.36</td>
<td>2.58</td>
<td>2.60</td>
</tr>
<tr>
<td>5. AND</td>
<td>2.28</td>
<td>2.61</td>
<td>2.68</td>
</tr>
<tr>
<td>6. A</td>
<td>2.14</td>
<td>2.17</td>
<td>2.21</td>
</tr>
<tr>
<td>7. ME</td>
<td>1.59</td>
<td>0.13</td>
<td>0.15</td>
</tr>
<tr>
<td>8. MY</td>
<td>1.35</td>
<td>0.14</td>
<td>0.24</td>
</tr>
<tr>
<td>9. IN</td>
<td>1.29</td>
<td>1.93</td>
<td>1.84</td>
</tr>
<tr>
<td>10. IT</td>
<td>1.21</td>
<td>0.91</td>
<td>1.15</td>
</tr>
<tr>
<td>11. OF</td>
<td>1.17</td>
<td>3.03</td>
<td>2.73</td>
</tr>
<tr>
<td>12. YOUR</td>
<td>0.99</td>
<td>0.13</td>
<td>0.11</td>
</tr>
<tr>
<td>13. ON</td>
<td>0.91</td>
<td>0.72</td>
<td>0.63</td>
</tr>
<tr>
<td>14. THAT</td>
<td>0.87</td>
<td>1.04</td>
<td>0.76</td>
</tr>
<tr>
<td>15. ALL</td>
<td>0.80</td>
<td>0.27</td>
<td>0.23</td>
</tr>
</tbody>
</table>


The use of inflected infinitive in a spoken corpus

Fernanda CANEVER
Universidade de São Paulo (USP)
Av. Prof. Luciano Gualberto, 403 - Sala 16 - Cidade Universitária - 05508-010- São Paulo - SP
fernandacanever@gmail.com

Abstract

In light of the usage-based approach (Langacker, 1987, 2000; Bybee, 2006a, 2006b, 2010) and the theory of utterance selection proposed by Croft (2000), this study intends to contribute to the investigation of the continuous update of linguistic knowledge that occurs through language use. Building upon prior research done by Canever (2012), which quantified the usage of the inflected infinitive in a written corpus, the focus of this study is on the use of the inflected infinitive in Brazilian Portuguese in a spoken corpus, namely a sample of the corpus Nurc/SP. The results show the presence of inflected infinitive in some innovative constructions in the 1970s, suggesting that a quantitative study with the complete Nurc/SP corpus should be likewise revealing. It is also argued that more studies with large spoken corpora of Brazilian Portuguese are needed to confirm Canever’s hypothesis that the infinitive inflection has received a positive social value, which, reinforced by the stigmatized lack of verbal agreement in Brazil and associated with the high frequency of occurrence of the infinitive inflection in other syntactic contexts, would be causing the inflection to spread to new infinitive constructions.

Keywords: Spoken Corpus; Usage-based Theories; Language Change; Inflected Infinitive; Automatic Data Extraction.

1. Introduction

Traditionally language use has not been the focus of linguistic investigation. Structuralism and generative grammar have given high priority to the langue, claiming that the linguistic system is self-contained and autonomous from other cognitive abilities and social factors (Croft, 2000). As a result, phenomena related to the parole such as variation have been considered peripheral.

Yet, Bybee (2006b) points that the interest for speech has increased in the last decades, and many theoretical approaches now claim that language structure should not be isolated from language use. Cognitive linguistics, which Langacker (1987, 2000) defines as usage-based, is one of them. According to this framework, language structure emerges from language use through general cognitive capabilities of the human brain, not because of an endowment exclusively related to language. But seen as symbolic, language represents a human biological adaptation for interactive goals (Tomasello, 2003). Thus, the role of experience in shaping both our linguistic knowledge and our concepts is highly emphasized in cognitive approaches to language studies.

Moreover, advances in computational and corpus linguistics have facilitated studies with real data. This means that those interested in capturing the more dynamic nature of language are now able to investigate linguistic phenomena by analyzing naturally-occurring data, and this is the realm this study belongs to. In light of the usage-based approach (Langacker, 1987, 2000; Bybee, 2006a, 2006b, 2010) and the theory of utterance selection proposed by Croft (2000), the aim of this study is to contribute to the investigation of how language use constantly shapes speaker’s grammar by quantifying variation in speech. Building upon prior research done by Canever (2012), which quantified the usage of the inflected infinitive in a written corpus, this study focuses on the usage of inflected infinitive in a spoken corpus, namely Nurc/SP, as well as on the challenges involved in such a task.

2. Usage-based theories

Coined by Langacker (1987), a language is a “structured inventory of conventional linguistic units” (p. 494). To understand how this inventory is structured, it is important to consider that in actual instances of language use, referred to by Langacker as usage events, the language user has to relate his linguistic system to these events. Either in order to produce an utterance with an intended meaning or to interpret someone else’s utterance, the language user establishes a connection between the usage event and his inventory, trying to find a similar structure. In case a compatible structure is found, the schema instantiated in the utterance is taken to be conventional. When a good match is not possible, the schema instantiated is considered non-conventional.

According to Langacker, novel structures may gradually become conventional and be stored in our linguistic inventory depending on their frequency of occurrence. When a non-conventional structure gets into the system, it might be reinforced by frequent use or disappear due to non-use. What is crucial in this process is the cognitive ability of habit formation, which Langacker refers to as entrenchment: the more frequent an element is, the more entrenched it becomes. Repetition, thus, affects speakers’ linguistic knowledge, and plays an important role in the characterization of a structure as being conventional.
The fact that the concrete use of language structures in the daily life of a speech community results in the emergence of new linguistic patterns may initially appear chaotic. However, it is undeniable that language is stable to a great extent. Such stability – or convention\(^1\) – is what allows communication and all the other social-interactive goals involved in language use to be achieved.

Even though Langacker recognized the role of use in the shaping of linguistic structure, his work has not discussed why some utterances propagate while others disappear. Considering that when a novel structure emerges, its frequency of occurrence is low, Blythe & Croft (2012) state that all innovations are expected to disappear if only the frequency of occurrence is considered. For this reason, these authors claim that frequency alone cannot explain how novel structures may survive and even replace former conventional structures.

Croft (2000), who proposes a usage-based theory for language change that is directly connected to theories of language use such as the one developed by Clark (1996), claims that social factors need to be taken into account in the investigation of language change. In presenting his theory of utterance selection, which is based on Hull’s generalized theory of selection (Hull, 1988), Croft (2000) proposes that language change is an evolutionary process, which is a model of change by replication. In this model, the replicator is a token of linguistic structure, which he calls a lingueme; the interactor is the speaker who replicates linguemes in interacting with other speakers; the population is a speech community, that is, a population of interactors; and the environment is the social context of the speech event, its goals as well as the other members of the population.

Based on the hypothesis that language change emerges from language use, the author claims that linguistic convention is central to the process of change. While interacting, when speakers are conforming to convention, they are doing what Croft called normal replication. However, even though speakers try to conform to convention, they often end up violating it by using non-conventional devices. Such non-conformity to convention is called altered replication, and is the first step to change – innovation. Once variation is generated through altered replication, different variants are made available for speakers to use, so they need to select among them, and this is called differential replication. To Croft, language change consists of these two steps: innovation and propagation/selection.

After innovations occur, they might be propagated or not. When propagation takes place, it means a new convention is established. As defended by Croft (2000), propagation is a social process, since it occurs according to the social values assigned to the variants, such as prestige, for example. However, in order to perpetuate, the cognitive structures on which linguistic utterances depend need to be entrenched in the speaker’s grammar.

The correlation between the degree of entrenchment and the social values assigned to linguistic variants in guiding language change posited by Croft seems to be the most appropriate way of approaching the issue, and therefore this idea underlies this investigation. Furthermore, since frequency of occurrence is crucial to determining the degree of entrenchment of linguistic constructions in speaker’s grammars, frequency studies are presumed to play a vital role in the investigation of natural languages.

3. **The Portuguese inflected infinitive**

According to Maurer (1968), the inflection of the infinitive has been documented since the first Portuguese documents, and has gradually spread to different constructions. Nowadays, the inflection is considered optional in numerous contexts, as in:

\[
\begin{align*}
(1) & \text{Estudamos para vencermos na vida.} \\
& \text{study.1PL to succeed.INF.1PL in life} \\
& \text{We study to succeed in life.} \\
(2) & \text{Estudamos para vencer na vida.} \\
& \text{study.1PL to succeed.INF in life} \\
& \text{We study to succeed in life.}
\end{align*}
\]

Bechara (2009), for instance, states that the infinitive inflection is used when the speaker intends to emphasize the grammatical person, as shown in (1), and the uninflected form is used when the emphasis is on the action, as shown in (2).

Recently, though, examples\(^2\) of the inflection of the infinitive in contexts where it is considered hypercorrection have been attested in spoken language, as in:

\[
\begin{align*}
(3) & \text{Viemos para SP para podermos lançarmos …} \\
& \text{came.1PL to SP to can.INF.1PL launch.INF.1PL} \\
& \text{We came to SP to be able to launch …} \\
(4) & \text{Nós temos que nos prepararmos…} \\
& \text{we have.1PL that REFL.1PL prepare.INF.1PL} \\
& \text{We need to prepare ourselves …}
\end{align*}
\]

Interested in infinitive constructions with optional inflection as well as in some more innovative contexts for the infinitive inflection, such as those illustrated by examples (3) and (4), Canever (2012) quantified the variation in a corpus of standard written language, more specifically a corpus of academic written Brazilian Portuguese that contained 11,000,000 words. The results

---

\(^1\) Reformulating Lewis (1969 in Clark 1996: 71). Clark defines convention as a partly arbitrary regularity in behavior that is common ground in a given community, but even though it is stable, it is not static (Croft, 2000: 132).

\(^2\) The examples (3) and (4) were collected by members of the LLIC/USP (http://www.linguistica.fflch.usp.br/llic), while the examples (5) to (9) were taken from Canever (2012). Because of space limitations, only excerpts of the examples are presented here.
reveal a high frequency of occurrence of the inflected infinitive, mainly in causal, final and temporal clauses, such as in:

(5) Tarefa que não podemos recusar, especialmente task that not can.1PL. refuse mainly para entendermos a falta de ... to understand.INF.1PL. the lack of
A task we cannot refuse, mainly in order to understand the lack of...

In constructions such as modal and aspect periphrases with an infinitive, Canever showed there is no preference for the inflection, as in:

(6) Podemos levantar a seguinte hipótese ... can.1PL. suggest.INF the following hypothesis
We can suggest the following hypothesis...

(7) As mulheres começam a ser felizes ... the women start to be.INF happy.PL
Women start to be happy ...

However, a few occurrences of inflected infinitive were found in those constructions, such as in:

(8) Não poderiam serem esquecidas ... not could.PL. be.INF.3PL. forgot.PL
Couldn’t be forgotten ...

(9) As virtudes começam a serem tratadas ... the virtues start.3PL. to be.3PL.INF treated.PL
The virtues start to be treated ...

Given the occurrence of such hypercorrect infinitive inflections in a written corpus of standard Portuguese, Canever claims that a positive social value might have been attributed to the inflected forms. Canever states that this positive value, reinforced by the stigma associated with the lack of verbal agreement in Brazil, and the high frequency of occurrence of infinitive inflection in other syntactic contexts could – together – be causing the inflection to spread to new infinitive constructions.

Although the results found by Canever suggest that in many constructions the inflected forms are highly entrenched in the grammars of the investigated speakers, further quantitative studies with spoken corpora are necessary to validate the hypothesis that the inflected infinitive is spreading in standard Brazilian Portuguese.

4. Quantification in a spoken corpus

4.1 Methods

4.1.1. Corpus
The spoken corpus used for this study was a sample of formal utterances – lectures, conferences, etc. – collected by the NURC project3 in São Paulo, Brazil. The sample, with approximately 30,000 words, consists of utterances produced by six participants, and has been published in a book (Castilho & Preti, 1986).

4.1.2. Data extraction
Because the original files were in .pdf format, they had to be converted to .txt format so the data extraction could be automatically done with the software R. In order to extract the occurrences of the infinitive inflection, a script containing the function exact.matches was used4. The script basically made R look for all the occurrences of words that ended either in –rmos or –rem, which are the infinitive plural inflections, and return the matches with some preceding and subsequent contexts. The output file was then handled in a spreadsheet program.

4.2 Results
Among the occurrences of infinitive inflection found, 20 were occurrences of the Third Person Plural (3PL) inflection –rem. Most of them occurred in contexts where a plural subject precedes the infinitive, such as in:

(10) (…) que levam as pessoas a demandarem … that lead.3PL. the people to demand.INF.3PL
(…) that lead people to demand ...

As for the inflection of First Person Plural (1PL) – rmos –, 8 occurrences were found, one of them being:

(11) Nós podemos utilizarmos desta reflexão … we can.1PL. use.INF.1PL. of. this reflection
We can use this reflection ...

4.3 Discussion
Given the small size of the sample, not many results were found. However, the quantification yielded some interesting results. The occurrence of an infinitive inflection after a modal verb such as in (11), for instance, suggests that the inflection of the infinitive in constructions such as modal periphrases, which Canever (2012) considered innovative and hypercorrect usage, already occurred in spoken language in the 1970s.

5. Conclusion and future directions
This study quantified the usage of inflected infinitive in a sample of the spoken corpus (Nurc/SP) in order to contribute to the investigation of how usage is constantly

3 NURC stands for Norma Urbana Culta (urban spoken standard language), and this project consisted of the investigation of spoken Portuguese in five state capitals in Brazil: São Paulo, Rio de Janeiro, Recife, Salvador and Porto Alegre in the 1970s.
4 The script can be found in Canever (2012), and the function the function exact.matches, developed by professor Stefan Th. Gries (University of California Santa Barbara), is available at: <http://www.linguistics.ucsb.edu/faculty/stgries/exact_matches.r>.
shaping our linguistic knowledge. The results found are revealing and suggest that a quantitative study with the complete Nurc/SP corpus should be likewise relevant to the investigation of the spread of the inflected infinitive in Brazilian Portuguese.

In order to do to that, some methodological challenges will have to be dealt with, though. First of all, it is crucial that the corpus Nurc/SP be in a machine-readable format, ideally in a format that is compatible with software such as R. Once this is done, it will be important to decide what annotation should be kept, as well as what kind of cleaning will be necessary, mainly because some speech annotation might be a problem in data extraction.

To support Canever (2012)'s hypothesis that the inflected infinitive is spreading in Brazilian Portuguese not only because of its high frequency of occurrence in optional contexts, but also because the inflection has received a positive social value, the use of the inflected infinitive needs to be quantified in different spoken corpora. For this reason, after the study with the whole Nurc/Sp corpus is ready, it will be also important to contrast its results with data obtained from more contemporary spoken corpora of Portuguese.

Given the lack of large spoken electronic corpora of Contemporary Brazilian Portuguese, a solution might be to work with different corpora formed by different research groups in Brazil.

6. Acknowledgements

I thank Professors Stefan Th. Gries, William Croft, Richard Blythe, Suzanne Kemmer, Michael Barlow and Kathrin Campbell-Kibler for their valuable help and suggestions during the 2011 LSA Linguistic Institute at University of Colorado at Boulder. I am equally grateful to Professor Evani de Carvalho Viotti for her inspiring guidance and encouragement throughout the course of this study.

This research was funded by CNPq (Grant 134950/2009-7).

7. References

A corpus-based analysis for superlative construction of body expression

Igor de Oliveira COSTA, Neusa Salim MIRANDA
Federal University at Juiz de Fora
Juiz de Fora, Minas Gerais, Brazil
igorsabo@yahoo.com.br, neusasalim@oi.com.br

Abstract
This work focuses on the corpus dimension of the Superlative Construction of Body Expression (“[…] solteirona e toda virgem, ignorava machezas, quase morreu de vergonha numa tarde de conversas”; “Padre Dito quase estourou de rir […]”; “O Lúcio rolou de rir com a explicação, e como consequência acabou virando a vítima e a cobaia do seminário.”), a major link in the network of constructions of Portuguese named by Miranda (2008a) as Superlative Constructions. The theoretical approach involves the Cognitive Linguistics and the Cognitive Construction Grammar. The corpus used is the Corpus do Português (http://www.corpusdoportugues.org/), composed of forty-five million words of fifty-seven thousand texts of the XIV-XX centuries. The results points, among other things, to the productivity of the construction under investigation, which instantiate, in the corpus investigated, 19 different types, and its conventionalization, outlined by the presence of 1.726 tokens, that corresponds to 43.9% of the usage of the searched verbs followed by the genitive preposition “de” in the corpus (3.929). The advantage in adopting a corpus based approach on the constructions’ investigation is also highlighted, once it offers access to the comprehension of the construction’s productivity and conventionalization in a language.

Keywords: cognitive linguistics; cognitive construction grammar; corpus-based approach; intensity; superlative constructions.

1. Introduction
The notion of degree is very rich to the grammar of languages. It is through scalar constructions that the language users denote the degree that speakers/writers can experience or believe they have experienced, among other things.

There are many structures in the Portuguese language (as in other languages) that serve this purpose of intensifying a statement. But against what speakers/writers use, the Grammatical Tradition and even Linguistic Tradition, little or almost nothing, devoted to the study of this phenomenon. Some examples of degree modifier constructions present, for example, in normative grammars of Portuguese are: Comparative Constructions (“Ele é tão rápido quanto o Bolt”;“Eu escrevo melhor/pior do que ele”); Construction with Adverbs of Intensity (“Maria Fernanda Cândido é perfeita demais”); (“Maria Fernanda Candido is too perfect”), pleonastic expressions (“Que jogada linda, linda, linda!”); What a pretty, pretty move!”).

In order to fill this gap, the present work, along with others, aims to expand the study of the manifestations of degree in Portuguese Language, as a way to contribute to a fuller description of the language. In this work, the object under investigation is the Superlative Construction of Body Expression (SCBE)

1) 19:Fic:Br:Cony:Piano Enquanto o sábado não chegasse, ele podia se fartar de ouvir todos os discos que quisesse […]

2) 19Or:Br:Intrv:ISP […] o meu clown não consegue cruzar os braços. A platéia morre de rir do que é, na verdade, uma tragédia para o meu personagem. “[…] my clown cannot cross his arms. The audience die of laughing about what is, indeed, a tragedy for my character.” (to die of laughing = to die laughing = to laugh too much)

3) 19:Fic:Br:Garcia:Silencio […] queria era apenas assustar, podemos telefonar para ele e dizer que eu estou me borrando de medo. “[…] s/he just want to scare, we can call him and say I am shitting of fear.” (to shit of fear = to scared shitless = to be very much afraid)

Because it is a very broad research (which, in addition to the formal description and semantic-pragmatic motivations, involves its conceptual motivation, its inheritance relations, its process of grammaticalization, among other issues2), this work cuts out the part of the SCBE study that is more directly related to the use of corpora.

This research is linked to the “Superlative Constructions of Brazilian Portuguese: a study about scale semantic” (Miranda, 2008 – CNPq), which, from its genesis to now, elucidated, with the study of the SCBE, seven nodes of this large network of constructions. Four other studies are still in progress.

The paper is organized as follows: the first section presents the theoretical perspective through which we develop our object; the following section discusses the research methodology chosen and the process of data

---

1 All the English “versions” of the examples and SCBE types are just an attempting to clarify the phenomenon being studying, presenting the semantic nature of words that compose the construction.

2 Costa (2010) covers most of these points.
collection; section 3, in turn, will bring the analyzes of SCBE, which involves the use of corpus; after that, we presented our conclusion, followed by the acknowledgments and the references.

2. Theoretical Bases

The theoretical framework of this study is composed of Cognitive Linguistics (Fauconnier, 1994; Fauconnier & Turner, 2002; Fillmore, 1982; Johnson, 1987; Lakoff, 1987; Lakoff & Johnson, 2002[1980], 1999; Miranda, 2002, 2008a, 2008b; Salomão, 1997, 2006; among others) and one of its models of grammar, the Cognitive Construction Grammar (Goldberg, 1995, 2006; Boas, in press).

The cognitive research program of language emerged at the end of the seventies last century, and strongly opposes to the Generative Grammar and Truth-conditions semantics. In general, Cognitive Linguistics considers (1) language as a non-autonomous cognitive faculty, governed by general cognitive apparatus; (2) advocates a central role for imaginative processes (metaphor, metonymy, blending) in human cognition and language; (3) sees grammar as conceptualization, as a way to profile a human scene; and (4) allows that the knowledge of language emerges from its use (Croft & Cruse, 2004: 1-4).

The Cognitive Construction Grammar (CCxG) (Goldberg, 1995, 2006; Boas, in press), defining constructs as pairs of form and function, gives these structures the status of basic units of language. Thus, the grammar and lexicon are defined as a network of constructions established by the use through culture. The description of such structures, therefore, is realized investigating not only their formal patterns, but also their dimensions of meaning and use.

A key point for the Goldberian model of grammar is the frequency of type and frequency of token variables, responsible respectively for the entrenchment of certain constructional pattern in the minds of speakers of a language and the conventionalization of a construction in a given language (that is, the capacity of a construction to be extended to new cases within the language). Once a corpus allows the verification of such data, the use of this tool in a study of an object like the one being investigated here is highly profitable and productive.

As a model of grammar fully immersed in the assumptions of Cognitive Linguistics, CCxG aims to provide psychologically plausible explanations for the language (Croft & Cruse, 2004: 272; Boas, in press: 12.) exploring the motivation and inheritance relations among constructions.

3. Methodology

Due to the importance of the use in the theoretical model adopted (CCxG is a use-based model of language, cf. Croft & Cruse, 2004: 291-327), we make use of a corpus-based approach (Aluísio & Almeida, 2006; Divjak & Gries, 2003; Sardinha, 2004; Stefanowitsch, 2006) in the investigation of the object.

The assembly of a database specifically for cases involving the SCBE is the first (and crucial) step in the study of a construction, because it is a way of letting the data speak, and not be hostage solely to our intuitions. Therefore, in order to be faithful to it, the search for cases of the construction was divided into two different phases: one in which we use different sources to get the most different types of the construction and another in which we make use of an annotated corpus for systematic study of the construction.

<table>
<thead>
<tr>
<th>Constructional types&lt;sup&gt;3&lt;/sup&gt;</th>
<th>CP</th>
<th>CE</th>
<th>Abril .com</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>01 acabar(-se) de rir</td>
<td>---</td>
<td>---</td>
<td>09</td>
<td>09</td>
</tr>
<tr>
<td>02 borrar(-se) de rir</td>
<td>01</td>
<td>---</td>
<td>---</td>
<td>01</td>
</tr>
<tr>
<td>03 cagar(-se) de rir</td>
<td>---</td>
<td>---</td>
<td>01</td>
<td>01</td>
</tr>
<tr>
<td>04 cair de rir</td>
<td>---</td>
<td>---</td>
<td>01</td>
<td>01</td>
</tr>
<tr>
<td>05 cansar(-se) de rir</td>
<td>01</td>
<td>02</td>
<td>---</td>
<td>03</td>
</tr>
<tr>
<td>06 chorar de rir</td>
<td>01</td>
<td>---</td>
<td>03</td>
<td>04</td>
</tr>
<tr>
<td>07 contorcer(-se) de rir</td>
<td>---</td>
<td>01</td>
<td>01</td>
<td>02</td>
</tr>
<tr>
<td>08 dobrar(-se) de rir</td>
<td>---</td>
<td>---</td>
<td>03</td>
<td>03</td>
</tr>
<tr>
<td>09 engasgar(-se) de rir</td>
<td>---</td>
<td>01</td>
<td>---</td>
<td>01</td>
</tr>
<tr>
<td>10 esbaladar(-se) de rir</td>
<td>---</td>
<td>---</td>
<td>01</td>
<td>01</td>
</tr>
<tr>
<td>11 esbrarchar(-se) de rir</td>
<td>---</td>
<td>---</td>
<td>01</td>
<td>01</td>
</tr>
<tr>
<td>12 escangalhar(-se) de rir</td>
<td>---</td>
<td>---</td>
<td>09</td>
<td>09</td>
</tr>
<tr>
<td>13 escrachar(-se) de rir</td>
<td>---</td>
<td>---</td>
<td>01</td>
<td>01</td>
</tr>
<tr>
<td>14 esganchar(-se) de rir</td>
<td>---</td>
<td>---</td>
<td>01</td>
<td>01</td>
</tr>
<tr>
<td>15 espemrer(-se) de rir</td>
<td>---</td>
<td>01</td>
<td>---</td>
<td>01</td>
</tr>
<tr>
<td>16 estourar(-se) de rir</td>
<td>01</td>
<td>---</td>
<td>---</td>
<td>01</td>
</tr>
<tr>
<td>17 farter(-se) de rir</td>
<td>10</td>
<td>19</td>
<td>---</td>
<td>29</td>
</tr>
<tr>
<td>18 finar(-se) de rir</td>
<td>01</td>
<td>---</td>
<td>---</td>
<td>01</td>
</tr>
<tr>
<td>19 mijar(-se) de rir</td>
<td>---</td>
<td>01</td>
<td>01</td>
<td>02</td>
</tr>
<tr>
<td>20 morrer de rir</td>
<td>14</td>
<td>20</td>
<td>185</td>
<td>219</td>
</tr>
<tr>
<td>21 não (se) aguentar de rir</td>
<td>---</td>
<td>---</td>
<td>01</td>
<td>01</td>
</tr>
<tr>
<td>22 passar mal de rir</td>
<td>---</td>
<td>---</td>
<td>02</td>
<td>02</td>
</tr>
<tr>
<td>23 rachar(-se) de rir</td>
<td>---</td>
<td>---</td>
<td>08</td>
<td>08</td>
</tr>
<tr>
<td>24 rasgar(-se) de rir</td>
<td>---</td>
<td>---</td>
<td>01</td>
<td>01</td>
</tr>
<tr>
<td>25 rebentar(-se) de rir</td>
<td>01</td>
<td>---</td>
<td>---</td>
<td>01</td>
</tr>
<tr>
<td>26 rolar de rir</td>
<td>---</td>
<td>08</td>
<td>52</td>
<td>60</td>
</tr>
<tr>
<td>27 torcer(-se) de rir</td>
<td>---</td>
<td>---</td>
<td>01</td>
<td>01</td>
</tr>
</tbody>
</table>

| TOTAL                           | 30 | 53 | 282     | 365   |

Table 1: SCBE Types

---

3 The particle “se” presented between parentheses is a Portuguese reflexive pronoun demanded by one of the uses of some verbs in the construction.
First phase: having the results of Sampaio (2007) – which point “rir” (“laughing”) as the most frequent Y element to the pattern ‘X DE Y’ (“chorar de rir”; “to cry of laughing”, “fartar-se de rir”; “to glut of laughing”, “morrer de rir”; “to die of laughing”, etc.) – as the start point, first we searched for the expression “de rir” in three different language database (the Corpus do Português, the Corpus Eye of the VISL project, and Abril.com) as a way to raise X elements of the constructional pattern being investigated. The initial hypothesis was that, starting from a most common form and therefore more conventional, it was possible to obtain wide and significant combinations of the variables which compose the construction. In fact, our hypothesis was confirmed. Table 1, below, shows the types collected in the searches.

<table>
<thead>
<tr>
<th>SCBE type</th>
<th>Results of the search</th>
<th>Tokens of SCBE</th>
<th>Productivity of the search</th>
</tr>
</thead>
<tbody>
<tr>
<td>acabar(-se) de Y</td>
<td>252</td>
<td>08</td>
<td>3.2%</td>
</tr>
<tr>
<td>borrar(-se) de Y</td>
<td>08</td>
<td>04</td>
<td>50%</td>
</tr>
<tr>
<td>cagar(-se) de Y</td>
<td>03</td>
<td>02</td>
<td>66.7%</td>
</tr>
<tr>
<td>cair de Y</td>
<td>835</td>
<td>96</td>
<td>11.5%</td>
</tr>
<tr>
<td>cansar(-se) de Y</td>
<td>437</td>
<td>372</td>
<td>85.1%</td>
</tr>
<tr>
<td>chorar(-se) de Y</td>
<td>196</td>
<td>112</td>
<td>57.1%</td>
</tr>
<tr>
<td>contorcer(-se) de Y</td>
<td>06</td>
<td>01</td>
<td>16.7%</td>
</tr>
<tr>
<td>dobrar(-se) de Y</td>
<td>75</td>
<td>01</td>
<td>1.3%</td>
</tr>
<tr>
<td>engasgar(-se) de Y</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>esbadar(-se) de Y</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>esboorchar(-se) de Y</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>escangalhar(-se) de Y</td>
<td>01</td>
<td>01</td>
<td>100%</td>
</tr>
<tr>
<td>escrachar(-se) de Y</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>escançar(-se) de Y</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>espremer(-se) de Y</td>
<td>06</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>estourar(-se) de Y</td>
<td>27</td>
<td>17</td>
<td>63%</td>
</tr>
<tr>
<td>fartar(-se) de Y</td>
<td>401</td>
<td>381</td>
<td>95%</td>
</tr>
<tr>
<td>fimar(-se) de Y</td>
<td>18</td>
<td>05</td>
<td>27.8%</td>
</tr>
<tr>
<td>mujar(-se) de Y</td>
<td>02</td>
<td>01</td>
<td>50%</td>
</tr>
<tr>
<td>morrer de Y</td>
<td>1,486</td>
<td>674</td>
<td>45.4%</td>
</tr>
<tr>
<td>não (se) aguentar de Y</td>
<td>01</td>
<td>01</td>
<td>100%</td>
</tr>
<tr>
<td>passar mal de Y</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>rachar(-se) de Y</td>
<td>18</td>
<td>01</td>
<td>5.6%</td>
</tr>
<tr>
<td>rasgar(-se) de Y</td>
<td>46</td>
<td>05</td>
<td>10.9%</td>
</tr>
<tr>
<td>rebentar(-se) de Y</td>
<td>52</td>
<td>34</td>
<td>65.4%</td>
</tr>
<tr>
<td>rolar de Y</td>
<td>29</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>torcer(-se) de Y</td>
<td>30</td>
<td>10</td>
<td>33.3%</td>
</tr>
</tbody>
</table>

TOTAL: 3,929, 1,726, 43.9%

Table 2: Data obtained in the second phase of the study

<table>
<thead>
<tr>
<th>SCBE Types</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>morrer de Y</td>
<td>674</td>
</tr>
<tr>
<td>fartar(-se) de Y</td>
<td>381</td>
</tr>
<tr>
<td>cansar(-se) de Y</td>
<td>372</td>
</tr>
<tr>
<td>chorar de Y</td>
<td>112</td>
</tr>
<tr>
<td>cair de Y</td>
<td>96</td>
</tr>
<tr>
<td>rebentar(-se) de Y</td>
<td>34</td>
</tr>
<tr>
<td>estourar(-se) de Y</td>
<td>17</td>
</tr>
<tr>
<td>torcer(-se) de Y</td>
<td>10</td>
</tr>
<tr>
<td>acabar(-se) de Y</td>
<td>08</td>
</tr>
<tr>
<td>fimar(-se) de Y</td>
<td>05</td>
</tr>
<tr>
<td>rasgar(-se) de Y</td>
<td>05</td>
</tr>
<tr>
<td>borrar(-se) de Y</td>
<td>04</td>
</tr>
<tr>
<td>cagar(-se) de Y</td>
<td>02</td>
</tr>
<tr>
<td>mujar(-se) de Y</td>
<td>01</td>
</tr>
<tr>
<td>escangalhar(-se) de Y</td>
<td>01</td>
</tr>
<tr>
<td>contorcer(-se) de Y</td>
<td>01</td>
</tr>
<tr>
<td>dobrar(-se) de Y</td>
<td>01</td>
</tr>
<tr>
<td>não (se) aguentar de Y</td>
<td>01</td>
</tr>
<tr>
<td>rachar(-se) de Y</td>
<td>01</td>
</tr>
</tbody>
</table>

TOTAL: 1,726

Table 3: Conventionalization of SCBE types in Corpus do Português

According to the occurrence of SCBE in the corpus, it was possible to more precisely understand the form of construction:

\[ X_Y \cdot Y_N Y \]

where X is filled with verbs that evoke the conceptual domains of physical impact (“acabar”) = “to finish”,

4. Analysis

In the description and explanation of SCBE, some findings are more strongly linked to the adoption of corpus research. As explained to the introduction, these findings are topics of the next lines.

In view of the data obtained from the corpus, the SCBE appears as a very productive construction, instantiating 19 different types in the corpus investigated. The construction can also be considered conventionalized since 1,726 tokens of the construction were found in Corpus do Português. This corresponds to 43.9% of the use of the 19 verbs followed by the preposition “de” in the corpus (3,929).

There is, however, a variation in the conventionalization of each type: only “Morrer de Y”, “Fartar(-se) de Y”, “Cansar(-se) de Y”, “Chorar de Y”, “Cair de Y” had a number of tokens that could attest to their conventionalization, as shown in Table 3.
“cair”/“to fall”, “rachar”/“to crack”, “rolar”/“to roll”) or physiological impact (“cagar”/“to shit”, “canzar”/“to be tired”, “mijar”/“to piss”, “morrer”/“to die”) and Y prototypically is an abstract name or a verb:

(4) 16: FM Melo: Letters Com as premissas de que haveria de seguir o Conde Ene ao Brasil, me acabei de destruir, empenhar e carregar de novas obrigações.

“With the assumptions that I should follow the Count Ene to Brazil, I finished of destroying, engage and load of new bonds.” (to finished of destroying = destroy a lot; finished of engage = to engage in a superlative way; finished of load = load a lot)

(5) 18: Azevedo: Japão [...] dragonas de ouro e desses chapéus de pluma que fizeram rebentar de medo o Imperador da China nas profundezas empedradas de Pekin.

 [...] gold epaulettes and these feather hats that made the Emperor of China burst of fear in the depths paved of Pekin. (to burst of fear = to have a lot of fear)

(6) 18: Álvares: Lira E quando eu morra de esperar por ela.../Deixai que eu durma ali [...] And when I die of waiting for her.../ Let me sleep here [...] (to die of waiting = to wait for a long, long time)

(7) 19N: Pt: Beira Maria do Carmo Borges, a presidente em exercício, não se cansou de valorizar esta festa, e tinha razões para isso.

Maria do Carmo Borges, the acting president, wasn’t tired of appreciate this feast, and she had reasons for this. (to not be tired of appreciate = to appreciate a lot)

(8) 19Or: Br: Intrv: ISP Aí Cacá fez Ubu, estourou e eu fiquei morrendo de inveja.

Then Caca made “bang”, he burst and I was dying of envy. (to die of envy = to have a lot of envy)

(9) 19: Fic: Br: Novaes: Mao Foi quando, quase se mijando de medo, o moleque o cutucou com a coroinha do bacamarte [...]

That’s when, almost pissing of fear, the boy nudged him with the butt of the blunderbuss [...] (to piss of fear = to have a lot of fear)

Corpus do Português, being a corpus consisted of more formal texts (cf. section 3) prevented the postulation of more broad generalizations about the habitat of the SCBE. Still, the data obtained allowed us to understand that SCBE is more pertinent to discursive contexts in which the speaker/writer has more freedom to express his subjectivity, since it is especially present in narrative sequences and dialogues (in fiction texts, 87.2% of its occurrence in the corpus used) and in excerpts of reports (other genres).

5. Conclusion

It was our intention here to expose the corpus dimension involved in the research of SCBE. By doing so, we presented an effective form for investigating constructional patterns in a language and the advantages that a corpus-based approach can offer to researches investigating this kind of objects.

To form this framework, beyond a very brief presentation of the theories that underpin our way of looking at the object, we presented the method used in the research and also the findings directly related to the choice of use corpus in the work: the conventionalization and productivity of the SCBE in Portuguese, the description of the construction and the texts in which the construction appears.

The results show that, in fact, it is advantageous to use corpora in language research, not only for providing access to information inaccessible to introspection, but also to allow more precise descriptions, and actual, of a given object, since that arise naturally information data. It is true that the use of corpus does not warrant a full analysis (in the study of the SCBE, for example, we found through the corpus research of common cases that we see in Portuguese, as “Pirar de rir”, something like “freak out laughing”), but, as stated by Fillmore (1992: 35),

“there can be any corpora, however large, that contain information about all of the areas [...] that I want to explore; all that I have seen are inadequate. [But] every corpus that I’ve had a chance to examine, however small, has taught me facts that I couldn’t imagine finding out about in any other way”.

6. Acknowledgments

“Superlative Constructions of Brazilian Portuguese: a study about the scale semantic” macroproject receives the financial support of CNPq and the project which this study integrates, the SCBE investigation, received the financial support – scholarship of FAPEMIG.

7. References


Past tense in Brazilian Portuguese: set of tense-aspect-modality features

Raquel Meister Ko. FREITAG
Universidade Federal de Sergipe, Centro de Educação e Ciências Humanas, Departamento de Letras Vernáculos
rko@uol.com.br

Abstract

In this paper, results from an investigation about the set of verbal features in Brazilian Portuguese are presented. Tense, aspect and modality features are described based on use of verbal forms in a sociolinguistic corpus of spoken Brazilian Portuguese. The verbal categories finding in the corpus are presented and the directions form > function and function > form. Results point that the IMP forms (simple and compound) are overlapping the most functions, specially the functions of modality domain, in irrealis.

Keywords: verbal categories; variation; Brazilian Portuguese.

1. Introduction

Normative grammars of Portuguese define the verbal paradigm as a tense: in the past scope there are the “pretérito perfeito” forms (simple and compound), “pretérito mais que perfeito” (simple and compound), “pretérito imperfeito” and future do pretérito), in indicative mode, and “pretérito imperfeito” in subjunctive mode. However descriptive and variacionist studies point that these forms pass for a) a semantic-discursive reset, with a single form expressing more than one function, losing the iconicity, and b) a morphosyntactic reset, with emergency and regularization of new forms and obsolescence of others. For example, there are evidences of obsolescence of simple “pretérito mais que perfeito” forms and the low frequency of compound “pretérito mais que perfeito” forms in context of anterior past; the simple “pretérito perfeito” forms assume this function (Coan, 1997). Other example is the emergency and regularization of form to expresses the imperfective progressive past, constituted by auxiliary verb “estar” + principal verb in gerund form, the compound “pretérito imperfeito” (Freitag, 2007). Still there are the switching between the “future do pretérito” and simple “pretérito imperfeito” forms (Costa, 1997), switching between “pretérito imperfeito” of indicative and subjunctive mode, and the specialization of compound “pretérito perfeito” form to expresses iterative perfect (Barbosa, 2008), and anymore. These switching contexts, emergency and regularization in verbal paradigm of Brazilian Portuguese are possibly due the reset processes of verbal paradigm, which origins are in the transition from Classical Latin to Vulgar Latin and to Romance languages. In this process language loses the aspectual distinction (“inflectum” and “perfectum” tenses), resulting in verbal paradigms in Romance languages that has an irregular paradigm as for the aspectual distinction. The emergency of compound forms, which codifies aspectual tense, is an evidence for this process.

In this paper, results from an investigation about the set of verbal features in Brazilian Portuguese are presented. Tense, aspect and modality features are described based on use’s description of verbal forms in a sociolinguistic corpus of spoken Brazilian Portuguese (Banco de dados Falantes Cultos de Itabaiana/SE). The sociofuncionalist assumptions (Tavares, 2003) are adopted for the analysis: the emergency of forms (grammaticalization follows Bybee, Perkings and Pagliuca, 1994) and the use regularization (linguistic change follows Labov, 1972). This approach postulates that clines of linguistic change presuppose stages of more or left stability in system, in so far as there are overlapping functions for one form and/or overlapping forms for a single function. First, TAM domain is presented; follows forms and functions correlation is.

2. TAM Domain

To analysis, we assumed the postulate that verbal form accumulate the tense, aspect and modality (TAM) features, in a complex functional domain (Givón, 1995, 2001), in which the features interacting. The complexity of the functional domains is due the fact that the boundaries between each feature are not always clear or precise, locking the separation, in fact, of each feature. However to pick up nuances of emergency, switching and regularization processes must be analyzing the verbal features globally, observing the discursive features that locking or favor any verbal form in any contexts.

2.1 Tense

Tense notion refers at the ordaining events (experiences) in points and intervals at a sequence; this concept is based on Reichenbach (1947): verbal tenses are determined for the ordaining of event point in function of the reference point and speech point. Based on speech point is possible establish three basic temporal relations: past, tense and future. Fixate only one point allows diagraming only three temporal relations; but others two parameters – event point and reference point – amplifying the temporal possibilities. Event point is the point when the event occurs; and reference point is a parameter point, a temporal reference, to determinate the event point, which is established according to the speech point. The speech point becomes the reference point when there is not temporal reference contextually explicit.

2.2 Aspect

Aspect linguistic category refers at the different modes to perceive the internal tense of an event (Comrie, 1976). Aspectual notion involves the internal set tense in events.
(initial, medial and final states/event presented as 
perfective/close or imperfective/open, and anymore 
possibilities). Perfective aspect is characterized for global 
perspective of event, which is expressed as closed, 
without internal reference, in a single united. Imperfective 
aspect focuses the internal constitution of events: its 
development (cursive, progressive imperfective aspect), 
or selecting stages of internal tense development (initial, 
medial and final), or expressing resultative states, and 
anymore. Imperfective aspect does not determine initial 
or final event points but focalizes its development, in 
contrast at perfective aspect, that emphasis the initial and 
the final points.

There is also other level of aspectuality: the inherent 
aspect of event. Bertinetto (2001) characterizes the event 
base on three aspectual proprieties: dynamicity 
durativity and homogeneity. Homogeneity refers at 
asbence of inherent internal limit in any event: a [+ 
homogeneity] event is this that does not change its nature; 
yet [-homogeneity] event presents an inherent 
achievement point. Dynamicity is a propriety 
characterized according to observation of dynamic atoms, 
which corresponding at minimal granularity of event and 
hence these are not divisible indefinitely [+ dynamicity]; 
the statics atoms can be divisible indefinitely [- 
dynamicity]. Durativity is a concept strictly operational, 
since any event, for so soon as far, has certain duration; 
nevertheless is possible distinguished events wit 
duration [+ durativity] from instant events [- durativity].

2.3 Modality

Modality is usually defined as the grammaticalization of 
speaker attitudes as the propositional content. In the 
languages it possible recognizes a grammatical category 
(the modality) which is similar at tense, aspect, number, 
gender. Givón (1995) divides the modality in epistemic, 
which refers at truth, belief, probably, certainty and 
evidence, or deontic, which refer at preference, desire, 
tention, ability, obligation and manipulation.

Epistemic modalities from Aristotelian logic 
tradition, follows Givón, have communicative equivalents: 
at the necessary truth corresponds the communicative 
equivalent of presupposition; at factual truth corresponds 
the realis assertion; at possible truth corresponds the 
irrealis assertion; and at non truth correspond the negative 
assertion. The communicative redefinition of epistemic 
modalities takes the presupposition as a proposition 
asumes as truth for anterior concordance, cultural 
convention or obvious at all participants in context of 
interaction. Realis assertion takes a proposition strongly 
asserted as truth; irrealis assertion is a proposition 
strongly asserted as possible, probably or uncertain; 
negative assertion takes the presupposition strongly 
asserted as false, in contradiction with explicit or assumed 
belief by hearing.

3. Prototypical tense features set in spoken 
Brazilian Portuguese

In a functionalist/cognitivist approach, the language 
structure reflects the experience structure, deriving from 
iconicity principle (cf. Bolinger, 1977; Givón, 1995). In a 
strong version of iconicity, model provides a one-to-one 
relation between form and function; however, in a 
moderate version the model provides the opacization 
between codification and function, andt becomes possible 
the variation between forms and functions. In Brazilian 
Portuguese spoken the past tense domain presents non 
univocal relations between forms and functions: one 
single form codifies more than one function and one 
single function is codified by more than one form.

The verbal categories identified in corpus are 
presented, first in form > function approach and follow in 
function > form approach.

The mapping of corpus results the follow forms (in 
indicative mode):

- Simple “Pretérito Perfeito” (simple PP)
- Compound “Pretérito Perfeito” (compound PP)
- Simple “Pretérito Imperfeito” (simple IMP)
- Compound “Pretérito Imperfeito” (compound IMP)
- Simple “Futuro do Pretérito” (simple FP)
- Compound “Futuro do Pretérito” (compound FP)
- Compound “Pretérito Mais que Perfeito” 
  (compound +QP)

These forms codifying follows functions:

- Anterior past: a past event which reference is 
other past event;
- Iterative perfective past: a past event which 
occurs systematically to past into the present;
- Imperfective past: a past event which reference 
is other simultaneous past event;
- Perfective past: a past event which reference is 
the speech point;
- Habitual past: an irregular past event recurrent;
- Conditional past: an event due of other past 
 event;
- Imminential past: an event which is presented at its 
pre-achievement.

Examples (1)-(12) illustrate the relation between 
forms and functions to expression of past tense in 
analysed corpus.

1) Inclusive conversei com alguns amigos meus que 
trabalham no escritório tal tudo e me ajudaram só 
a confirmar mesmo... que o curso era aquilo 
mesmo que eu já ESTAVA ESPERANDO se ita 
bmb lq 10
d

1The acronym in italics refers to source of data extrating from 
Sociolinguistic interview sample from Banco de dados Falantes 
Cultos de Itabaiana/SE. Two first letters are the county (Sergipe) 
and the three follow letters are the city (Itabaiana); follow letters
‘Also I talk with my friends which work in the office and they help me confirm the course was that even though I WAS EXPECTING (Compound IMP – Imperfective past)’

2) Olhe até ontem eu ACHAVA que seria um curso... né? que... dá as condições de emprego se ita fp sq 02
   ‘Look until yesterday I THOUGHT (Simple IMP – Imperfective past) it would be a course... right? that... gives employment conditions’

3) Chegou um menino colega dele “me dê aí um geladinho” ele... “vá lá pegar por favor” ele foi pegar quando ele ABRIU a geladeira que PEGOU o geladinho se ita mbh 08
   ‘Arrived a boy his colleague ”Give me a chilled” he... ”Please come pick up” when he was caught the fridges’s door opened (Simple PP – anterior past) the fridge that TOOK (Simple PP – perfective past) the chilled’

4) Uma vez meu colega me CONTOU que a mãe dele TINHA IDO para a rua se ita mbh 08
   ‘Upon time my friend TOLD (Simple PP – perfective past) me that yours mother WENT (compound PP – anterior past) out’

5) Se eu me formasse e visse que não que eu não dava pra ensinar que não era o meu ramo... eu não FARIA... eu não EXERCIA a profissão melhor dizendo se ita mb sq 09
   ‘If I graduated and I see that I could not to teach because it was not my business... I did not DO (Simple FP – conditional past)... I did not PURSUE (Simple IMP – conditional past) the profession rather’

6) Se a prova trouxesse questões desse tipo questões relacionadas ao dia-a-dia das pessoas questões problemas todos os professores de escolas particulares IAM se ADAPTAR também né? se ita mb sq 09
   ‘If the test brought issues matters to the day-to-day problems of people questions all private school teachers WOULD ADAPT (Compound FP – conditional past) also right?’

7) Ele achava que sendo universitário já era algo a mais que IA ACRESCENTAR no currículo dele se ita mb sq 10
   ‘He thought that being university student was already something else that WOULD ADD (Compound FP – iminental past) to his resume’

8) Desde a oitava série do ensino fundamental eu já tinha certeza de que a minha carreira seria na área da computação eu ENXERGUEI a área de tecnologia em geral como uma área bastante promissora e eu estava certo se ita mp sq 01
   ‘Since eighth midle school I yet had ’certain that my career would be in computation area I SAW (Simple PP – iterative perfective past) the technological area as a promissory area and I was right’

9) Eu acho que eu vou conseguir colher os frutos que eu TENHO PLANEJADO se ita mp sq 01
   ‘I think I will get to reap the fruits I HAVE PLANNED (Compound PP – iterative perfective past)’

10) Bom... eu pensei que o curso SERIA um curso voltado pra formação de professores né? se ita mb sq 08
    ‘Well I guess the course WOULD BE (Simple FP – iminental past) a course to teacher formation right?’

11) É preciso saber escrever muito bem no idioma inglês e no seu próprio idioma inclusive pessoas de outros países a Google COSTUMAVA também contratar para fazer as traduções se ita mp lq 10
    ‘You need to know how to write well in English and in your own language also people from other countries Google USED HIRE (Simple IMP – habitual past) to do the translations’

12) Como foi uma turma que sempre ESTEVE ENVOLVIDA... eu vejo que uma grande parte... né? está... realmente pensando e já criando os seus projetos... né? se ita fp sq 02
    ‘As was a class that was always WAS INVOLVED (Compound IMP – habitual past)... I see that a large part... right? is... really thinking and already creating their projects... right?’

---

are the sex (F = feminine and M = masculine), age (P = 16 at 20 year old; B = 26 at 35 years old); school grading (S = college completed; B college in course) and the last numbers refer to informant identification.
Each form and each function are analyzed separately in a quantitative approach and after the general results was correlated, as in table 1. This summarization is based on the studies about these verbal categories in the corpus of “Variation in expression of past tense: concurrent functions and forms” project researchers’ papers: Araujo & Freitag (2010, 2012), Cardoso & Santos (2011), Freitag & Araujo (2011), Freitag (2011), Freitag, Santos & Araujo (2011). Results showed at table 1 point that the IMP forms (simple and compound) are polysemic, recovering a range of functions of imperfective aspect and irrealsis modality. In perfective aspect, the actual verbal paradigm points the obsolescence of simple “pretérito mais que perfeito” form and the low productivity of compound “pretérito mais que perfeito” form; this form occurs in context of counter factuality. The realignment of verbal paradigm follow the specialization of forms based on distinction simple/compound: the IMP forms are distributed according the tendency simple IMP > habitual past and compound IMP > imperfective past. The correlation between forms and TAM set contributes to elucidate the clines of grammaticalization of semantic-discursive functions which the verbal forms codify; these results contribute to the refinement of the theoretical model. The analyses also subsides the application in tagger corpus processes.

### 4. Conclusion

Empirical analysis of linguistic change phenomena in different grammatical levels provides reflections about the theoretical models of grammaticalization, and contributes to point the limits and limitations of theory, reinforcing interface approaches. If at first time the grammaticalization studies focus the design of clines change of constructions (forms), actually the functional domains (function) has been highlight also at object of investigation. In verbal categories domain this approach has been showed productive and evidencing the need of more studies to priming the model.

### 5. Acknowledgements

This paper summarizes the results of the research project “Variation in expression of past tense: concurrent functions and forms” which was funded by Fundação de Apoio à Pesquisa e Inovação Tecnológica do Estado de Sergipe – FAPITEC (Proc. 019.203.00910/2009-0) and Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq (Proc. 401564/2010-0).

### 6. References


(Mestrado em Linguística) – Programa de Pós-graduação em Linguística da Universidade Federal de Santa Catarina.


7. Appendix

![Figure 1: Form and function relations in past tense domain in spoken Portuguese](http://www.ufmg.org.br/revistadogel/volumes/7/gel_7_2_t06.pdf)
Lexical and grammatical features of spoken and written Japanese in contrast: 
exploring a lexical profiling approach to comparing spoken and written corpora

Itsuko FUJIMURA, Shoju CHIBA, Mieko OHSO
Nagoya University; Reitaku University; Nagoya University
Furo-cho, Chikusa-ku, Nagoya, Japan
fujimura@nagoya-u.jp, schiba@reitaku-u.ac.jp, ohsomk@ac.auone-net.jp

This paper statistically demonstrates the lexical and grammatical characteristics of conversational Japanese by comparing a 100 hour spontaneous spoken corpus: the NUCC (Nagoya University Conversation Corpus) with a written corpus: the Balanced Corpus of Contemporary Written Japanese (monitor version). 1) The conversation corpus contains more involved production than the compared written corpus. 2) The comparison between the spoken and written interactional corpora shows that the participants leave much more metalinguistic and illocutionary traces in their speech than their writing. This is explained by the difference of degree of elaboration of the emitted messages and the difference of degree of closeness between/among participants of exchanges. 3) Fragmented utterances are much more frequent in spoken conversation than written texts. In Japanese, because of its grammatical structure (=SOV type language; particles come after their head), fragmentation, omnipresent conversational phenomenon, easily causes a functional and grammatical change in the role of particles.

Keywords: conversation; internet exchanges; metalinguistic; norm; linguistic change; Japanese; fragmentation.

1. Introduction
In this paper, we describe the lexical and grammatical characteristics of Japanese face-to-face spoken conversation and show how they differ from written registers. The aim of this research is to elucidate the characteristics of spoken Japanese, so we can later compare them with the results piled in the literature of this domain (Blanche-Benveniste, 1990; Biber, 1995 among others). For this purpose, we compare a spoken corpus: the NUCC (Nagoya University Conversation Corpus) with a written corpus: the BCCWJ (Balanced Corpus of Contemporary Written Japanese, monitor version). The former is a corpus of 100 hours built by our research team. The latter is a 45 million morpheme-sized written corpus. Our method is mainly quantitative. We perform this research with a tool named Lexical Profiling System, devised by one of the co-authors of this paper.

2. Corpora and tool

2.1 NUCC
The NUCC was constructed between 2001 and 2003, and is available for research purposes from the site (https://dbms.ninjal.ac.jp/nuc/index.php?mode=viewnuc) free of charge. It is composed of transcriptions of 129 uncontrolled, natural conversations between or among friends, family members or colleagues. Each conversation has 2 to 4 participants and lasts 30 to 60 minutes. The participants are 198 native speakers of Japanese of various ages and from diverse academic backgrounds. Each conversation constitutes a file so that the corpus NUCC consists of 129 files.

Conversations were recorded and transcribed in standard Japanese orthography. The Japanese orthography currently used is quite phonemic, but suprasegmental features are not captured. Hence, accent, intonation, and prominence are not transcribed. Only the rising intonation that indicates questioning is marked with a question mark at the end of an utterance.

The corpus contains about 1.5 million morphemes (“short unit words” according to UniDic (cf. Ogiso et al., 2012)), which shows that this is the largest corpus currently available of spontaneous spoken Japanese. As a caveat, there are more female participants (161) than male (37), and many of the participants are graduate students majoring in linguistic subjects. The lack of balance of the participants may be reflected in the data taken from this corpus.

2.2 BCCWJ (monitor version)¹
The integral BCCWJ, published in 2012, includes about 170,000 samples of written texts, which are classified into carefully designed subcorpora (genres), namely books, newspapers, magazines, whitepaper texts, Internet texts, Diet minutes, among others. We see the BCCWJ as a good sample of written Japanese, because the corpus contains the samples from many genres, each of which is relatively large. It also utilizes unique sampling strategies so that the corpus represents the most recent status of contemporary written Japanese (Maekawa, 2007).

In this work, we used the monitor version of the BCCWJ earlier released in 2009, which is a part of the integral version. The monitor version consists of 4 subcorpora indicated in Table 1. We use the BCCWJ in two ways. One is the whole BCCWJ (monitor version) for the grammatical study in section 4, and the other, its subcorpora: Books (BK) and Internet Bulletin Boards (IBB) for the lexical studies in section 3. The BK is composed of 10423 samples taken from various genre of books published between 1971-2005. We used it because it is the largest part of the BCCWJ and for its standardized nature as written corpus. The IBB consists of “Questions and Answers” type written exchanges between anonymous writers and readers, published on Yahoo Japan’s web site in 2005. The IBB is an interesting material to compare with the NUCC, because of their shared characteristics and for its novelty as a medium of communication. Both of them involve interaction

¹ Cf: http://www.ninjal.ac.jp/english/products/bccwj/. The BCCWJ refers to the BCCWJ (monitor version) from section 3 below.
between/among participants. The relation between/among participants is different though: the participants in the latter have close relationships while those in the former are strangers. They made real-time interactions in the latter, while there is a time lag between questions and answers in the former.

Table 1 indicates the characteristics of the studied corpora.

<table>
<thead>
<tr>
<th>Subcorpus of BCCWJ and NUCC</th>
<th>Number of morphemes (millions)</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Books (BK)</td>
<td>36.0</td>
<td>No interaction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Elaborated production</td>
</tr>
<tr>
<td>White Paper</td>
<td>5.8</td>
<td>Long-distance interaction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Prepared production</td>
</tr>
<tr>
<td>Internet Bulletin Boards (IBB)</td>
<td>6.7</td>
<td>Close interaction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Real-time production</td>
</tr>
<tr>
<td>Minutes of the National Diet</td>
<td>5.5</td>
<td></td>
</tr>
<tr>
<td>NUCC</td>
<td>1.5</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Subcorpora of the BCCWJ (monitor version) and the NUCC

2.3 Lexical profiling system
The Lexical Profiling System is designed to compare corpora of different size, genre, or even an individual part of a corpus with the whole. The data to be compared are morphologically analyzed by a GUI program Chamame (ver. 1.71) (composed by a part-of-speech and morphological analyzer: Mecab (ver. 0.98) and a dictionary: UniDic (ver. 1.3.12)), and the frequency of lemmas, word forms, bigrams are counted and stored in a database. The tool then computes the frequencies of these units using different statistical measures such as LLR (Log-Likelihood Ratio) among others.

3. Lexical studies
3.1 60 Basic morphemes in the NUCC
First of all, we identified the 60 morphemes employed in all 129 conversations of the NUCC as in Table 2 in order to compare later the use of these morphemes in the NUCC and the IBB and the BK. We could say that these are basic morphemes of Japanese conversation. These consist of 6 adjectives, 4 adverbs, 1 conjunction, 4 interjections, 6 nouns, 18 particles, 1 prefix, 2 pronouns and 12 verbs. Among the 18 particles, there are 4 utterance-final interactional particles, 13 sentence-internal casual or conjunctive particles and “no”, “No”, one of the most frequently used morphemes in Japanese, is subcategorized into three according to the dictionary UniDic: genitive (of in English), quasi-nominal (thing, nominalizer) and interactional. The first two are sentence-internal particles and the last one, utterance-final particle.

<table>
<thead>
<tr>
<th>POS</th>
<th>No</th>
<th>Morpheme</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADJ</td>
<td>6</td>
<td>nai (not to exist), yoi (good), you (to look like), sugoi (superb), sonna (that kind of), sôno (that)</td>
</tr>
<tr>
<td>ADV</td>
<td>4</td>
<td>mou (already), dou (how), you (so, in such a way), kou (this way)</td>
</tr>
<tr>
<td>AUX</td>
<td>6</td>
<td>da, desu (DEC), ruru (PASS/POT/HON), ta (PAST), nai (NEG), teru (PROG, PERF)</td>
</tr>
<tr>
<td>CONJ</td>
<td>1</td>
<td>de (and)</td>
</tr>
<tr>
<td>INTJ</td>
<td>4</td>
<td>un (yeah, I see), ah, a! (wow), ano (well)</td>
</tr>
<tr>
<td>NOUN</td>
<td>18</td>
<td>Uterance-final, interactional:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ne (TAGQ, you know), yo (I tell you), ka (Q), na (I tell you)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sentence–internal:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>wo (ACC), ga (SUB), wa (TOP), ni (DAT, LOC, TEMP, ADVL), to (and with), keredo (although), kara (from), mo (also), kurai (about)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>te, de (and (Vi/ADJ Suffix)) te (QUO), made (until), mo; GEN, QN (sentence-internal), INTA (utterance-final)</td>
</tr>
<tr>
<td>PREFIX</td>
<td>1</td>
<td>yo (POLITE)</td>
</tr>
<tr>
<td>PRO</td>
<td>2</td>
<td>nant (what), sore (that)</td>
</tr>
<tr>
<td>VERB</td>
<td>12</td>
<td>kuru (to exist, to be), dekiru (to be able to), miru (to see, to look at), naru (to become), wakaru (to understand), omou (to think), kuru (to exist), kuru (to come), suru (to do), yaru (to do), iku (to go), iu (to say)</td>
</tr>
</tbody>
</table>

Table 2: 60 Morphemes used in all 129 conversations of the NUCC

The fact that there are no personal pronouns in the list should not be interpreted as lack of active interaction. In Japanese, one can speak even for 30 minutes long without mentioning “me” or “you”. Especially the

2 These are the output of the Analyzer Chamame. We only modified the result of the automatic analysis by grouping “Rentai-shi”, “Keijo-shi” and “Keiyo-shi” in Adjective, since the major function of these three categories is noun modification.

reference to the interlocutor with a personal pronoun meaning “you” is considered to be rude. The frequent uses of interactional particles like ne, yo, deictic verbs like iku (to go), kuru (to come) and honorific expressions fill the gap caused by the lack of personal pronouns.

### 3.2 NUCC compared with Books (BK)

The statistic measure: LLR demonstrates the degree of typicality for these 60 morphemes compared with the BK. Even if they are used in every conversation of the NUCC, their degree of typicality is not homogeneous. The most typical 10 morphemes relative to the BK with the highest degree of LLR and the least typical 5 with the lowest degree of LLR are shown in Table 3. The MPM indicates the number of morphemes per million.

<table>
<thead>
<tr>
<th>no</th>
<th>Morpheme</th>
<th>Function</th>
<th>LLR</th>
<th>MPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>un</td>
<td>Yeah, I see</td>
<td>310,539</td>
<td>30,003</td>
</tr>
<tr>
<td>2</td>
<td>ne</td>
<td>TAGQ.</td>
<td>127,327</td>
<td>19,754</td>
</tr>
<tr>
<td>3</td>
<td>tte</td>
<td>QUO (contracted)</td>
<td>80,628</td>
<td>12,575</td>
</tr>
<tr>
<td>4</td>
<td>kai</td>
<td>Q</td>
<td>67,541</td>
<td>22,884</td>
</tr>
<tr>
<td>5</td>
<td>teru</td>
<td>PROG/PERS (contracted)</td>
<td>59,022</td>
<td>9,714</td>
</tr>
<tr>
<td>6</td>
<td>sou</td>
<td>so</td>
<td>51,485</td>
<td>11,024</td>
</tr>
<tr>
<td>7</td>
<td>yo</td>
<td>I tell you</td>
<td>44,561</td>
<td>9,790</td>
</tr>
<tr>
<td>8</td>
<td>nani</td>
<td>what</td>
<td>39,340</td>
<td>9,820</td>
</tr>
<tr>
<td>9</td>
<td>keredo</td>
<td>although</td>
<td>36,307</td>
<td>6,436</td>
</tr>
<tr>
<td>10</td>
<td>al</td>
<td>INTJ</td>
<td>36,090</td>
<td>4,273</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>56</td>
<td>suru</td>
<td>to do</td>
<td>-2,899</td>
<td>14,343</td>
</tr>
<tr>
<td>57</td>
<td>wa</td>
<td>TOP</td>
<td>-4,030</td>
<td>25,419</td>
</tr>
<tr>
<td>58</td>
<td>ni</td>
<td>IO etc.</td>
<td>-4,301</td>
<td>29,498</td>
</tr>
<tr>
<td>59</td>
<td>iru</td>
<td>to exist, to be</td>
<td>-6,440</td>
<td>1,200</td>
</tr>
<tr>
<td>60</td>
<td>wo</td>
<td>ACC</td>
<td>-20,037</td>
<td>3,939</td>
</tr>
</tbody>
</table>

Table 3: Typical and atypical morphemes in the NUCC compared with the BK

We can easily see that interactional expressions and contracted forms are typical in face-to-face conversation. The backchannel un appears 30,000 times per million. This is 3% of the morphemes used in the NUCC. In contrast, the least typical 5 are indispensable grammatical morphemes in any Japanese utterance regardless of spoken or written. Negative value means that the morpheme is less used in the conversation than in books. In fact, the least typical morpheme with the lowest degree of the LLR, the accusative marker “wo” is often not pronounced in conversation.

### 3.3 NUCC compared with the IBB

We then compare the uses of these 60 morphemes in the NUCC with the IBB in order to show the difference in spoken and written interactional exchanges. These interactions are characterized by two points of view: social closeness and physical distance between two participants of communication.

#### 3.3.1 Typical Morphemes

The most typical 10 morphemes of the NUCC compared with the IBB are following (LLR is in bracket).

1. unyeah, I see (324,691)
2. daDEC (159,975)
3. ne TAGQ, you know (146,670)
4. no/n GEN, QN or INTA (108,044)
5. ka Q (101,483)
6. sou so, in such a way (95,564)
7. tte QUO (contracted) (85,429)
8. to PAST (75,684)
9. nani what (67,687)
10. iu to say (61,961)

The high frequency of da (declarative marker) is noteworthy. Its occurrence seems to derive from the frequent use of short turn taking in face-to-face conversation, especially the large number of casual backchannel feedback finishing with “da”, such as “sou-na-n-da” (so-DEC-QN-DEC, “Indeed”), whereas this is not the case in written correspondence on the Internet. The participants are not in real-time interactions in “Questions and Answers” type exchanges, so that the frequent use of short turn taking is not common. Also the participants of the IBB do not have a close relationship between them, because in fact they do not know each other and in general the written communication does not allow them to make intimate interactions in Japanese. These are the reasons for which the informal declarative form “da” is typical in the NUCC, whereas the formal one “desu” is numerous in the IBB.

#### 3.3.2 Verb: To Say in the Conversation

Among the 12 verbs in the Table 1, “iu” (to say) is the most typical one of the NUCC with LLR: 61,961, followed by iku (to go, LLR: 20,919), yaru (to do, LLR: 17,603), suru (to do, LLR: 14,343), kuru (to come, LLR: 13,558), aru (to exist, LLR: 12,403), omou (to think, LLR: 10,903), wakaru (to understand, LLR: 8,613), naru (to become, LLR: 5,970), miru (to see, to look at, LLR: 5,599), dekiri (to be able to, LLR: 1,489) and iru (to exist, to be, LLR: 1,200) in descending order. This metalinguistic verb to say is used much more often in oral conversation than in written correspondence. It may be explained at least partially by the fact that in real-time exchanges, we talk a lot about “how to say” something. The speaker leaves traces of metalinguistic activity in his speech. For example, when we hesitate in seeking an expression, we say: “How should I say?”. In the example

\[\text{Table 3: Typical and atypical morphemes in the NUCC compared with the BK}\]

<table>
<thead>
<tr>
<th>no</th>
<th>Morpheme</th>
<th>Function</th>
<th>LLR</th>
<th>MPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>un</td>
<td>Yeah, I see</td>
<td>310,539</td>
<td>30,003</td>
</tr>
<tr>
<td>2</td>
<td>ne</td>
<td>TAGQ.</td>
<td>127,327</td>
<td>19,754</td>
</tr>
<tr>
<td>3</td>
<td>tte</td>
<td>QUO (contracted)</td>
<td>80,628</td>
<td>12,575</td>
</tr>
<tr>
<td>4</td>
<td>kai</td>
<td>Q</td>
<td>67,541</td>
<td>22,884</td>
</tr>
<tr>
<td>5</td>
<td>teru</td>
<td>PROG/PERS (contracted)</td>
<td>59,022</td>
<td>9,714</td>
</tr>
<tr>
<td>6</td>
<td>sou</td>
<td>so</td>
<td>51,485</td>
<td>11,024</td>
</tr>
<tr>
<td>7</td>
<td>yo</td>
<td>I tell you</td>
<td>44,561</td>
<td>9,790</td>
</tr>
<tr>
<td>8</td>
<td>nani</td>
<td>what</td>
<td>39,340</td>
<td>9,820</td>
</tr>
<tr>
<td>9</td>
<td>keredo</td>
<td>although</td>
<td>36,307</td>
<td>6,436</td>
</tr>
<tr>
<td>10</td>
<td>al</td>
<td>INTJ</td>
<td>36,090</td>
<td>4,273</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>56</td>
<td>suru</td>
<td>to do</td>
<td>-2,899</td>
<td>14,343</td>
</tr>
<tr>
<td>57</td>
<td>wa</td>
<td>TOP</td>
<td>-4,030</td>
<td>25,419</td>
</tr>
<tr>
<td>58</td>
<td>ni</td>
<td>IO etc.</td>
<td>-4,301</td>
<td>29,498</td>
</tr>
<tr>
<td>59</td>
<td>iru</td>
<td>to exist, to be</td>
<td>-6,440</td>
<td>1,200</td>
</tr>
<tr>
<td>60</td>
<td>wo</td>
<td>ACC</td>
<td>-20,037</td>
<td>3,939</td>
</tr>
</tbody>
</table>

4 The occurrence of numerous “no” in conversation primarily comes from the frequent use of the interactional usage of this morpheme placed at the end of utterances. However there are also many “no” placed before the declarative “da” often realized “n-da”. This frequently used bigram is often analyzed as a compound auxiliary in Japanese linguistics. This is not the case in this study, as to our morphological analyzer processes them as QN-DEC.
(1), having once used the word "room", the speaker corrects it with the word "entrance" while talking about the process of this correction: heya-tte-i-ka (Can-I say "room"?). In this type of metalinguistic utterance, the verb: to say plays the main role.

(Ex.1) conversation 019

Gozenchu-wa zuutto heya-ni morning-TOP throughout room-LOC heya-tte-IU-ka genkan-ni haitte-ta-n-da room-QUO-SAY-Q entrance-LOC enter-PAST-QN-DEC “I was in a room all morning, can-I SAY “room”?, in the entrance.”

In contrast, in the activity of writing, even private texts like those found in the IBB are prepared and elaborated. That would be why there is a big gap in the use of the verb: to say between the IBB and the NUCC.

4. Grammatical study: fragmentation

Finally, we will discuss how to end an utterance in Japanese conversation.

4.1 13 basic utterance-final morphemes in the NUCC compared with the BCCWJ

We analyze 13 morphemes employed at the utterance-final position in all 129 conversations of the NUCC. This position is defined by a period or a question mark in the transcription. We can consider these 13 items as the basic utterance-final morphemes in Japanese informal face-to-face exchanges. The Table 4 indicates that when compared with the BCCWJ, the most typical utterance-final morpheme of the NUCC is the interactional particle: “ne”, while the least typical one is the auxiliary: “ta (Past Tense)”. These are classified into three groups. The first includes 4 final interactional particles (Final PRT): “ne, yo, na, ka”. The second, 3 auxiliaries (AUX): “da, nai, ta” and the third, 6 sentence-internal conjunctive particles (PRT): “te, keredo(kedo), tte, kara, de, ni” as indicates the Table 4.

Of these three groups, the frequent use of interactional particles in conversation is entirely predictable. The normal position of these morphemes is at the end of utterances. The use of auxiliaries at the final position is also ordinary in every type of text. The most interesting phenomenon is the use of sentence-internal conjunctive particles at the utterance-final position. It is not normative in Japanese traditional grammar and absent in the written formal texts, while it is found in every conversation of the NUCC.

<table>
<thead>
<tr>
<th>POS</th>
<th>morpheme</th>
<th>function</th>
<th>LLR</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRT</td>
<td>ne</td>
<td>TAGQ, Alignment</td>
<td>55,092</td>
</tr>
<tr>
<td>PRT</td>
<td>te</td>
<td>and</td>
<td>22,516</td>
</tr>
<tr>
<td>PRT</td>
<td>keredo(kedo)</td>
<td>although</td>
<td>14,129</td>
</tr>
<tr>
<td>PRT</td>
<td>tte</td>
<td>QUO</td>
<td>13,949</td>
</tr>
<tr>
<td>Final PRT</td>
<td>yo</td>
<td>I tell you</td>
<td>12,305</td>
</tr>
<tr>
<td>Final PRT</td>
<td>na</td>
<td>I tell you, I know</td>
<td>10,520</td>
</tr>
<tr>
<td>PRT</td>
<td>kara</td>
<td>because</td>
<td>7,526</td>
</tr>
<tr>
<td>PRT</td>
<td>de</td>
<td>and</td>
<td>6,583</td>
</tr>
<tr>
<td>Final PRT</td>
<td>ka</td>
<td>Q</td>
<td>6,329</td>
</tr>
<tr>
<td>PRT</td>
<td>ni</td>
<td>DAT, LOC, TEMP, ADVL</td>
<td>4,672</td>
</tr>
<tr>
<td>AUX</td>
<td>da</td>
<td>DEC</td>
<td>1,027</td>
</tr>
<tr>
<td>AUX</td>
<td>nai</td>
<td>NEG</td>
<td>270</td>
</tr>
<tr>
<td>AUX</td>
<td>ta</td>
<td>PAST</td>
<td>-7,774</td>
</tr>
</tbody>
</table>

Table 4: LLR of final morphemes of the NUCC compared with the BCCWJ

4.2 From sentence-internal particle to utterance-final particle or vice versa

We could say first that there are many syntactically incomplete sentences in Japanese conversation as in other languages. This could be due to the pragmatics of conversation: the participants of communication collaborate to finish a sentence as in example (2). The utterance of the speaker A stops at the end of the subordinate clause marked by an adverasive conjunction KEDO (= KEREDO “although”). The speaker B completes A’s utterance by adding the main clause.

(Ex.2) conversation 035

A: sensei-ni mikkahodo tomatte-morae-ba professor-IO several days stay-make-if ii-n-desu KEDO good-QN-DEC(formal) ALTHOUGH “Although it would be better if we could ask the professor stay here for several days.”

B: A! deki-nai-n-desu-ka. ah can-NEG-QN-DEC(formal)-Q “Ah, you can not do so.”

However in most cases, this kind of collaboration between the participants of conversation is not obvious. The particle at the end of the utterance no longer has the conjunctive function linking the subordinate and main clauses but rather has a modal function. The example 3 shows that the utterance emitted by speaker B does not adverasive with that of speaker A, despite the existence of KEDO. The function of KEDO in this case is to attenuate the assertive power of the predication and to show the intention of continuing the dialogue to the interlocutor (cf. Saegusa, 2007).

5 Syntactic fragmentation does not necessarily correspond to informational fragmentation (cf. Matsumoto 2010).
typical interactional particles, interjections, markers of agreement and "what", reflect the involved nature of this activity, when compared with books.

3) The typical auxiliary of conversation, compared with written correspondence, is “da (declarative)”. It may reflect the high frequency of short answers and backchannels in conversation.

4) The typical verb in conversation is “iu (to say)”. This could come from frequent metalinguistic use of this verb in spontaneous speech, which, unlike written discourse, is not elaborated.

5) 13 basic utterance ending forms within conversation have been identified. Some of them are only used at the sentence-internal position in written texts. This is due to close and frequent exchanges between participants which cause incomplete utterances. In Japanese, because of its grammatical structure the fragmentation easily causes a functional and grammatical change in the role of particles.

Lastly, we summarize some of the features of conversational Japanese in contrast with written Japanese. It has more involved production, more metalinguistic and illocutionary traces. It also has more fragmented structures, which could cause a dynamic linguistic change. These are universal characteristics of spoken exchanges mentioned in Biber (1995), primarily due to the lack of time in real-time interactions (Biber, 2010) and secondarily to the closeness between two participants during exchanges. We also found some specific characteristics of Japanese conversation, like the absence of personal pronouns. This is explained only by the individual language structure.

6. Acknowledgements

This work was supported by MEXT/JSPS KAKENHI Grant Number (23520504).

7. References


In search of modality: a spontaneous speech corpus-based study

Heliana MELLO, Luciana ÁVILA, Priscila OSÓRIO, Raíssa CAETANO, Adriana RAMOS

Universidade Federal de Minas Gerais, Faculdade de Letras – UFMG
Av. Antônio Carlos, 6627 – Pampulha – Belo Horizonte, MG 31270-901 Brazil
heliana.mello@gmail.com

Abstract

Modality in speech can be taken to be a speaker’s evaluation of an uttered locutive material. This paper explores the semantic notion of modality through the analysis of a Brazilian Portuguese spontaneous speech corpus. The building of the corpus took into account the utterance unit, as it is proposed in the Language into Act Theory (Cresti, 2000). This paper aims at briefly presenting modality studies developed so far within the C-ORAL-BRASIL corpus. The studies presented in this paper focus on: the identification of morpho-lexical modality indexes in tone units, a comparative study between modal adverbs of certainty in a sample of Brazilian and European spontaneous speech corpora and the mapping of modal adverbial constructions in Brazilian Portuguese. In all these studies, we carried a qualitative analysis, in order to describe the occurrences of the different modal indexes, such as for example: (semi-)auxiliary modal verbs, modal adverbs, verbs of propositional attitude, volitional verbs, modal adjective constructions and emerging forms.

Keywords: modality; C-ORAL-BRASIL; corpus-based research; spoken Brazilian Portuguese.

1. What is modality?

Modality in speech can be taken to be a speaker’s evaluation of an uttered locutive material following the Ballyan view that modality is the evaluation (“Modus”) of the speaker towards his own locutionary content (“Dictum”) (Bally, 1932). However, precisely defining this category is a difficult task, since, according to Venn (1888: 245), “[modality is] a variety of place upon that most thorny and repulsive of districts in the logical territory.” This difficulty stems from different factors: (a) in its study tradition, modality has been the subject matter of both logical studies and natural language studies (Lyons, 1977), which implies a methodological maze not always productive for the research on its actual linguistic use; (b) this category interrelates with a number of grammatical phenomena such as time, aspect and mood (Palmer, 1986), prosody, information organization, among others; and (c) the concept of modality itself overlaps those of attitude, illocution and emotion (Mello & Raso, 2012). Therefore, for the purposes of this paper, modality in speech will be understood as the conceptualizer’s evaluation of an uttered locutary material, anchored in a communicative situation.

2. The C-ORAL-BRASIL

The investigation of modality reported in this paper was carried through the analysis of a Brazilian Portuguese Spontaneous Speech Corpus, the C-ORAL-BRASIL I (Raso & Mello, 2010, 2012). This corpus is the fifth branch of the C-ORAL-ROM project (Cresti & Moneglia, 2005), a set of corpora representative of European Portuguese, French, Italian and Spanish spontaneous speech. The C-ORAL-BRASIL follows the same architecture and technical specifications found in the C-ORAL-ROM corpora, therefore being entirely comparable to the latter.

The C-ORAL-BRASIL I is presented through a DVD in which the following files can be found: sound files (wav); metadata featuring textual, situational, participants’ information; transcriptions (rtf) segmented in tone units and utterances following the Language into Act parameters (Cresti, 2000); PoS tagged transcriptions in txt and XML formats through the PALAVRAS parser (Bick, 2000), speech to text alignment in XML format through the WinPitch aligner (Martin, 2004).

The C-ORAL-BRASIL I, the informal part of the C-ORAL-BRASIL project, features a very broad diaphasic variation, that is, speech situation variation, in view of representing as accurately as possible, a range of different speech acts through actual spontaneous linguistic activity.

The corpus textual typology is branched into monologues, dialogues and conversations, which on their part, are divided into public and private.

The C-ORAL-BRASIL I also features a balanced and informationally tagged subcorpus for study purposes. The information tagging was carried following the Language into Act Theory (Cresti, 2000) and the Information Patterning Theory (Cresti & Moneglia, 2010). Searches in the subcorpus can be carried through the search interface IPIC (http://lablita.dit.unifi.it/ipic/).

3. In search of modality

The C-ORAL-BRASIL subcorpus was used as data source for the search of modal indexes since it is balanced for textual typology and it is informationally tagged, which allows for the identification of information units that carry modal indexes. The subcorpus is composed by 20 texts of three interactional typologies: dialogic (7), monologic (7) and conversational (6), divided into private and public, in a total of approximately 30,000 words.

The procedure adopted for analysis was to manually search for modal indexes and classify them in their context of occurrence according to their typological characteristics, which are: part of speech, information unit of placement, semantic label (aletic, epistemic or deontic modality), textual typology, gender and speaker schooling level. This qualitative classification was followed by a quantitative analysis, which took into consideration
type-token ratio and a multivariate analysis supported by the R environment (http://www.r-project.org/). The semantic label assigned to each token was validated through group discussion. Cases which presented disagreements or difficulties in labeling were reassessed until reaching satisfactory classification agreement.

Among the studies that resulted from this research effort are: identification of morpholexical modality indexes in tone units (Mello et al., 2010), a comparative study between modal adverbs of certainty in a sample of Brazilian and European spontaneous speech corpora (Mello et al., 2011), a study about the epistemic character of conditional constructions (Ávila & Côrtes, 2011), the description of modal indexes and their pragmatic-cognitive consequences (Ávila, 2012), and the mapping of modal adverbial constructions in Brazilian Portuguese (Mello & Caetano, in progress).

The research has shown the following distribution for modal types: from 2,573 utterances examined, 250 have some kind of modal marking (9.71%). The majority of modal markings are epistemic (57.85%), with deontic marking featuring 23.57% and aletic marking exhibiting 18.57%. The modal indexes found and their morpholexical classification, along with percentage of occurrence are shown in Table 1 below.

In order to illustrate the data analyzed, some examples follow below.

1. =S [171] no /=PHA= thirty reals /=TOP= then I &j [2]=SCA= I /=INT= I suppose that he thinks like that /==EXP_r= maybe at my place one need to go shopping and everything /==COM_r= right// /=PHA=$ (bpubmn01)

   =S [171] não /=PHA= trinta reais /=TOP= aí eu &j [2]=SCA= eu /=INT= eu fico imaginando que e’ fica pensando assim /==INT= Nossa Sio’ /=EXP_r= às vezes lá em casa tá precisando de fazer uma compra e tudo /==COM_r= né // /=PHA=$ (bpubmn01)

2. *LUC: [74] <if on the first time that you say a word /=SCA= it doesn’t work /=TOP= it never will=/=COM= got it/=PHA=$ (bfamcv04)

   *LUC: [74] <se na primeira vez que cê falou uma palavra /=SCA= não> for /=TOP= nunca mais vai ser /==COM= entendeu // /=PHA=$ (bfamcv04)

3. *PAU: [153] because it’s most likely that I’ll build a wall there /==COM=

   *PAU: [153] porque é capaz d’ eu subir uma parede lá // /==COM=

As for the comparison between Brazilian and European Portuguese modal adverbs of certainty, the results indicate an overall rate of occurrence higher in EP than in BP. The explanatory hypothesis for this finding is discussed in Mello et al. (2010) and is related to social hierarchization and education level differences in the two cultures. In Table 2 below the overall token numbers are presented for both language varieties, exhibiting the higher usage of modal marking in EP vis-à-vis comparative situations in BP.

<table>
<thead>
<tr>
<th>Modality morpholexical strategies</th>
<th>Types</th>
<th>Percentages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjectives (or nominals in adjectival function) in predicative position</td>
<td>(é) lógico, é provável, é importante, (é) verdade</td>
<td>1.42%</td>
</tr>
<tr>
<td>Adverbs and adverbiale expressions</td>
<td>Talvez, certamente, realmente, às vezes, também, logicamente, sinceramente, com certeza, completamente, sem dúvida, possivelmente, na verdade, na realidade</td>
<td>6.42%</td>
</tr>
<tr>
<td>Conditionals</td>
<td>[if X then Y]</td>
<td>13.21%</td>
</tr>
<tr>
<td>Modal constructions</td>
<td>ten condição (de), tem chance de, o que acontece, ter que, ficar imaginando, ficar pensando, (é) para + inf., dá para + inf., ter certeza, vai saber, tem jeito</td>
<td>22.14%</td>
</tr>
<tr>
<td>Future</td>
<td>vou + inf.</td>
<td>1.07%</td>
</tr>
<tr>
<td>Preterit future</td>
<td>ia ser, ia dar, seria</td>
<td>3.21%</td>
</tr>
<tr>
<td>Other forms</td>
<td>Digamos que, de certa forma</td>
<td>3.57%</td>
</tr>
<tr>
<td>Verbs (indicative mood – present, perfect and imperfect; infinitive)</td>
<td>Dever, poder, achar, acreditar, acontecer, ver, conseguir, precisar, pensar, dar e parecer</td>
<td>48.92%</td>
</tr>
</tbody>
</table>

Table 1: Morpholexical strategies, types and percentages
The marking of modality in conditional constructions has evidenced epistemic values as predominant. As for the information structure organization, the most frequent structuring brings protasis in Topic and apodosis in Comment units. The cognitive value of this organization needs further study in order to determine if and how modality indexes within different informational units interact at a higher semantic level.

On a pragmatic-discursive level, especially as far as modal verbs are concerned, the major functions found in our data were: (a) mitigation of previous assertion when the modalizer occurs in Parenthetical units; (b) mark agreement or disagreement; (c) mitigation of sociocultural differences among participants in a given interaction.

4. Provisional Conclusions

So far, our research has shown that verbs are the major modality agent in BP and epistemic modality is the most frequent semantic type found. Another interesting finding is that BP allows for multiple modal valency utterances and tone units. What that means is that the same modal index may carry different semantic values depending on the utterance and tone unit in which it is found.

The preliminary study on adverbs of certainty in a sample of BP and EP has shown an upward curve representing an increased use of modal adverbs in lower diastaty in BP if compared to higher ones, which may indicate socioculturally-based differences in the expression of politeness in the two groups. Additionally, the comparison between EP and BP indicated differences in lexical choices in these two varieties along with a much higher usage of modal markings in EP than in BP.

Modal adverbs in BP spontaneous speech have complex usage patterns. The bare modal semantic meaning of adverbials is associated with other notions such as temporality, which should be further investigated.

Additionally, we have observed a strong interface between semantics and pragmatics which we address in face of participants’ roles in speech events and their stance.

Last but not least, the epistemic character of conditionals seems to indicate the different degrees of “actuality” between the protasis and the apodosis.

5. Acknowledgements

We are grateful to the following for research grants: CNPq, FAPEMIG, UFMG.

6. References


Bally, C. (1932). Linguistique générale et linguistique


C-ORAL-BRASIL. Available at: <http://www.c-oral-brasil.org>.

C-ORAL-ROM. Available at: <http://lablita.dit.unifi.it/coralrom>.


Temporal and causal uses of the connector *come* in spoken Italian

Francesca GATTA

SSLIMIT, Università di Bologna, polo scientifico didattico di Forlì
Francesca.Gatta@unibo.it

Abstract

This paper is part of a larger research project on Italian connectors. The aims is to study the contribution of connectors to the encoding of conceptual relationship between two processes. The point of view to study the relationship between encoding and inference is the conceptual framework proposed by Prandi (2004). The occurrences of *come* in spoken Italian (LIP) allow us to describe the value of the connector as proposition and conjunction. As proposition *come* has a basic modal / comparative meaning; the temporal and the causal value of *come* derives from inferences which overlays other relationship: when the contents of the connected propositions allow, the meaning of the connector may be enriched by a temporal or a causal value.

Keywords: ‘Come’ (conjunction); connector; encoding; inference; LIP.

1. Introduction

This paper is a small part of a larger research project on Italian connectors. The project aims to study the contribution of connectors to the encoding of conceptual relationship between two processes. The general questions we are posing are: if the relationship between two processes can be inferred, what is the function of the connector? And can the contents of the connected propositions attribute a “new” value to the connector, extending the meaning of the latter?

These are questions which concern the relationship between encoding and inference, and that between content and expression. A conceptual framework for examining such questions has been proposed by Prandi (2004, III; 2006), who argues that in some areas of language, for instance in the nucleus of the sentence, encoding is relational (roles are assigned by a grammatical relation, so the grammatical relation assigns a content), while in others, such as the more outlying parts of the sentence, coding is punctual and the conceptual content prevails over the grammatical relation. In other words, there are some cases where the grammatical relation imposes itself on the contents and is independent of them, whereas in other cases the content is independent of the linguistic expression, and the latter merely encodes a conceptual relationship which is created outside the expression as such.

We believe our findings on the temporal and the causal value of *come* in spoken Italian support this theoretical position.

2. Data

Our data is taken from corpora of spoken Italian. This first step is based only on LIP (De Mauro et al., 1993), but in future the analysis will be extended to CLIPS (Leoni et al., 2006), C-Coral ROM (Cresti & Moneglia, 2005) and PIXI (Gavioli & Mansfield, 1990). Looking only at transcripts, we lack reliable information on prosody, and it remains to be seen how far prosodic features may also influence the interpretation of connectors and of the clauses they link.

The LIP corpus (queryable online at badip.unigraz.at) contains transcripts of 469 encounters for a total of approximately 500,000 orthographic words, divided into similarly sized components from four geographical areas (Milan, Florence, Rome, Naples). The corpus is part-of-speech tagged, making for a slightly higher number of pos units than the number of orthographic words.

For each geographical area, the corpus contains five types of speech: A, B, C, are two-way encounters (face-to-face and telephone conversations, interviews, etc.: 320,331 pos units); D, E are one-way encounters (lectures, radio monologues, etc.: 203,334 pos units).

In the corpus, the forms *com’* and *come* are tagged either as prepositions (Pz) or conjunctions (C). Table 1 shows their relative frequencies in two-way and one-way encounters.

<table>
<thead>
<tr>
<th></th>
<th>2-way</th>
<th>1-way</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Freq.</td>
<td>Freq/1000 pos units</td>
<td>Freq</td>
</tr>
<tr>
<td>Pz</td>
<td>442</td>
<td>1.38</td>
<td>427</td>
</tr>
<tr>
<td>C</td>
<td>1284</td>
<td>4.01</td>
<td>631</td>
</tr>
<tr>
<td>Tot.</td>
<td>1726</td>
<td>5.39</td>
<td>1058</td>
</tr>
</tbody>
</table>

Table 1: Frequencies of *come/com’* in the LIP corpus

Cases where *come* is tagged as a preposition are relatively straightforward:

Come donna ti senti realizzata o no (As a woman, do you feel realised or not?) (F B 17 61 C)
Volevo sapere come informatica a che punto siamo noi con tutti i programmi (As a computer expert, I wanted to know where we are with all the programmes) (F A 12.5 A)
Eh vedono vedono la loro vita come spezzata e alora ricucirla ci vuol tempo (They see they see their life as torn apart and needing time to put it together again) (F E 15 253 A)

It is more difficult to identify the value of *come* where it is tagged as a conjunction: we manually analysed the occurrences in order to identify the transphrasic relationships involved, distinguishing two-way and one-way
encounters.

Traditional Italian grammars list *come* as a conjunction in the following uses:

- introducing (a) direct interrogatives, (b) indirect interrogatives, (c) completing subordinates:
  
a) ciao *come* va (R B 6 4 B)
  
b) questi condoni non si sa *come* andranno a finire (we don’t know how these new regulations will turn out) (F A 10 82 B)
  
c) il dibattito sull’opinione pubblica vediamo *come* è determinato dalla domanda se è giusto o non giusto la guerra (the debate on public opinion we will see how it is dominated by the question of whether the war is just or unjust.) (M E 8 8 G)

- introducing adverbial clauses which are (d) comparisons or analogies (e) temporal, or (f) causal:
  
d) diceva trattare l’ammalato *come* se fosse la madre *come* se tu infermiere o tu medico fossi sua madre e fosse lui l’unico tuo figlio (as if she was his mother) (M E 12 10 C)
  
e) allora *come* esce [incomprehensible word] dal comune *come* esce lo porta su all’archivio (as soon as he walks out of the office …) (F A 5 1 A)
  
f) ma *come* non è un ragazzo di questo (but since he’s not this kind of boy) (N B 65 23 A)

Some examples, particularly those with adverbial clauses, are however ambiguous, in particular between the causal and the temporal meanings.

The temporal use of *come* is documented since Dante (“‘Si tosto come il vento a noi li piega / mossi la dre e fosse lui l’unico tuo figlio – ma come non è un ragazzo di questo (but since he’s not this kind of boy) (N B 65 23 A)

Some examples, particularly those with adverbial clauses, are however ambiguous, in particular between the causal and the temporal meanings.

The temporal use of *come* is documented since Dante (“‘Si tosto come il vento a noi li piega / mossi la voce …”, *Inferno* V, vv. 81-82). For the dictionary GRADIT, the temporal value belongs to basic Italian (“uso fondamentale”); on the contrary, Serianni (1988) considers it typical of written and especially literary Italian. In LIP the temporal sense appears only in bidirectional encounters, supporting GRADIT’s proposal that it is also a colloquial usage.

As far as concerns the causal value of *come*, GRADIT states that it is relatively infrequent (“basso uso”); similarly, Serianni claims that *come* assumes a causal value only occasionally. In LIP we found fewer causal than temporal examples, some being particularly ambiguous.

The causal interpretation appears to depend on either (a) the contents of the connected propositions; and or (b) position in the dialogue sequence. The following examples illustrate causal linking between connected propositions: in both cases there is some ambiguity between a causal interpretation and one of analogy:

Io penso che gente *come* gioca alle lotterie gioca anche al totocalcio perché insegue proprio il miracolo dl due miliardi del tre miliardi del miliardo (I think people bet on the lottery for the same reasons/in the same ways they bet on the pools) (M E 7 26 A)

Sì ma se tu me seguiti a di’ sempre quando troverò *come* so’ passati circa sette anni ne passeranno altri sette e io non ce sto più allora io vado a fini sotto tera o mezzo a ‘n campo de patate (As about seven years have passed, another seven will) (R E 11 86 D)

The next three examples illustrate the importance of position in the dialogue sequence in suggesting a causal value (in these cases, LIP tags *come* as a preposition, while for other grammars it would be an interrogative adverb). *Come* is used to question the previous affirmation of the other speaker, in the causal sense of “why do you say that?” This is particularly clear in the second example, where speaker A explicitly confirms the causal value of his previous *come* by reformulating it with *perché* in the next utterance:

B: no tesoro non posso
A: *come* non puoi* (why can’t you*)
B: tu non fossi amico di XYZ forse sì ma così non posso (M B 46 356 B)
A: e non lo vendono quella roba lì dal rivenditore
B: *come* non li vendono* (why don’t they sell it*)
A: eh non capisco perché non devono venderlo be’ $$$ ce li ha (M B 70 15 A)
A: *come* ci metto la tuta e vengo [incomprehensible word]
B: ti infili la tuta*
A: *come* vengo in tuta
B: ma che schifo *come* vieni in tuta* (how disgusting why do you come in a tracksuit*)
A: vengo in tuta da ginnastica
B: Bleah
A: ’n ti piace*
B: no (R B 1 120 B)

3. Conclusions

To sum up, our research on a corpus of spoken Italian has provided evidence that the temporal and causal senses of *come* belong to colloquial usage as well as literary Italian. We would argue that these senses of *come* are the result of processes of inferential enrichment. From our point of view, the temporal/causal value of the connector is under-coded, and the attribution of this value derives from inferencing which overlays other relationships. If we see *come* as having a basic modal/comparative meaning, then *come* can encode this kind of relation between two clauses
without considering the contents of the propositions involved. When the contents of the connected propositions allow, however, the meaning of the connector may be enriched by a temporal or a causal value. Such enrichment is possible because – according to the theoretical viewpoint of Prandi (2004) – when we speak of adverbial clauses, we are in an area of the language in which conceptual contents are dominant with respect to grammatical relations.

4. Acknowledgements

I discussed this paper with my colleagues (and friends) Guy Aston and Daniela Zorzi. I would like to thanks both for their great help.

5. References


De Mauro et al. (1993). Lessico di frequenza dell’italiano parlato. Available at: <www.badip.uni-graz.at>.


La variazione dei verbi generali nei corpora di parlato spontaneo. L’ontologia IMAGACT

Massimo MONEGLIA, Gloria GALIARDI, Lorenzo GREGORI, Alessandro PANUNZI, Samuele PALADINI, Andrew WILLIAMS
Università di Firenze (Italia)
moneglia@unifi.it, gloria.gagliardi@unifi.it, lorenzo.gregori@unifi.it, alessandro.panunzi@unifi.it, samuele.paladini@drwolf.it, andrewwilliams.esq@gmail.com

Abstract
I verbi di azione, ad alta frequenza nel parlato, sono molto spesso “generali”, perché si estendono produttivamente ad azioni che individuano oggetti ontologici diversi, ed ogni lingua presenta categorizzazioni idiosincratiche dello spazio ontologico dell’azione. Per questo motivo i verbi d’azione costituiscono un problema per la disambiguazione e per la traduzione delle lingue naturali. Questo lavoro presenta le linee di sviluppo del progetto IMAGACT, che si propone di derivare da corpora di parlato spontaneo multilingui informazioni essenziali sulla categorizzazione linguistica dell’azione, non prevedibili allo stato attuale delle conoscenze. Il progetto utilizza campioni di corpora di parlato spontaneo italiano e inglese, da cui induce l’ambito di variazione produttiva dei circa 500 verbi di azione più alti in frequenza in ciascun corpus. In IMAGACT la variazione si oggettiva in una ontologia interlinguistica le cui entrate sono costituite da scene prototipiche. L’utilizzo del linguaggio universale delle immagini evita problemi di indeterminatezza delle definizioni e facilita sia lo sviluppo, sia lo sfruttamento della base dati.

Keywords: verbi di azione; ontologie; corpora di parlato multilingui.

1. Introduzione
I verbi di azione sono gli elementi più frequenti di strutturazione del discorso parlato e contengono l’informazione essenziale per dare senso agli enunciati (Moneglia & Panunzi, 2007). Ma i verbi d’azione sono anche i tipi linguistici meno prediciibili per i dizionari bilingui e per le tecnologie di traduzione automatica (Moneglia, 2011). Questi verbi, infatti, molto spesso sono “generali”, in quanto si estendono ad azioni appartenenti a differenti tipi ontologici. Per esempio in inglese ed italiano i verbi ad alta frequenza to put e mettere appartengono a questa categoria. La Tabella 1 esemplifica la varietà di atti che ricadono nella loro estensione. In 1 ad un oggetto è data locazione, in 2 un oggetto è dotato di attributi funzionali, in 3 un oggetto è modificato, in 4 una parte del corpo assume una posizione.

La diversità sostanziale tra i tipi di atti riferiti dal verbo, evidenziata dalla figura, è marcatamente linguistica dalla possibilità di identificare ciascuna azione con verbi equivalenti diversi, che si applicano in modo differenziale a ciascun tipo (collocare, inserire, aggiungere, alzare).

Malgrado una forte relazione di traduzione, to put e mettere non sono però coestensivi, dal momento che to put può essere esteso a 4, ma non mettere.

Questa differenza, individuata in seguito a lavoro su corpus, non è chiaramente identificata allo stato attuale delle conoscenze sul lessico verbale d’Azione ed è un esempio delle ragioni cruciali per cui le predicazioni del sottalbero naturale non sono idonee alla traduzione automatica. non sono identificate le entità ontologiche a cui i verbi d’azione si riferiscono nelle frasi semplici e non vi è quindi garanzia che due predicati in un dizionario bilingue selezionino la stessa entità.

Ogni lingua, con la sua varietà di verbi generali, categorizza l’azione in un modo specifico e perciò il riferimento cross-linguistico alle attività di ogni giorno risulta scarsamente prevedibile (Moneglia & Panunzi, 2007).

<table>
<thead>
<tr>
<th>ACTION TYPE</th>
<th>INSTANCES</th>
<th>EQUIVALENT VERBS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type 1</strong></td>
<td>John puts the glass on the table</td>
<td>to locate</td>
</tr>
<tr>
<td></td>
<td>John mette il bicchiere sul tavolo</td>
<td>colloca</td>
</tr>
<tr>
<td><strong>Type 2</strong></td>
<td>John puts the cap on the pen</td>
<td>to fasten</td>
</tr>
<tr>
<td></td>
<td>John mette il tappo alla penna</td>
<td>inserire</td>
</tr>
<tr>
<td><strong>Type 3</strong></td>
<td>John puts water into the whisky</td>
<td>to add</td>
</tr>
<tr>
<td></td>
<td>John mette l’acqua nel whisky</td>
<td>aggiungere</td>
</tr>
<tr>
<td><strong>Type 4</strong></td>
<td>*Mary mette su la mano</td>
<td>to raise</td>
</tr>
<tr>
<td></td>
<td>Mary puts her hand up</td>
<td></td>
</tr>
</tbody>
</table>

Tabella 1: Tipi azionali dei verbi to put e mettere
E’ rilevante notare che tale variazione cross-linguistica non è dovuta alle fraseologie proprie di ogni lingua, ma è conseguenza del modo peculiare con cui le lingue categorizzano gli eventi, ovvero deriva da fattori semantic (Moneglia, 1998; Majid et al., 2008).

Infatti l’applicazione dei verbi generali ai tipi azionali nella loro estensione è produttiva: in qualunque evento del tipo 1 to put sarà tradotto in Italiano con mettere, e in nessuna istanza del tipo 4 il verbo Inglese to put, risulterà traducibile in Italiano con mettere, come mostrano i seguenti esempi:

(1) John puts a glass / a pot / a dress on the table / on the stove / on the harm chair

(1’) John mette un bicchiere / la pentola / sul tavolo / sul fornello / sulla poltrona

(2) Mary puts her hand / her finger / her leg / up / aside / down

(2’) *Mary mette la mano / il dito / la gamba / su / di lato / giù

Se l’applicazione di un verbo ad un tipo è produttiva, dovrebbe in linea di principio essere anche predicibile: il range di variazioni produttive dei verbi generali nelle diverse lingue è però, al momento, largamente sconosciuto; non è chiara, inoltre, la distinzione tra variazioni produttive e variazioni non produttive nell’estensione dei verbi generali.

Le risorse esistenti, e in particolare WordNet, che costituisce la principale e più ricca base di dati lessicale oggi disponibile (Fellbaum, 1998), non contengono informazione sufficiente a questo scopo per una varietà di ragioni (Moneglia et al., 2012). Per esempio il numero di tipi (synset) registrati per ciascuna entrata è alto ma, non essendo la risorsa derivata da corpora, i significati periferici non sono distinti da quelli con alta probabilità di occorrenza. Inoltre, per lo stesso motivo, non esiste certezza che le variazioni principali di un verbo generale nell’uso linguistico siano censite. In aggiunta, le descrizioni date per ciascun synset sono vaghe e difficili da utilizzare perfino da annotatori esperti (Ng et al., 1999).

Più in generale deve essere notato un problema teorico che affligge le risorse che riflettono la varietà dell’uso linguistico e rendono poco prevedibile la possibilità di traduzione, ovvero che la produttività dell’applicazione del verbo non può essere garantita da tutti i synset nella stessa misura. I verbi hanno infatti vari usi che si distaccano dal loro significato effettivo, ed in questi significati la relazione di traduzione non può essere predetta.

Ad esempio, tra i synset di WordNet del verbo to put è riportato il seguente:

S: (v) arrange, set up, put, order (arrange thoughts, ideas, temporal events)

In questa entrata dell’ontologia, diversamente da quanto avviene in (1) e (2), la possibilità di traduzione non corre in parallelo in tutte le istanze del tipo. Funziona in (3), ma per qualche ragione idiosincratica non in (4):

(3) I put my schedule in a certain way > Ho messo i miei impegni in un certo modo

(4) *I put my life in a certain way > * Ho messa la mia vita in un certo modo

La distinzione tra tipi produttivi e tipi idiosincratici è cruciale: solo gli usi primari (come quelli nella Tabella 1) sono sicuramente produttivi, mentre gli usi fraseologici o metaforici spesso non lo sono. In altri termini, mentre la variazione in Tabella 1 identifica le variazioni in estensione su tipi di azioni diverse che un parlante nativo deve poter assentire o rifiutare sulla base della sua sola competenza linguistica, lo stesso non vale per usi marcati come in (3). Solo l’identificazione degli usi produttivi costituisce una base di conoscenza per la previsione degli ambiti di estensione dei verbi di lingue diverse nello spazio dell’azione e per rendere obiettive le relazioni di traduzione.

Il progetto IMAGACT utilizza metodologie corpus-based e competence-based per l’estrazione simultanea da risorse multilingui di parlato spontaneo di una ontologia dell’azione indipendente dal linguaggio, e permetterà la disambiguazione dei verbi di azione ad alta frequenza nel parlato rispetto ai tipi azionali in cui una applicazione produttiva può essere prevista.

Questo lavoro descrive le caratteristiche chiave del progetto. Il paragrafo 2. mostrerà la strategia corpus-based scelta per l’induzione delle proprietà variazionali dei verbi d’azione e presenterà in allegato le entrate verbali oggetto di analisi; il paragrafo 3. illustrerà, sulla base di un esempio concreto (la variazione di to roll in inglese e parallelamente la variazione di rotolare e arrotole in italiano), la metodologia di costruzione dell’ontologia interlinguistica, specificamente basata sull’utilizzo dell’immagine.

2. Lo sfruttamento di risorse di parlato spontaneo

Le azioni specificate dai verbi usati con maggior frequenza nella comunicazione quotidiana sono anche le azioni più rilevanti per le nostre attività di ogni giorno e, in quanto tali, costituiscono l’universo di riferimento per il linguaggio. L’uso effettivo di tali verbi può pertanto essere apprezzato nella performance linguistica mediante l’osservazione delle loro occorrenze nel parlato spontaneo, in cui il riferimento all’azione è primario. I corpora di parlato spontaneo pubblicati negli ultimi due decenni sono sfruttati in IMAGACT a questo fine: la variazione di un set di predicati generali verrà infatti identificata nel corpus BNC (sezione di parlato) e, in parallelo, in una collezione di corpora italiani (C-ORAL-ROM; LABLITA, LIP, CLIPS).

IMAGACT si focalizza sui verbi ad alta probabilità di occorrenza, ovvero i 500 verbi di azione più alti in rank nelle liste di frequenza, che rappresentano il lessico
verbale di base nelle due lingue. Un’ampia selezione di questo lessico è riportata nella liste di frequenza disponibili in appendice.

Saranno annotate attraverso una infrastruttura web circa 50.000 occorrenze per lingua, derivate da un campione di 2 milioni di parole di entrambi i corpora.

Gli enunciati in cui le occorrenze compaiono nei corpora, necessariamente frammentari dal punto di vista semantico, vengono interpretati da annotatori madrelingua e ricondotti a frasi semplici nelle quali è saturata la struttura valenziale e da cui l’azione riferita risulta in modo trasparente. La presenza di una serie ampia di frasi semplici derivate dall’uso orale consente di individuare i punti essenziali della variazione d’uso di ciascun verbo e di raggrupparne in tipi gli usi produttivi.

A tal fine è adottata una metodologia specifica e una procedura di annotazione guidata dall’infrastruttura web IMAGACT a disposizione degli annotatori.

3. Formazione dell’ontologia interlinguistica dell’azione e immagine. Uno scenario “alla Wittgenstein”

Lavorando con più di una lingua, IMAGACT deve produrre un inventario di tipi language-independent. Precedenti esperienze nella costituzione di Ontologie hanno evidenziato però che il livello di consenso raggiungibile nella definizione delle entità riferite dalle espressioni linguistiche è generalmente basso, e che l’accordo nell’annotazione varia in relazione alla granularità semantica dei sensi (Brown et al., 2010).

L’innovazione chiave di IMAGACT è di fornire una metodologia che sfruti la capacità, indipendente dal linguaggio, di apprezzare somiglianze tra scene, distinguendo di fatto l’Identificazione dei tipi azionali dalla loro Definizione.

Ad esempio, la distinzione tra i tipi 1-4 nella Tabella 1 è rilevante per prevedere la variazione cross-linguistica dei concetti azionali. La differenza tra i tipi è facilmente riconosciuta dai parlanti e non richiede la definizione di un set di caratteristiche differenziali, che sono, come si diceva, radicalmente sottodeterminate.

Crucialmente solo l’identificazione, e non la definizione delle entità individuate, è richiesta per stabilire le relazioni cross-linguistiche.

In termini Wittgensteiniani: come posso spiegare a qualcuno cos’è un gioco? Semplicemente indicando un gioco e dicendo “Questo e simili cose sono giochi” (Wittgenstein, 1953).

Lo scenario “alla Wittgenstein” è utilizzato in IMAGACT sia per distinguere le variazioni produttive dalle variazioni non produttive all’interno dell’uso linguistico dei verbi, sia per identificare tipi azionali a livello cross-linguistico, consentendo la comparazione diretta dei tipi derivati dall’annotazione dei corpora di lingue diverse.

Per l’induzione della variazione semantica dei verbi di azione dai corpora di parlato italiano e inglese IMAGACT si sviluppa sui seguenti passi:

- distinguere gli usi primari dagli usi marcati;
- identificare in ciascun corpus di parlato i punti focali di variazione dei verbi generali su tipi di azione diversi;
- rappresentare i concetti azionali attraverso scene prototipiche a cui rapportare la variazione riscontrata nei verbi delle due lingue.

3.1 Variazione primaria vs. Variazione marcata

Il primo compito sfrutta lo scenario “alla Wittgenstein” come banco di prova della effettiva produttività dei concetti. Si deve notare, infatti, che solo gli usi che ad un parlante competente appaiono adeguati a rappresentare il significato di un predicato possono essere indicati come prototipi per l’uso del predicato stesso. In parallelo, gli usi non primari o comunque metaforici o fraseologici non possono essere indicati come istance prototipiche di ciò che viene significato.

Si consideri ad esempio il verbo italiano rollolare. L’istanza (5), derivata dal corpus, può essere ragionevolmente indicata come una istance prototipica del concetto espresso dal verbo, in altri termini un parlante competente può indicare l’istanza a qualcuno che non conosce la lingua fornendo l’informazione: “questa e simili cose sono ciò che noi intendiamo con rollolare”. Al contrario, l’istanza (6) non potrà ragionevolmente essere indicata come un’istanza di “ciò che noi intendiamo con rollolare”.

(5) Cristina si rotola nell’erba umida
(6) Il bambino rotolò in terra dal soggionlo

Infatti, nonostante la frequenza con cui può comparire in quel contesto, in (6) il verbo è usato palesemente in senso non proprio (il bambino non rotolare, bensì cade). Ciò risulta evidente ad un parlante competente. Il test consente quindi, salvo casi limite, di isolare la gran parte degli usi strettamente propri del verbo, identificando poi la loro variazione.

Lo stesso avverrà con le frasi derivate dal corpus inglese. Ad esempio, per quanto riguarda la variazione del verbo to roll (7), potrà essere indicata come un’istanza prototipica di ciò che si intende con to roll, ma non (8).

(7) John rolls a cigarette
(8) John rolls the words around in his mind

Lo studio della variazione produttiva di un verbo inizia quando gli usi non produttivi sono esclusi dal campo di analisi.

3.2 Variazione verticale vs. variazione orizzontale

La variazione dei verbi generali si configura in modo simile a quanto ipotizzato originariamente da Wittgenstein, ovvero l’uso si raccoglie in una serie di famiglie, ciascuna delle quali contiene variazioni granulari rappresentabili ad una istance prototipica (Givon, 1986). Ogni concetto istanziato da un prototipo è
produttivo e distinto dagli altri dal punto di vista cognitivo, nonostante lo stesso verbo si applichi a tutte le famiglie (proprietà per cui il verbo si dice “generale”). A tale variazione si unisce poi la variazione non produttiva, non identificata nel lavoro originale del filosofo, che ovviamente non definisce entrate nell’ontologia.

L’annotazione del verbo inglese to roll e dei verbi italiani apparentemente in relazione di traduzione con questo, ovvero arrotolare e rotolare, può essere riassunta in breve nelle tabelle seguenti derivate dalla annotazione dei corpora attraverso l’infrastruttura IMAGACT. Nel corpus sono identificati una serie di tipi (variazione verticale del verbo), ognuno dei quali contiene una serie di istanze (variazione orizzontale del tipo).

### Tabella 2: Tipi azionali del verbo to roll

<table>
<thead>
<tr>
<th>Type</th>
<th>John rolls his sleeve up</th>
<th>John rolls a cigarette</th>
<th>The sailors roll the sail up</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type 1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type</th>
<th>The horse rolls around the field</th>
<th>Mary rolls onto her side</th>
<th>John rolls along the floor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type 2</td>
<td>The horse rolls around the field</td>
<td>Mary rolls onto her side</td>
<td>John rolls along the floor</td>
</tr>
<tr>
<td>Type 3</td>
<td>John rolls the barrel along the floor</td>
<td>John rolls the girl onto her side</td>
<td>John rolls the thread around</td>
</tr>
<tr>
<td>Type 4</td>
<td>John rolls the ball across the room</td>
<td>John rolls the wheel into the scrapheap</td>
<td>John rolls the apple across the table to Mary</td>
</tr>
<tr>
<td>Type 5</td>
<td>John rolls his ankle around</td>
<td>John rolls his eyes</td>
<td>John rolls his wrist around its socket</td>
</tr>
<tr>
<td>Type 6</td>
<td>The car rolls into the fence</td>
<td>The ball rolls over to the wall</td>
<td>The car rolls into the lake</td>
</tr>
<tr>
<td>Type 7</td>
<td>John rolls the clay in his hands</td>
<td>John rolls the dough into a ball</td>
<td>John rolls the playdoh on the table</td>
</tr>
</tbody>
</table>

### Tabella 3: Tipi azionali del verbo arrotolare

<table>
<thead>
<tr>
<th>Tipo</th>
<th>Cristina arrotola il filo intorno alla ruota</th>
<th>Cristina arrotola la benda intorno al braccio</th>
<th>Fabio arrotola la corda intorno alla gamba</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tipo 1</td>
<td>Cristina arrotola una sigaretta</td>
<td>Cristina arrotola il poster</td>
<td>Cristina arrotola il filo</td>
</tr>
<tr>
<td>Tipo 2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Tabella 4: Tipi azionali del verbo rotolare

<table>
<thead>
<tr>
<th>Tipo</th>
<th>Matteo si rotola per terra</th>
<th>Cristina si rotola nell’erba umida</th>
<th>Fabio e Cristina si rotolano</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tipo 1</td>
<td>La ciambella di gomma rotola</td>
<td>L’arancia rotola</td>
<td>Il cilindro rotola</td>
</tr>
<tr>
<td>Tipo 2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Dopo la procedura di annotazione dei corpora, IMAGACT rilascerà un database di tipi azionali associati alla loro codifica linguistica in inglese e in italiano. L’insieme delle frasi derivate da corpora istanzieranno ogni tipo rappresentato.

#### 3.3 Immagine e Ontologia Cross-linguistica

Sulla base dell’induzione della variazione verticale across-types dei verbi di azione nei corpora, IMAGACT fa uso del linguaggio universale delle immagini per riconciliare in una sola ontologia i tipi derivati dall’annotazione di corpora di diverse lingue. Ad esempio i tipi estratti dalla annotazione di to roll sono rappresentati dalle scene B-H, come in Figura 1 di seguito.

La costituzione delle scene permette una rappresentazione dell’universo dell’azione valido indipendentemente dalla lingua. Per cui, a livello della costituzione dell’ontologia cross-linguistica sulla base dei dati derivati da corpus, si scoprirà che la scena B è estesa anche dal tipo 2 del verbo italiano arrotolare, e che i tipi 1 e 2 del verbo rotolare estendono rispettivamente sui tipi C e G.

Nell’insieme possiamo osservare che la variazione del verbo inglese to roll è più ampio rispetto alle sue controparti italiane, dato che i due verbi italiani in linea teorica corrispondenti a questo verbo inglese (arrotolare e rotolare) trovano applicazione solo in un sottoinsieme dei tipi azionali estesi da to roll.

Il differenziale nel significato sarà ulteriormente evidenziato nel momento in cui, dovendo identificare una scena per il tipo 1 di arrotolare (il tipo A di Figura 1) diventerà evidente che c’è almeno un tipo esteso da arrotolare che non è una possibile estensione di to roll. La relazione cross-linguistica risulta quindi in una intersezione tra tipi.

La corrispondenza tra tipi derivati da differenti corpora linguistiche seguirà perciò dal riferimento dei tipi estratti dai corpora alla stessa galleria di scene. Questo risultato è ottenuto senza far ricorso alla comparazione tra definizioni date da differenti annotatori: identificare la corrispondenza cross-linguistica dei verbi d’azione su una ontologia language-indipendent, aggiura la sottodeterminazione delle definizioni.

IMAGACT rilascerà una base dati di tipi azionali individuati nel riferimento linguistico alle azioni quotidiane attraverso la rappresentazione di scene prototipiche. Ogni scena sarà associata a uno o più verbi italiani e inglesi che risulteranno in relazione di traduzione stretta in tutte le istanze del tipo.

IMAGACT renderà chiaro sia l’ambito di variazione dei predicati generali nelle lingue considerate, sia il differenziale semantico tra entrate lessicali a livello cross-linguistico e permetterà di basare processi di disambiguazione e traduzione su tipi ontologici produttivi oltrecché rilevanti in quanto derivati da corpora rappresentativi dell’uso linguistico quotidiano.
4. Referimenti


CLIPS Corpus. Available at: <http://www.clips.unina.it>.


IMAGACT. Available at: <http://www.imagact.it/>.

LABLITA Corpus of Spontaneous Spoken Italian. Available at: <http://lablita.dit.unifi.it/corpora/>.


5. Appendice

Tabella 5: Verbi italiani di azione ad alta frequenza

<table>
<thead>
<tr>
<th>LEMMA</th>
<th>FREQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>leggere</td>
<td>163</td>
</tr>
<tr>
<td>scrivere</td>
<td>158</td>
</tr>
<tr>
<td>parlare</td>
<td>157</td>
</tr>
<tr>
<td>bere</td>
<td>156</td>
</tr>
<tr>
<td>mangiare</td>
<td>155</td>
</tr>
<tr>
<td>dormire</td>
<td>154</td>
</tr>
<tr>
<td>bere</td>
<td>153</td>
</tr>
<tr>
<td>lavorare</td>
<td>152</td>
</tr>
<tr>
<td>mangiare</td>
<td>151</td>
</tr>
<tr>
<td>bere</td>
<td>150</td>
</tr>
<tr>
<td>parlare</td>
<td>149</td>
</tr>
<tr>
<td>leggere</td>
<td>148</td>
</tr>
<tr>
<td>scrivere</td>
<td>147</td>
</tr>
<tr>
<td>dormire</td>
<td>146</td>
</tr>
<tr>
<td>bere</td>
<td>145</td>
</tr>
<tr>
<td>parlare</td>
<td>144</td>
</tr>
<tr>
<td>leggere</td>
<td>143</td>
</tr>
<tr>
<td>scrivere</td>
<td>142</td>
</tr>
<tr>
<td>mangiare</td>
<td>141</td>
</tr>
<tr>
<td>bere</td>
<td>140</td>
</tr>
<tr>
<td>parlare</td>
<td>139</td>
</tr>
<tr>
<td>leggere</td>
<td>138</td>
</tr>
<tr>
<td>scrivere</td>
<td>137</td>
</tr>
<tr>
<td>mangiare</td>
<td>136</td>
</tr>
<tr>
<td>bere</td>
<td>135</td>
</tr>
<tr>
<td>parlare</td>
<td>134</td>
</tr>
<tr>
<td>leggere</td>
<td>133</td>
</tr>
<tr>
<td>scrivere</td>
<td>132</td>
</tr>
<tr>
<td>mangiare</td>
<td>131</td>
</tr>
<tr>
<td>bere</td>
<td>130</td>
</tr>
<tr>
<td>parlare</td>
<td>129</td>
</tr>
<tr>
<td>leggere</td>
<td>128</td>
</tr>
<tr>
<td>scrivere</td>
<td>127</td>
</tr>
<tr>
<td>mangiare</td>
<td>126</td>
</tr>
<tr>
<td>bere</td>
<td>125</td>
</tr>
<tr>
<td>parlare</td>
<td>124</td>
</tr>
<tr>
<td>leggere</td>
<td>123</td>
</tr>
<tr>
<td>scrivere</td>
<td>122</td>
</tr>
<tr>
<td>mangiare</td>
<td>121</td>
</tr>
<tr>
<td>bere</td>
<td>120</td>
</tr>
<tr>
<td>parlare</td>
<td>119</td>
</tr>
<tr>
<td>leggere</td>
<td>118</td>
</tr>
<tr>
<td>scrivere</td>
<td>117</td>
</tr>
<tr>
<td>mangiare</td>
<td>116</td>
</tr>
<tr>
<td>bere</td>
<td>115</td>
</tr>
<tr>
<td>parlare</td>
<td>114</td>
</tr>
<tr>
<td>leggere</td>
<td>113</td>
</tr>
<tr>
<td>scrivere</td>
<td>112</td>
</tr>
<tr>
<td>mangiare</td>
<td>111</td>
</tr>
<tr>
<td>bere</td>
<td>110</td>
</tr>
<tr>
<td>parlare</td>
<td>109</td>
</tr>
<tr>
<td>leggere</td>
<td>108</td>
</tr>
<tr>
<td>scrivere</td>
<td>107</td>
</tr>
<tr>
<td>mangiare</td>
<td>106</td>
</tr>
<tr>
<td>bere</td>
<td>105</td>
</tr>
<tr>
<td>parlare</td>
<td>104</td>
</tr>
<tr>
<td>leggere</td>
<td>103</td>
</tr>
<tr>
<td>scrivere</td>
<td>102</td>
</tr>
<tr>
<td>mangiare</td>
<td>101</td>
</tr>
<tr>
<td>bere</td>
<td>100</td>
</tr>
<tr>
<td>parlare</td>
<td>99</td>
</tr>
<tr>
<td>leggere</td>
<td>98</td>
</tr>
<tr>
<td>scrivere</td>
<td>97</td>
</tr>
<tr>
<td>mangiare</td>
<td>96</td>
</tr>
<tr>
<td>bere</td>
<td>95</td>
</tr>
<tr>
<td>parlare</td>
<td>94</td>
</tr>
<tr>
<td>leggere</td>
<td>93</td>
</tr>
<tr>
<td>scrivere</td>
<td>92</td>
</tr>
<tr>
<td>mangiare</td>
<td>91</td>
</tr>
<tr>
<td>bere</td>
<td>90</td>
</tr>
<tr>
<td>parlare</td>
<td>89</td>
</tr>
<tr>
<td>leggere</td>
<td>88</td>
</tr>
<tr>
<td>scrivere</td>
<td>87</td>
</tr>
<tr>
<td>mangiare</td>
<td>86</td>
</tr>
<tr>
<td>bere</td>
<td>85</td>
</tr>
<tr>
<td>parlare</td>
<td>84</td>
</tr>
<tr>
<td>leggere</td>
<td>83</td>
</tr>
<tr>
<td>scrivere</td>
<td>82</td>
</tr>
<tr>
<td>mangiare</td>
<td>81</td>
</tr>
<tr>
<td>bere</td>
<td>80</td>
</tr>
<tr>
<td>parlare</td>
<td>79</td>
</tr>
<tr>
<td>leggere</td>
<td>78</td>
</tr>
<tr>
<td>scrivere</td>
<td>77</td>
</tr>
<tr>
<td>mangiare</td>
<td>76</td>
</tr>
<tr>
<td>bere</td>
<td>75</td>
</tr>
<tr>
<td>parlare</td>
<td>74</td>
</tr>
<tr>
<td>leggere</td>
<td>73</td>
</tr>
<tr>
<td>scrivere</td>
<td>72</td>
</tr>
<tr>
<td>mangiare</td>
<td>71</td>
</tr>
<tr>
<td>bere</td>
<td>70</td>
</tr>
<tr>
<td>parlare</td>
<td>69</td>
</tr>
<tr>
<td>leggere</td>
<td>68</td>
</tr>
<tr>
<td>scrivere</td>
<td>67</td>
</tr>
<tr>
<td>mangiare</td>
<td>66</td>
</tr>
<tr>
<td>bere</td>
<td>65</td>
</tr>
<tr>
<td>parlare</td>
<td>64</td>
</tr>
<tr>
<td>leggere</td>
<td>63</td>
</tr>
<tr>
<td>scrivere</td>
<td>62</td>
</tr>
<tr>
<td>mangiare</td>
<td>61</td>
</tr>
<tr>
<td>bere</td>
<td>60</td>
</tr>
<tr>
<td>parlare</td>
<td>59</td>
</tr>
<tr>
<td>leggere</td>
<td>58</td>
</tr>
<tr>
<td>scrivere</td>
<td>57</td>
</tr>
<tr>
<td>mangiare</td>
<td>56</td>
</tr>
<tr>
<td>bere</td>
<td>55</td>
</tr>
<tr>
<td>parlare</td>
<td>54</td>
</tr>
<tr>
<td>leggere</td>
<td>53</td>
</tr>
<tr>
<td>scrivere</td>
<td>52</td>
</tr>
<tr>
<td>mangiare</td>
<td>51</td>
</tr>
<tr>
<td>bere</td>
<td>50</td>
</tr>
<tr>
<td>parlare</td>
<td>49</td>
</tr>
<tr>
<td>leggere</td>
<td>48</td>
</tr>
<tr>
<td>scrivere</td>
<td>47</td>
</tr>
<tr>
<td>mangiare</td>
<td>46</td>
</tr>
<tr>
<td>bere</td>
<td>45</td>
</tr>
<tr>
<td>parlare</td>
<td>44</td>
</tr>
<tr>
<td>leggere</td>
<td>43</td>
</tr>
<tr>
<td>scrivere</td>
<td>42</td>
</tr>
<tr>
<td>mangiare</td>
<td>41</td>
</tr>
<tr>
<td>bere</td>
<td>40</td>
</tr>
<tr>
<td>parlare</td>
<td>39</td>
</tr>
<tr>
<td>leggere</td>
<td>38</td>
</tr>
<tr>
<td>scrivere</td>
<td>37</td>
</tr>
<tr>
<td>mangiare</td>
<td>36</td>
</tr>
<tr>
<td>bere</td>
<td>35</td>
</tr>
<tr>
<td>parlare</td>
<td>34</td>
</tr>
<tr>
<td>leggere</td>
<td>33</td>
</tr>
<tr>
<td>scrivere</td>
<td>32</td>
</tr>
<tr>
<td>mangiare</td>
<td>31</td>
</tr>
<tr>
<td>bere</td>
<td>30</td>
</tr>
<tr>
<td>parlare</td>
<td>29</td>
</tr>
<tr>
<td>leggere</td>
<td>28</td>
</tr>
<tr>
<td>scrivere</td>
<td>27</td>
</tr>
<tr>
<td>mangiare</td>
<td>26</td>
</tr>
<tr>
<td>bere</td>
<td>25</td>
</tr>
<tr>
<td>parlare</td>
<td>24</td>
</tr>
<tr>
<td>leggere</td>
<td>23</td>
</tr>
<tr>
<td>scrivere</td>
<td>22</td>
</tr>
<tr>
<td>mangiare</td>
<td>21</td>
</tr>
<tr>
<td>bere</td>
<td>20</td>
</tr>
<tr>
<td>parlare</td>
<td>19</td>
</tr>
<tr>
<td>leggere</td>
<td>18</td>
</tr>
<tr>
<td>scrivere</td>
<td>17</td>
</tr>
<tr>
<td>mangiare</td>
<td>16</td>
</tr>
<tr>
<td>bere</td>
<td>15</td>
</tr>
<tr>
<td>parlare</td>
<td>14</td>
</tr>
<tr>
<td>leggere</td>
<td>13</td>
</tr>
<tr>
<td>scrivere</td>
<td>12</td>
</tr>
<tr>
<td>mangiare</td>
<td>11</td>
</tr>
<tr>
<td>bere</td>
<td>10</td>
</tr>
<tr>
<td>parlare</td>
<td>9</td>
</tr>
<tr>
<td>leggere</td>
<td>8</td>
</tr>
<tr>
<td>scrivere</td>
<td>7</td>
</tr>
<tr>
<td>mangiare</td>
<td>6</td>
</tr>
<tr>
<td>bere</td>
<td>5</td>
</tr>
<tr>
<td>parlare</td>
<td>4</td>
</tr>
<tr>
<td>leggere</td>
<td>3</td>
</tr>
<tr>
<td>scrivere</td>
<td>2</td>
</tr>
<tr>
<td>mangiare</td>
<td>1</td>
</tr>
<tr>
<td>bere</td>
<td>0</td>
</tr>
</tbody>
</table>

Tabella 6: Verbi inglesi di azione ad alta frequenza

<table>
<thead>
<tr>
<th>LEMMA</th>
<th>FREQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>take</td>
<td>163</td>
</tr>
<tr>
<td>put</td>
<td>158</td>
</tr>
<tr>
<td>work</td>
<td>157</td>
</tr>
<tr>
<td>keep</td>
<td>156</td>
</tr>
<tr>
<td>feel</td>
<td>155</td>
</tr>
<tr>
<td>look</td>
<td>154</td>
</tr>
<tr>
<td>know</td>
<td>153</td>
</tr>
<tr>
<td>use</td>
<td>152</td>
</tr>
<tr>
<td>move</td>
<td>151</td>
</tr>
<tr>
<td>play</td>
<td>150</td>
</tr>
<tr>
<td>read</td>
<td>149</td>
</tr>
<tr>
<td>write</td>
<td>148</td>
</tr>
<tr>
<td>see</td>
<td>147</td>
</tr>
<tr>
<td>hear</td>
<td>146</td>
</tr>
<tr>
<td>speak</td>
<td>145</td>
</tr>
<tr>
<td>listen</td>
<td>144</td>
</tr>
<tr>
<td>talk</td>
<td>143</td>
</tr>
<tr>
<td>listen</td>
<td>142</td>
</tr>
<tr>
<td>speak</td>
<td>141</td>
</tr>
<tr>
<td>listen</td>
<td>140</td>
</tr>
<tr>
<td>speak</td>
<td>139</td>
</tr>
<tr>
<td>listen</td>
<td>138</td>
</tr>
<tr>
<td>speak</td>
<td>137</td>
</tr>
<tr>
<td>listen</td>
<td>136</td>
</tr>
<tr>
<td>speak</td>
<td>135</td>
</tr>
<tr>
<td>listen</td>
<td>134</td>
</tr>
<tr>
<td>speak</td>
<td>133</td>
</tr>
<tr>
<td>listen</td>
<td>132</td>
</tr>
<tr>
<td>speak</td>
<td>131</td>
</tr>
<tr>
<td>listen</td>
<td>130</td>
</tr>
<tr>
<td>speak</td>
<td>129</td>
</tr>
<tr>
<td>listen</td>
<td>128</td>
</tr>
<tr>
<td>speak</td>
<td>127</td>
</tr>
<tr>
<td>listen</td>
<td>126</td>
</tr>
<tr>
<td>speak</td>
<td>125</td>
</tr>
<tr>
<td>listen</td>
<td>124</td>
</tr>
<tr>
<td>speak</td>
<td>123</td>
</tr>
<tr>
<td>listen</td>
<td>122</td>
</tr>
<tr>
<td>speak</td>
<td>121</td>
</tr>
<tr>
<td>listen</td>
<td>120</td>
</tr>
<tr>
<td>speak</td>
<td>119</td>
</tr>
<tr>
<td>listen</td>
<td>118</td>
</tr>
<tr>
<td>speak</td>
<td>117</td>
</tr>
<tr>
<td>listen</td>
<td>116</td>
</tr>
<tr>
<td>speak</td>
<td>115</td>
</tr>
<tr>
<td>listen</td>
<td>114</td>
</tr>
<tr>
<td>speak</td>
<td>113</td>
</tr>
<tr>
<td>listen</td>
<td>112</td>
</tr>
<tr>
<td>speak</td>
<td>111</td>
</tr>
<tr>
<td>listen</td>
<td>110</td>
</tr>
<tr>
<td>speak</td>
<td>109</td>
</tr>
<tr>
<td>listen</td>
<td>108</td>
</tr>
<tr>
<td>speak</td>
<td>107</td>
</tr>
<tr>
<td>listen</td>
<td>106</td>
</tr>
<tr>
<td>speak</td>
<td>105</td>
</tr>
<tr>
<td>listen</td>
<td>104</td>
</tr>
<tr>
<td>speak</td>
<td>103</td>
</tr>
<tr>
<td>listen</td>
<td>102</td>
</tr>
<tr>
<td>speak</td>
<td>101</td>
</tr>
<tr>
<td>listen</td>
<td>100</td>
</tr>
<tr>
<td>speak</td>
<td>99</td>
</tr>
<tr>
<td>listen</td>
<td>98</td>
</tr>
<tr>
<td>speak</td>
<td>97</td>
</tr>
<tr>
<td>listen</td>
<td>96</td>
</tr>
<tr>
<td>speak</td>
<td>95</td>
</tr>
<tr>
<td>listen</td>
<td>94</td>
</tr>
<tr>
<td>speak</td>
<td>93</td>
</tr>
<tr>
<td>listen</td>
<td>92</td>
</tr>
<tr>
<td>speak</td>
<td>91</td>
</tr>
<tr>
<td>listen</td>
<td>90</td>
</tr>
<tr>
<td>speak</td>
<td>89</td>
</tr>
<tr>
<td>listen</td>
<td>88</td>
</tr>
<tr>
<td>speak</td>
<td>87</td>
</tr>
<tr>
<td>listen</td>
<td>86</td>
</tr>
<tr>
<td>speak</td>
<td>85</td>
</tr>
<tr>
<td>listen</td>
<td>84</td>
</tr>
<tr>
<td>speak</td>
<td>83</td>
</tr>
<tr>
<td>listen</td>
<td>82</td>
</tr>
<tr>
<td>speak</td>
<td>81</td>
</tr>
<tr>
<td>listen</td>
<td>80</td>
</tr>
<tr>
<td>speak</td>
<td>79</td>
</tr>
<tr>
<td>listen</td>
<td>78</td>
</tr>
<tr>
<td>speak</td>
<td>77</td>
</tr>
<tr>
<td>listen</td>
<td>76</td>
</tr>
<tr>
<td>speak</td>
<td>75</td>
</tr>
<tr>
<td>listen</td>
<td>74</td>
</tr>
<tr>
<td>speak</td>
<td>73</td>
</tr>
<tr>
<td>listen</td>
<td>72</td>
</tr>
<tr>
<td>speak</td>
<td>71</td>
</tr>
<tr>
<td>listen</td>
<td>70</td>
</tr>
<tr>
<td>speak</td>
<td>69</td>
</tr>
<tr>
<td>listen</td>
<td>68</td>
</tr>
<tr>
<td>speak</td>
<td>67</td>
</tr>
<tr>
<td>listen</td>
<td>66</td>
</tr>
<tr>
<td>speak</td>
<td>65</td>
</tr>
<tr>
<td>listen</td>
<td>64</td>
</tr>
<tr>
<td>speak</td>
<td>63</td>
</tr>
<tr>
<td>listen</td>
<td>62</td>
</tr>
<tr>
<td>speak</td>
<td>61</td>
</tr>
<tr>
<td>listen</td>
<td>60</td>
</tr>
<tr>
<td>speak</td>
<td>59</td>
</tr>
<tr>
<td>listen</td>
<td>58</td>
</tr>
<tr>
<td>speak</td>
<td>57</td>
</tr>
<tr>
<td>listen</td>
<td>56</td>
</tr>
<tr>
<td>speak</td>
<td>55</td>
</tr>
<tr>
<td>listen</td>
<td>54</td>
</tr>
<tr>
<td>speak</td>
<td>53</td>
</tr>
<tr>
<td>listen</td>
<td>52</td>
</tr>
<tr>
<td>speak</td>
<td>51</td>
</tr>
<tr>
<td>listen</td>
<td>50</td>
</tr>
<tr>
<td>speak</td>
<td>49</td>
</tr>
<tr>
<td>listen</td>
<td>48</td>
</tr>
<tr>
<td>speak</td>
<td>47</td>
</tr>
<tr>
<td>listen</td>
<td>46</td>
</tr>
<tr>
<td>speak</td>
<td>45</td>
</tr>
<tr>
<td>listen</td>
<td>44</td>
</tr>
<tr>
<td>speak</td>
<td>43</td>
</tr>
<tr>
<td>listen</td>
<td>42</td>
</tr>
<tr>
<td>speak</td>
<td>41</td>
</tr>
<tr>
<td>listen</td>
<td>40</td>
</tr>
<tr>
<td>speak</td>
<td>39</td>
</tr>
<tr>
<td>listen</td>
<td>38</td>
</tr>
<tr>
<td>speak</td>
<td>37</td>
</tr>
<tr>
<td>listen</td>
<td>36</td>
</tr>
<tr>
<td>speak</td>
<td>35</td>
</tr>
<tr>
<td>listen</td>
<td>34</td>
</tr>
<tr>
<td>speak</td>
<td>33</td>
</tr>
<tr>
<td>listen</td>
<td>32</td>
</tr>
<tr>
<td>speak</td>
<td>31</td>
</tr>
<tr>
<td>listen</td>
<td>30</td>
</tr>
<tr>
<td>speak</td>
<td>29</td>
</tr>
<tr>
<td>listen</td>
<td>28</td>
</tr>
<tr>
<td>speak</td>
<td>27</td>
</tr>
<tr>
<td>listen</td>
<td>26</td>
</tr>
<tr>
<td>speak</td>
<td>25</td>
</tr>
<tr>
<td>listen</td>
<td>24</td>
</tr>
<tr>
<td>speak</td>
<td>23</td>
</tr>
<tr>
<td>listen</td>
<td>22</td>
</tr>
<tr>
<td>speak</td>
<td>21</td>
</tr>
<tr>
<td>listen</td>
<td>20</td>
</tr>
<tr>
<td>speak</td>
<td>19</td>
</tr>
<tr>
<td>listen</td>
<td>18</td>
</tr>
<tr>
<td>speak</td>
<td>17</td>
</tr>
<tr>
<td>listen</td>
<td>16</td>
</tr>
<tr>
<td>speak</td>
<td>15</td>
</tr>
<tr>
<td>listen</td>
<td>14</td>
</tr>
<tr>
<td>speak</td>
<td>13</td>
</tr>
<tr>
<td>listen</td>
<td>12</td>
</tr>
<tr>
<td>speak</td>
<td>11</td>
</tr>
<tr>
<td>listen</td>
<td>10</td>
</tr>
<tr>
<td>speak</td>
<td>9</td>
</tr>
<tr>
<td>listen</td>
<td>8</td>
</tr>
<tr>
<td>speak</td>
<td>7</td>
</tr>
<tr>
<td>listen</td>
<td>6</td>
</tr>
<tr>
<td>speak</td>
<td>5</td>
</tr>
<tr>
<td>listen</td>
<td>4</td>
</tr>
<tr>
<td>speak</td>
<td>3</td>
</tr>
<tr>
<td>listen</td>
<td>2</td>
</tr>
<tr>
<td>speak</td>
<td>1</td>
</tr>
<tr>
<td>listen</td>
<td>0</td>
</tr>
</tbody>
</table>
Fictive self-quotation: quantitative and qualitative aspects of fictivity in European and Brazilian Portuguese

Luiz Fernando Matos ROCHA
UFJF
Rua José Lourenço Kelmer, s/n - Campus Universitário, 36036-900, Juiz de Fora - MG, luiz.rocha@ufjf.edu.br

Abstract
Studies on fictivity point out that certain linguistic expressions are only indirectly related to their meant referents and that unreal scene is often presented by language users as a means of mentally accessing the real scene. By overlapping cognitive and interactional frames, the fictive self-quotation phenomenon is a discursive type of fictivity, by which its conceptualisers pose a subjectifying assessing perspective to the direct speech in the first person. The objective of this work is to analyse fictive self-quotation and its factive co-extension in oral corpora of European and Brazilian Portuguese, focusing on the construction “(I) said X-clause”. As for the data, the C-ORAL-ROM Portuguese corpus (Bacelar do Nascimento et al., 2005), the C-ORAL Brazilian corpus (Raso & Mello, 2010, 2012), and a database from the reality show Big Brother Brasil (2002) are used, all of which subjected to electronic tools. The results point out meaningful conceptual, diatopic and diaphasic contrasts between the uses of “disse” and “falei” in the national varieties, since the verb “falar” is not often used to build a reported speech mental space in the European Portuguese and that, from a constructional standpoint, certain interactional frames seem to favour fictive self-quotation more promptly.

Keywords: cognition; fictivity; reported speech; self-quotation.

1. Introduction
Studies on fictivity point out that certain linguistic expressions are only indirectly related to their meant referents and that unreal scene is often presented by language users as a means of mentally accessing the real scene. By overlapping cognitive and interactional frames, the fictive self-quotation phenomenon is a discursive type of fictivity, by which its conceptualisers pose a subjectifying assessing perspective to the direct speech in the first person, differently from its factive counterpart. This is mainly due to the mismatched use between the traditional way of reporting self-speech and thought and the meaning of dicendi verbs like “dizer” and “falar”, which take an exclusively epistemic cognitive conflict, the image assessed as unreal is fictive.

By overlapping cognitive and interactional frames, the fictive self-quotation phenomenon is a discursive type of fictivity, by which its conceptualisers pose a subjectifying assessing perspective to the direct speech in the first person, differently from its factive counterpart. This is mainly due to the mismatched use between the traditional way of reporting self-speech and thought and the meaning of dicendi verbs like “dizer” and “falar”, which take an exclusively epistemic status (e.g. “I said (thought) “Oh, God!”). Therefore, by means of an unreal scene of discourse reporting, the illocutionary agent reports himself to a previous and assumed speech scene, aiming at allowing mental access to the real scene of thought.

The historical methodological track followed by the studies on fictivity is analogous to the one made by Cognitive Linguistics as a whole. It begins with works which are solely based on the linguists’ intuition, who developed epistemological constructs prompted by both imagery and linguistic illustrations, either made up or faked, though plausible, for postulating both psychological and cognitive state of affairs. Within this context, the main objective of this work is to describe and analyse fictive self-quotation and its factive co-extension in oral corpora of European and Brazilian Portuguese, focusing on the construction “(I) said X-clause”, devoid of any directional phrases (Goldberg, 1995) or active zones (Langacker, 1991), which would unquestionably point to its factive interpretation.

As for the data, the C-ORAL-ROM Portuguese corpus (Bacelar do Nascimento et al., 2005) and the C-ORAL Brazilian corpus (Raso & Mello, 2010, 2012) are used, as they have similar basic architectures. A database from the reality show Big Brother Brasil (2002) is also used. They were subjected to the TextSTAT or Contextes electronic tools. On the whole, the results point out meaningful conceptual, diatopic and diaphasic contrasts between the uses of “disse” and “falei” in the national varieties, since the verb “falar” is not often used to build a reported speech mental space in the European Portuguese and that, from a constructional standpoint, certain interactional frames seem to favour fictive self-quotation more promptly, as in the case of the reality show.

However, from a discursive point of view, fictivity affects self-quotation in both varieties of the Portuguese language, mapped by clues which include monological self-report, subjectification, epistemic co-text, deictic mismatch, mental scanning, the metaphor “THINKING IS SAYING” (Rocha, 2004, 2006, 2010), speech acts such as promises, planning and appreciation. Such signs form a set of semantic and pragmatic trends extracted from the one-to-one case analysis of real interactions, making interactional and cognitive frames to converge, thus supporting the multidimensional feature of the phenomenon, basically split into epistemic and pragmatic dimensions.

This contributes to an innovative view on fictivity which, according to Talmy (2000), only refers to cognitive conflicts between discrepant (fictive and factive) ways of perceiving or conceiving the same object. On the
other hand, if we take into consideration the associative force between a given construction and a given lexical item, and if we treat it from a discursive standpoint, we conclude that a fictive cognitive frame is evoked whenever a fictive interactional frame is.

2. Fictive and Factive self-quotation

The present study investigates how discursive and prosodic aspects contribute to the recognizing of fictive self-quotation as a virtual instance of direct speech, a grammatical construction, whose features are indirectly tied with the referents, referring to the worlds, entities mentally constructed, as well as the exclusively epistemic events. Fictive self-quotation is a kind of mismatch between form and meaning. This case represents form–function mappings which are “incongruent with respect to more general patterns of correspondence in the language” (cf. Francis & Michaelis, 2003: 2). Since this construction is a non-canonical pattern, it can be a direct consequence of a grammaticalization process and mainly a product of general fictivity pattern (Talmy, 1996: 212), in which “two discrepant representations disagree with respect to some single dimension, representing opposite poles of the dimension”. That is: FACTIVE AND FICTIVE SELF–QUOTATION.

We can find similar examples like these in English, as in Henry Kravis’ interview:

**Henry Kravis’ interview (1)**

**FICTIVE SELFQUOTATION (FIC-SELF):**
My dad was reading an article in Time magazine about the Oxford/Cambridge of the West Coast. It’s part of a group of small colleges in Claremont, along with Pomona, Scripps, and Harvey Mudd. I wanted to go to the West Coast. I’m from Oklahoma originally, but I had been in an Eastern boarding school for five years and I said, “I want to see how the other half of the United States lives.” I tell people I went there to play competitive golf. I liked it. I used to say the first year was like a prep school with ash trays. I really went there because it was very strong in economics and political science, and those were the two areas that I wanted to focus my future on. (http://www.achievement.org/autodoc/page/kra0 int-1)

In the boldface fragment the verb “said” has an epistemic meaning, as “think” or “consider”. “Said” is a dicendi and sentiendi verb at the same time. But it is not in the next example:

**Henry Kravis interview (2)**

**FACTIVE SELFQUOTATION (FAC-SELF):**

After I graduated from college, that summer, I was given a job at the Madison Fund, which was a closed-end mutual fund here in New York. Ed Merkle ran it. What a terrific guy he was! After I was there for about three weeks, he said, "Kid," (they used to call me kid all the time), "I want you to go out and call on a company called Tri-State Motor Transit, in Joplin, Missouri. And I said, "That’s interesting, but who is going to go with me?" He said, "What do you mean, who is going to go with you? You are going to go by yourself.

(http://www.achievement.org/autodoc/page/kra0 int-1)

In this case, “said” is just dicendi. It is not an epistemic use.

There are some discursive and prosodic clues which suggest that fictive selfquotation (FIC-SELF) is abnormal in relation to canonical factive self-quotation (FAC-SELF) although FIC-SELF keeps some features inherited from this traditional pattern, as we see in the next picture. Because of it, there is a dotted arrow linking FIC-SELF and FAC-SELF as a continuum. This process involves some grammatical means of coding formal, semantic or pragmatic functional domains. In terms of argumental structure, both cases are the same (I SAID X-clause). But the last feature is different when we submitted data to PRAAT, a free scientific software program for the analysis of speech in phonetics.

Formal tendencies:

<table>
<thead>
<tr>
<th>FICTIVE</th>
<th>FACTIVE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject + Sentiendi/dicendi verb + Speech clause (direct object)</td>
<td>Subject + Dicendi verb + Speech clause (direct object)</td>
</tr>
<tr>
<td>Tendency: verb in the past tense or in historical present</td>
<td>Tendency: verb in the past tense or in historical present</td>
</tr>
<tr>
<td>No complementizer (direct speech)</td>
<td>No complementizer (direct speech)</td>
</tr>
<tr>
<td>Prosody (1)</td>
<td>Prosody (2)</td>
</tr>
</tbody>
</table>

**Table 1: Subjitive and factive**

Considering the scope of tested fragments made by Professor Pablo Arantes, from Federal University of Minas Gerais (Brazil), fictive selfquotation is different from the factive one in some aspects. Such difference is provided by the comparison between five factive selfquotation occurrences and four fictive self-quotation occurrences. All these instances were uttered by male voices and extracted from Brazilian reality shows available on YouTube. According to the nine examples, in terms of fundamental frequency movement, which means a major acoustic manifestation of suprasegmental structures such as tone, pitch accent, and intonation, there is no outstanding differences between both selfquotations. In general, fictive and factive selfquotation show soft curves.

Even though this corpus is small, in global sense, it
shows consistent differences in terms of (i) register, a quality voice element whose purpose can make speech more expressive, and emphatic; and (ii) tessitura, a speech melody element whose melodic height variations represent cohesive function. Fictive selfquotation curves occupy low tone region (bass-pitched). Factive selfquotation curves occupy high tone region. These numbers are statistically meaningful and contribute to the fact that we have distinct vocal construals. Besides, the variability of F0 is different in both cases. In Factive selfquotation, there is more F0 curve variance than in the fictive one. As a robust and perceptual parameter, the variation range of curves in each selfquotation is too different: fictive cases (6.8 semitones); factive cases (13.8 semitones), which means there are distinct kinds of half step, as the interval between two adjacent notes in music.

The graphic below shows F0 curves of factive and fictive according to time normalization technique, whose purpose is to try to set up equivalence among sentences with different extensions and facilitate direct comparison among different points of F0 curves making them similar. Basically, on the left, this graphic presents five factive curves that occupy a large extension in terms of hertz; on the right, the four fictive cases do not. This means more tone variability in factive cases than in fictive ones.

![F0 curves of factive and fictive occurrences](image)

**3. Meaning tendencies**

In this section, we have a comparison between meaning tendencies of fictive and factive self-quotation, which we have found in the corpora:

1. FIC-SELF and FAC-SELF constructions occur mainly in narrative textual types;
2. The frame of reporting scenario is monologic in FIC-SELF; in FAC-SELF, dialogic;
3. There is previous co-textual information before fictive selfquotation, like other epistemic verbs; in FAC-SELF, there is none;
4. In FIC-SELF, there is an epistemic space-builder whose semantic value is sentiendi and dicendi at the same time in the sense of “think” or “consider”; but in the factive case, this value is only dicendi; in FIC-SELF, there is the metaphor THINKING IS SAYING and the metonymy SAYING FOR THINKING.

5. The first one evokes an assessing frame and the second one a speech communication frame;
6. Fictive selfquotation tends to present speech acts in terms of promising, planning, evaluation, and concluding; factive tends to present speech acts in terms of requests, advice, suggestion, instruction, and asserting;
7. Considering all the scenario around the verb “falei” or “disse” in corpora, there is a strong tendency: fictive self-quotation is pairing with a fellowship face. On the other hand, factive selfquotation is pairing with competence face;
8. In fictive selfquotation, addressee in reported narrative is the speaker himself; but in factive, it is another character;
9. In fictive, vocative is a generic entity, for example, “Deus” (God), “gente” (folks), but in factive, we commonly have a person’s name;
10. Even though we do not find such clues, deixis phenomena in the embedded clause can help us to distinguish both constructions. Let us see an example:

**BRAZILIAN PORTUGUESE:**

JUL: <teve um dia que alguém me falou assim / Nossa / cê tá velha / hein / sua menina tá com dez anos / eu falei / velha é ela //

(C-ORAL Brasil - RASO & MELLO, 2012)

**TRANSLATION:**

JUL: someday someone told me: "You're old! Your daughter is ten!". I said: she is old!

The exchange of "you", second person, as "she", third person, in the X-clause (VELHA É ELA = SHE IS OLD, not YOU ARE OLD) becomes the direct speech (I said: she is old) a fictive self-quotation, although we have a previous direct speech frame: “someday someone told me: ‘You're old! Your daughter is ten!’.” The third-person deixis 'she' is inconsistent with that scenario marked by past tense verbs “told” and “said”. Besides, if it would be a case of factive self-quotation, in the reported interaction, the speaker JUL would have to use YOU and to say: YOU ARE OLD!, as the character “someone” does. It means we have just one clue to read all the self-quotation as fictive, which is discrepant with respect to a single deitic dimension.

**4. Quantitative analysis**

For the quantitative analysis from those corpora, I have searched the pattern (EU) DISSE/FALEI X-ORACIONAL (I SAID X-clause, in English) to find self-quotations in first person, using TextSTAT
concordance software and Contexts concordance from C-ORAL-ROM project.

In European Portuguese, the verb “falar” (to say), in general, does not profile dicendi substructure. In this sense, it is similar to the verb “speak”, in English. In European Portuguese, this function belongs to the verb “dizer” (to say). In Brazilian Portuguese, the verbs “dizer” and “falar” can profile dicendi substructure. In relation to selfquotation, all these numbers that we will see signalize important contrasts between national varieties of Portuguese, for example, the preference for “dizer” instead “falar” as a dicendi verb in European Portuguese than in Brazilian one. The former profiles a punctual process of demonstration by word of beliefs and convinced attitudes. The latter profiles a general process of verbalization, which refers to skills and abilities of speech production.

In European Portuguese Corpus, we have found 50 types of the pattern (EU) DISSE X-ORACIONAL, being 44 FAC-SELFs and only six FICTIVE-SELFs. The mainly reason for that is the specificity of this pattern, which is semi-instantiated. On the other hand, in the same corpus, we have found just 21 occurrences of “falei” associated with prepositional phrase in general, which for us means that there is no dicendi function. This is a kind of counter-evidence of FIC-SELF.

In terms of Brazilian Corpus, the word form “disse” (I said) occurs two times, being two cases of FAC-SELF and there are no FICTIVE-SELF cases with this form. But the word form “falei” (I said”, in English); 43 are cases of fictive self-quotation and 26 are cases of factive self-quotation. These numbers can’t be understood as a mere generalization. It signals that we use it a lot, depending on the interactional frame. Note that in a reality show, reported speech frame is a powerful and pervasive construction as “war” strategy. In this sense, fictive selfquotation justifies the reporting thoughts through an epistemic and discrepant use of “say” (“falei”) with the purpose of profiling more action and confidence than the mere use of “think” or “consider”.

5. Conclusion

It is important to highlight that the abundance of virtual computational architectures to study linguistics has a single purpose: to gain more precise access to language and to what is psychologically real in processing it. In other words, the fictivity of the proper linguistics investigation seems to be the current point of no return in the history of linguistics. In the case of this work, PRAAT and Corpus Linguistics instruments have permitted that fictive selfquotation is understood as a phenomenon which depends on its integrated features to be mapped. With PRAAT, we can say that fictivity has an specific melody when we can constrast it with factivity occurrences through similar constructional patterns. With Corpus Linguistics, we can see the integration of grammatical constructions with discourse more clearly; and show in more details how it happens; and verify how the conceptualizer sets up alternatives forms of construal for the same referent or situation, conventionalizing language changes. The comparison between European Portuguese Corpus and Brazilian Portuguese Corpus has revealed that both national varieties have their proper way of profiling fictive selfquotation. As their corpora architectures are the same (both under C-ORAL-ROM project), the numbers of fictive self-quotation occurrences are not very different proportionally, but when we compare these corpora with another one (reality show), we can see how fictivity depends on the interactional frame to be more or less productive. Cognitive frames of fictivity are strongly in action when interactional frames of fictivity are in action, too.

6. Acknowledgements

This paper is part of the results of the research project Construções gramaticais de conversa fictiva: convergências entre frames cognitivos e interacionais (Fictive talking as grammatical construction: convergences between cognitive and interactional frames) (CAPES Scholarship - Case: BEX 4084/10-1), developed during the post-doctoral internship under the supervision of Augusto Soares da Silva, PhD. (2010-2011), in collaboration with the Centro de Estudos Filosóficos e Humanísticos da Universidade Católica Portuguesa (Center of Philosophical and Humanistic Studies of the Portuguese Catholic University) (CEFH-UCP), Braga Regional Center (Portugal), in the scope of the research line Estudos Linguísticos e Literários (Linguistic and Literary Studies) and the project Linguística Cognitiva do Português: semântica e gramática (Portuguese Cognitive Linguistics: semantics and grammar).

7. References

Brandt, L. (2010). Language and enunciation – A cognitive inquiry with special focus on conceptual integration in semiotic meaning construction. Faculty of the Humanities: Aarhus University.
Linguistics Sciences. V. 29, n. 2.
Modeling the grammar of ASSESSMENT of casual conversation in Brazilian Portuguese: the design of a corpus to investigate language probabilities functioning in context

Giacomo Patrocinio FIGUEIREDO, Adriana Silvina PAGANO
Federal University of Ouro Preto; Federal University of Minas Gerais
Rua do Seminário, s/n, Mariana, MG, Brazil 35420-000
giacomojakob@yahoo.ca, apagano@ufmg.br

Abstract

This paper presents a corpus-based model for the interpersonal system of ASSESSMENT in the clause grammar of casual conversation (Eggins & Slade, 1997) in Brazilian Portuguese. More specifically, it examines Modal Particle use. Data were obtained from a sample of casual conversation retrieved from CALIBRA, a monolingual corpus of Brazilian Portuguese designed following a context-based typology of texts. The texts were analyzed according to systemic functional theory categories (Halliday & Matthiessen, 2004; Figueredo, 2011) and semi-automatically annotated for grammar categories with the software CorpusTools (O’Donnell, 2008). Variation in the patterns of Particle use was found for the whole corpus and those in the subsection where casual conversation is located. Results pointed to a more frequent use of Modal Particles for Assent, Understand, Confirm and Conclude and therefore to a more variation in the patterns of Particle use was found for the whole corpus and those in the subsection where casual conversation is located. On the other hand, Modal Particles related to the systems of PERSUASION and PROSODY were observed to contribute less to the variation found in casual conversation. More specifically, it examines Modal Particle use. Data were obtained from a sample of casual conversation retrieved from CALIBRA, a monolingual corpus of Brazilian Portuguese designed following a context-based typology of texts. The texts were analyzed according to systemic functional theory categories (Halliday & Matthiessen, 2004; Figueredo, 2011) and semi-automatically annotated for grammar categories with the software CorpusTools (O’Donnell, 2008). Variation in the patterns of Particle use was found for the whole corpus and those in the subsection where casual conversation is located. Results pointed to a more frequent use of Modal Particles for Assent, Understand, Confirm and Conclude and therefore to a more variation in the patterns of Particle use was found for the whole corpus and those in the subsection where casual conversation is located. On the other hand, Modal Particles related to the systems of PERSUASION and PROSODY were observed to contribute less to the variation found in casual conversation.

Keywords: casual conversation; assessment; modal particles; monolingual corpus; Brazilian Portuguese.

1. Introduction

Drawing on the notion of probabilistic grammar (Halliday, 1991), this paper presents a corpus-based model for the interpersonal system of ASSESSMENT in the clause grammar of casual conversation (Eggins & Slade, 1997) in Brazilian Portuguese. ‘Modeling’, as construed here, can be defined as the description of grammar features and statement of their probabilities of instantiation for the text type under investigation. Halliday (1978) conceives of language as a naturally evolved semiotic system, its main purpose being to offer a reservoir of meaning-making resources for humans to interpret and organize both our natural world and our social relations. Grammar is, in turn, the stratum of language responsible for creating meaning. Since meaning is, in fact, the contrast of paradigmatic features (Saussure, 2006), for any given language subsystem, the job done by the grammar is to change (responding to the pressure of new contextual demands) the systemic (paradigmatic) organization of features in order to create meaning. This process of specialization leads to language variation. As a result, language is modeled in terms of (Halliday, 1991): (i) its relations to the context of culture – the “environment” in which it takes place, in which it is meaningful; and (ii) the process through which language as a reservoir of meaning-making potential (the system) becomes, via grammar operations, language in context (the text). Consequently, the modeling of “actual” grammar – the grammar that creates meaning functioning in the context of situation – needs to account for (a) the way context is materialized in language (examining the systemic dimension of realization) and (b) the probabilities for a potential grammatical feature to be instantiated as text (the dimension of instantiation) (Halliday, 1991). Thus, to model the contextual pressure that ultimately causes language variation – in other words, to model any given text type – including casual conversation, it is necessary to account for the dimensions of realization and instantiation. Following Halliday’s (1978) conceptualization of language, a great number of studies have explored grammar from a realizational point of view (cf. Martin, 1992; Caffarel, Martin & Matthiessen, 2004, among others and Eggins & Slade, 1997, specifically for casual conversation). A smaller number of studies have explored the instantiational process (cf. Matthiessen, 2001; Martin, 2008, among others). There are fewer studies still drawing on the realization-instantiation complementarity (Matthiessen, 2004). To a large extent, this is due to the fact that the process of instantiation leading to the modeling of specific text types is not fully understood (Martin, 2008). By presenting a modeling of casual conversation interpersonal grammar systems, this paper aims at exploring the complementarity of realization and instantiation, as well as contributing to the understanding of probabilities in the constitution of text types. More specifically, it presents a study of the interpersonal grammatical system of ASSESSMENT in Brazilian Portuguese, including its distribution across text types and relates that to the distribution of ASSESSMENT functions in casual conversation. Such relation can ultimately lead to the modeling of casual conversation in Brazilian Portuguese and contribute to consolidating corpus-based investigation as a necessary step towards the understanding of the instantiation process.

2. Theoretical underpinnings

2.1 The design of a corpus to investigate language probabilities functioning in context

Drawing on the concept of text as “language functioning in context” (Halliday & Hasan, 1976), Matthiessen, Teruya & Wu (2008) propose a typology based on the
A corpus design based on the typology above allows for the study of language frequencies of grammatical systems, both globally in the language system as a whole, and “broken down” according to typological features of language in the context of culture (Halliday, 1992). CALIBRA, which stands for Catálogo da Língua Brasileira, is one such corpus, designed on the basis of the language typology proposed in Matthiessen, Teruya & Wu (2008). CALIBRA is a monolingual corpus of Brazilian Portuguese, which compiles language produced in a natural communicative setting and representative with respect to each of the socio-semiotic processes mentioned above. It is a raw corpus with minimal header annotation and encoding in UTF-8. Texts compiled in CALIBRA were produced within the 1990–2010 decades. As regards the spoken mode, texts were recorded from spontaneous speech and subsequently transcribed to be incorporated. The corpus design allows for mapping a particular language variety. For the purposes of the present study, which targets casual conversation, texts can be located in the typology as non-specialized, spoken, dialogic texts within the sharing process. A detailed account of this variety is provided in the following sections.

### 2.2 Casual conversation

As a species, human beings are part of the animal world. This means that our biological constitution needs food and shelter; safety and companionship. No human can live their whole life alone apart from other humans. It is also part of our species programing to be able to keep track and record of time towards the past by building and storing personal and collective memories and to the future by predicting, planning and realizing projects, such as finding food, building shelter, or maintaining relationships. As a result, our biology determines only partly what humans are, since it is embedded in our social world and in our history – not only individual histories of each human being, but the history of our social world.
(Malinowski, 1935).

The shaping of biology [by society [shaped by history] lies at the core of a functional theory of culture. The process is called symbolic modeling and ultimately explains why ‘mating’ becomes ‘marriage’, a ‘pack’ becomes a ‘family’, and ‘feeding’ becomes a ‘dinner party’. Culture, then, is a symbolic system of conditioning for human beings, turning the specimens into people with a place in society for a given period of history.

Language has a crucial part to play in symbolic modeling. It is through language that culture conditions human beings. Education, the law, religion and all institutions responsible for passing on a means of survival, a code of values and so on to the next generation are all fully dependent on language. Malinowski (1935) states that language creates the symbols of a social group, it organizes institutions by developing particular discourses and stores knowledge in the texts that are taught and shared among its members.

Casual conversation, thus, assumes a special status in this process, given that it responds for creating and passing on knowledge and values efficiently in a very specific context – that of people who are closest to each other. Eggnis and Slade (1997) state that casual conversation is a resource frequently deployed in negotiating our social identity and establishing our “social geography” – the people (along with their values and social relations) who are close or distant from us. The reiteration and multiplication of such texts through a period of time contribute to social stratification and distribution of power among people in a social group.

### 2.2.1. The grammar of casual conversation

Language can serve as the most resourceful tool in symbolic modeling because it has a grammar (Halliday, 1978). Semiotic systems are bi-stratal, in which a symbol is characterized by the univocal correspondence between its content plane (“semantics”) and its expression plane (“phonetics”) (Saussure, 2006). Language, however, has evolved to formally organize the content (Hjelmslev, 1969). The content is, as it were, divided into two: the substance of content (semantics) and the form of content (grammar). Grammar, then, is defined as the formal organization of language content plane. Consequently, the meaning of a linguistic symbol is not conveyed by the univocal correspondence between content and expression; rather, the understanding of content can change depending on its formal organization. Since meaning is, in fact, a paradigmatic contrast between symbols, grammar operates altering the organization of systems in order to create new meanings. Whenever there is need for a reshaping of some aspect of human life – different aspects of symbolic modeling – there is also a contextual pressure for new meanings and new texts. Grammar reorganizes features of systems, changing both their paradigmatic contrast and their probability, thus creating new meanings through variation of text types.

The grammar of casual conversation is one example of such process. Responding to the contextual pressure of negotiating social identity and drawing social geography maps, the grammar of casual conversation has created meanings to materialize such contexts (cf. Eggins & Slade, 1997). For example, interpersonal systems (MOOD, MODALITY, ASSESSMENT and POLARITY) are deployed to establish a “sympathy relation” towards the speaker’s values and positions. Ideational systems (TRANSITIVITY and EPITHESIS) help building the narrative underlying casual conversation as well as passing judgment and ascribing voice and thought to other people. Textual systems (THEME and INFORMATION) help staging phases of casual conversation, as well as giving prominence to interpersonal and ideational systems relevant to the construction of typical features of casual conversation such as sympathy, narrative and judgment (cf. Eggins & Slade, 1997).

From the point of view of instantiation, the grammar changes the typical, non-prominent, ratio of feature instantiation, due to contextual pressure. Although fewer features in relation to the whole of the system are deployed, these are relatively more frequent in casual conversation. One such case is found in the interpersonal system of ASSESSMENT in Brazilian Portuguese.

### 2.2.2. The system of ASSESSMENT

Any interaction between people can be viewed as the negotiated process of converting interactants’ [personal] opinions into [interpersonal] shared knowledge. The amount of opinion converted into shared knowledge is likely to determine the social proximity/distance among interactants for a given interaction. By the same token it is likely to indicate distribution of power, knowledge, expertise, authority, etc., contributing to determine their social identity. In general, a speaker tends deploy resources (from systems such as MODALITY, ASSESSMENT, INTONATION, etc.) which may increase the chances of his/her opinions being accepted. In this sense, the concept of ‘valid or not-valid’ is a very important feature of social relations, since it is an outcome of negotiation (cf. Halliday & Matthiessen, 2004). The interpersonal grammar deploys a set of sub-systems precisely to negotiate positioning, power, values and “social geography”. These are collectively responsible for exchanging evaluation and can be characterized by two features: (i) extension of evaluation – the speaker marks his/her position towards what s/he is saying; (ii) orientation of evaluation – the speaker marks his/her position towards his/her own role as speaker, or demands an assessment from the listener to do so. The interpersonal systems mostly associated with (i) are MODALITY, POLARITY and partially MODAL ADJUNCTS (mood and comment). The interpersonal systems associated with (ii) are partially MODAL ADJUNCTS (mood and comment) and ASSESSMENT.

Martin and White (2005: 95) describe the semantic region of engagement among interactants as:

> “when speakers/writers announce their own attitudinal positions they not only
self-expressively ‘speak their own mind’, but simultaneously invite others to endorse and to share with them the feelings, tastes or normative assessments they are announcing. Thus declarations of attitude are dialogically directed towards aligning the addressee into a community of shared value and belief’.

Engagement meanings, in turn, are grammaticalized by the system of ASSESSMENT, defined by Halliday and McDonald (2004) as: “a grammatical system ... whereby the speaker signals attitude to, and degree of involvement in, the proposition or proposal of the clause (p. 341).”

In Brazilian Portuguese, the system of ASSESSMENT is realized by Modal Particles (Lam, Figueredo & Espíndola, 2010) as displayed in Figure 1.

Figure 1: The system of ASSESSMENT in Brazilian Portuguese

Particles function adding further options to MOOD selection, shaping statements, questions, commands and offers according to the speaker’s need for their interlocutor’s assessment of a move, such as exhorting, agreeing, concluding, etc. By using Modal Particles in Brazilian Portuguese a speaker can not only assess what is being said, but also invite the listener to assess the speaker’s own role as speaker [the one who assess what is being said]. Modal Particles can be more strongly associated with propositions – the exchange of information – realized by Indicative Mood; and those associated with proposals – the exchange of goods-and-services – realized by Imperative Mood. Modal Particles carry two complementary interpersonal functions in the clause: they indicate how the clause should be valued in terms of agreement, assent, exhortation, etc.; and they are picked up by the listener as a means of propelling dialogue. Examples of ASSESSMENT functions in Brazilian Portuguese retrieved from CALIBRA can be found below, a gloss and a free translation being provided for each of them.

Ó João você toma conta deles ATTEND João you take care of them
“Listen to me João, you take good care of them.”

Todo mundo lá gostava dele né All world there liked of he ASSENT
“Everybody liked him, don’t you think so too?”

S1 Não deve de ser para ligar para elas Not must of be to call for them
“We are not supposed to call them.”
S2 Eu acho que é sim sô. I think that be yes INSIST
“I think that be yes.”

“*But I do think we are.*” Vocês não voltam pra lá VIU

“You not return to there UNDERSTAND

“You should never go back there, is it clear?”
Você grava as minhas aulas é
You record my lectures CONFIRM
“So you tape my lectures, do you?”

E eu tava animado sô And I was excited SYMPATHIZE
“And I was extremely excited.”

Ôxe quem tá ligando pra isso? EXCLAIM who is caring to this
“Why on earth would anyone care about it?!”

Fala aí o quê que você faz Speak ATTENUATE the what that you do
“Please, tell me what you do.”

Então o que é têchê? Speak ATTENUATE the what that you do
“Just say what it is.”

3. Methodology

To model casual conversation, the following methodology was adopted. A spoken language corpus of 10,000 tokens (10 texts of 1,000 tokens) of casual conversation was compiled from CALIBRA (Catalogue of the Language of Brazil). As previously mentioned, casual conversation texts are located in the typology as non-specialized, spoken, dialogue texts within the sharing process. For CALIBRA, spoken texts, including casual conversation texts used in this research, are recorded from spontaneous conversations and subsequently transcribed. Few features are inserted into the transcriptions, including basically clause/information unit separation: ‘.’ falling tone; ‘?’ rising tone; ‘…’ level tone; ‘,’ short pause; ‘--’ hesitation or turn-taking; ‘[‘] more than one speaker speaking simultaneously.

After compilation, texts in the casual conversation
corpus were analyzed according to systemic functional theory categories (Halliday & Matthiessen, 2004; Figueredo, 2011) and semi-automatically annotated for grammar categories with the software CorpusTools (O’Donnell, 2008). This software allows researchers to annotate texts with categories of interest and retrieve their frequency along the corpus tested for statistical significance.

Drawing on Halliday (1991b), who states that counting frequencies in a text is, in fact, stating instantiation probabilities in the grammar, the frequencies obtained were analyzed in order to reach a probabilistic grammatical profile of Particles based on the generalization of frequencies found in the corpus.

4. Modeling ASSESSMENT for casual conversation in Brazilian Portuguese

The concept of ‘modeling’ implies that the results of a study carried out for a sample allow us to make estimates for the whole of the population. When performing a modelling of a grammar feature, an account of the functional distribution of a particular resource (realization), together with its variation across text types (instantiation), is needed. The sample in this case is defined by two complementary steps. First, a grammar description is needed, so the “strings of sounds” found in the corpus can be converted into grammar features. As a result, the corpus under investigation – the “true” corpus – is a sample of grammar patterns. When querying CALIBRA for the categories in the system of ASSESSMENT, the patterns in Table 2 were found:

Table 2: Global model for ASSESSMENT

<table>
<thead>
<tr>
<th>Particle function</th>
<th>Occurrence No.</th>
<th>Relative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>assent</td>
<td>498</td>
<td>60.8%</td>
</tr>
<tr>
<td>conclude</td>
<td>73</td>
<td>8.9%</td>
</tr>
<tr>
<td>attend</td>
<td>61</td>
<td>7.4%</td>
</tr>
<tr>
<td>exclaim</td>
<td>60</td>
<td>7.3%</td>
</tr>
<tr>
<td>agree</td>
<td>48</td>
<td>5.9%</td>
</tr>
<tr>
<td>attenuate</td>
<td>41</td>
<td>5.0%</td>
</tr>
<tr>
<td>understand</td>
<td>12</td>
<td>1.5%</td>
</tr>
<tr>
<td>exhort (ans.)</td>
<td>10</td>
<td>1.2%</td>
</tr>
<tr>
<td>confirm</td>
<td>6</td>
<td>0.7%</td>
</tr>
<tr>
<td>exhort (und.)</td>
<td>5</td>
<td>0.6%</td>
</tr>
<tr>
<td>challenge</td>
<td>4</td>
<td>0.5%</td>
</tr>
<tr>
<td>sympathize</td>
<td>1</td>
<td>0.1%</td>
</tr>
<tr>
<td>TOTAL</td>
<td>819</td>
<td>100%</td>
</tr>
</tbody>
</table>

Secondly, a distribution of grammar features across text types is needed, so variation patterns can be observed. The results obtained from CALIBRA are shown in Table 3.

Based on these complementary distributions, it is possible to see if there is significant variation between the patterns in Particle use found for the language and those in casual conversation, evidenced by texts located in CALIBRA within the sharing process sharing. This variation is seen in Table 4.

Table 3: Typological variation for ASSESSMENT

<table>
<thead>
<tr>
<th>Process Function</th>
<th>Expo</th>
<th>Rep</th>
<th>Rec</th>
<th>Sha</th>
<th>Do</th>
<th>Recom</th>
<th>Ena</th>
<th>Expl</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>attend</td>
<td>2</td>
<td>3</td>
<td>11</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>29</td>
<td>4</td>
</tr>
<tr>
<td>exhort (und.)</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>attenuate</td>
<td>5</td>
<td>1</td>
<td>13</td>
<td>5</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>41</td>
</tr>
<tr>
<td>exhort (ans.)</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>challenge</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>exclaim</td>
<td>0</td>
<td>5</td>
<td>11</td>
<td>18</td>
<td>10</td>
<td>2</td>
<td>2</td>
<td>14</td>
<td>60</td>
</tr>
<tr>
<td>sympathize</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>assure</td>
<td>15</td>
<td>110</td>
<td>25</td>
<td>142</td>
<td>16</td>
<td>33</td>
<td>45</td>
<td>112</td>
<td>498</td>
</tr>
<tr>
<td>understand</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>agree</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>10</td>
<td>0</td>
<td>11</td>
<td>14</td>
<td>48</td>
</tr>
<tr>
<td>confirm</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>conclude</td>
<td>13</td>
<td>0</td>
<td>15</td>
<td>24</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>73</td>
</tr>
<tr>
<td>TOTAL</td>
<td>42</td>
<td>120</td>
<td>92</td>
<td>209</td>
<td>67</td>
<td>42</td>
<td>66</td>
<td>181</td>
<td>819</td>
</tr>
</tbody>
</table>

Table 4: ASSESSMENT model for casual conversation

The data presented in Table 4 show how casual conversation departs from the expected ratio for the whole language. In terms of ASSESSMENT more specifically, it is possible to see skewing in the use of Modal Particles, there being more frequency for Assent, Understand, Confirm and Conclude, all belonging to the sub-system of ROLE TYPE (see Figure 2, above). As a result, it is possible to estimate that this region of ASSESSMENT is contributing more intensely to the process of negotiation among interactants in casual conversation. On the other hand, there is skewing to a less frequent use for the other Modal Particles, suggesting that the sub-systems of PERSUASION and PROSODY contribute less to the variation found in casual conversation.

5. Conclusion

The results obtained for Modal Particle use in casual conversation texts drawing on a corpus sample validate the methodology used to model the interpersonal system.
of ASSESSMENT and can be further applied in order to describe other grammar features in Brazilian Portuguese and state their probabilities of instantiation in a particular text type. The idea of text variation – including the modeling of grammar systems – needs to account for the small perturbation in the average feature choices for any given text. The results presented in this paper can show how ASSESSMENT is deployed in such fashion, as to point how feature choices are skewed to vary the system towards casual conversation.

6. References


Fragments as instantiation of syntactic slots: complexity of the interface between lexicon, grammar and discourse in spoken French

Marie-Noëlle ROUBAUD, José DEULOFEU
EA 4671 ADEF; Aix Marseille Université; ENS de Lyon, IFE; 13248, Marseille, France
LIF-CNRS UMR 7279; Aix Marseille Université
mnroubaud@yahoo.fr, jose.deulofeu@gmail.com

Abstract

A particular feature of spontaneous speech syntax is the abundance of “fragments” (non clausal text units) which are generally analysed as independent syntactic text units. The main purpose of the paper is to show that many of them are in fact licensed by a verb of a preceding text unit, directly or by means of complex constructions, one of which will be discussed in detail: PSS. We will show that by reconsidering the syntax lexicon interface. We assume following Blanche-Benveniste et al. (1984) that this interface is highly complex. There are many ways in which a syntactic slot can be filled: null, pronominal, simple lexical, list of lexical items. And finally by means of a “discourse grafting” device (Deulofeu, 2010). One subcase, PSS, already investigated (Roubaud, 2000; Blanche-Benveniste, 1986, 2010), displays a combination of two fillers: a pronoun or a “light” lexical unit followed by a second one bringing a progressive semantic specification. In those patterns, the second clause does not obligatorily meet the subcategorisation requirements of the main verb. Such patterns pose the question of the limits between syntax and discourse. And also between structural and “online” syntax. Finally, we will show that PSS is combined with higher level discourse patterns in order to overcome processing problems.

Keywords: spoken French; corpus; syntax; progressive semantic specification.

1. Introduction

The main purpose of our paper is to revisit the way “fragments” and more generally syntactic structures are linked to the linguistic context through a corpus based study. This study is a piece of a more general project aiming to develop a competence grammar compatible with descriptive generalizations captured through spontaneous speech analysis This amounts to specifying the interface between grammar, lexicon and discourse.

Our empirical domain can be defined as “extended fragments”. “Fragments” are non sentential utterances syntactically autonomous but linked to a host construction by means of syntax-semantics interface rules (Culicover & Jackendoff, 2005):

L1 who came yesterday L2 Bo

We further define “extended fragments” as lexical items or constructs linked to a syntactic slot of a construction within « discourse patterns ». We look at defining the nature of that link.

2. Framework

We rely on the theoretical framework of Approche Pronominale (Blanche-Benveniste et al., 1984) revisited with Basic Linguistic Theory (Dixon, 2009). This framework, which can be compared with the one presented in chapter 8 of Biber et al. (1999); has been applied to spoken language analysis in numerous studies (Blanche-Benveniste & Jeanjean, 1987; Blanche-Benveniste et al., 1990; Blanche-Benveniste, 1986, 1997, 2010-b; Deulofeu, 2010).

The main Approche Pronominale (AP) assumptions are the following:

AP stands as a lexicalist approach of syntactic structures: lexical items licence syntactic slots: manger (eat) [P0, P1]; Pronouns and not full lexical items or phrases are default fillers of syntactic slots: je le mange ;

The paradigms of pronouns which can be built in the slots determine their grammatical features.

Lexical heads (constructeurs) with their underspecified syntactic slots are the basic components of syntactic constructions (skeletons). Syntactic skeletons slots are filled with lexical features to give full-fledged constructions.

Lexicalization can be "direct" lexical items fill directly the slots or “indirect”, involving additional grammatical devices (dispositifs).

As for the interface with performance, we assume, departing from the view that fragments are self contained syntactic units that an abstract syntactic construction (competence) can be uttered at once or in several times by the same speaker or several ones, which can result in a concatenation of fragments.

This lexicalization strategy can be linked to various competence performance interaction studies (Apotheloz, 2008; Auer, 2005; Blanche-Benveniste, 1990; Deulofeu, 2011).

More specifically, as particular structures are concerned, we propose to include lexicalization within the “performance” patterns identified by Iwasaki & Ono (2002): “to eyes used to the constructed data in linguistic literature, sentences in Japanese conversation look rather chaotic… though these types of utterances have been traditionally regarded as performance errors, careful examination reveals several clearly identifiable patterns, which we call “on line mechanisms”... We think these patterns are systematic enough to deserve a place in grammar…: phenomena of interpolation, incrementation, reformulation, local management and bridging… furthermore it is our hope that continuing analysis of spoken data in different languages will allow us to
3. Indirect filling

3.1 The case of lists of fillers

Consider the syntactic skeleton: faire P0 [-pers], P1 [-pers, -verbal]. The lexicalization of this abstract pattern can be Direct or Indirect.

A Direct pronominal filling will give the following construct: ça fait ceci. In the same way, a Direct lexical filling will give: son truc faisait une minerve [his stuff was (like) a neck brace].

Various types of “Indirect” lexical filling are possible as the utterance is processed: double filling, list filling, zero filling (contextual inferences).

Example:

Indirect lexical filling of a syntactic skeleton ça fait ceci by a « list » of lexical items with two speakers:

(1) L1: ça fait un + un + comment on dit je sais plus + une chose + là
L2: une écharpe + un col roulé + un
L1: mais non + le truc blanc là + qu’ils ont ceux qui se sont cassé là + le
L2: ah oui
L1: là + là + là + la cheva-
L2: la minerve
L1: voilà + la minerve (oral, privé)

According to our assumptions all the NPs which look like independent fragments are to be linked as indirect lexicalizations to the object syntactic slot of faire. This results in a fragmented filling (Deulofeu, 2011) of a syntactic slot. The link of the structural skeleton and the on line processing can be visualized by means of a graphic device: a « grid » as defined by Blanche-Benveniste & Jeanjean (1987). The structure SVO can be read horizontally whereas one can see vertically how the “Indirect filling” is processed:

(1) L1 ça fait un
un
comment on dit
je sais plus
une chose là
L2
une écharpe
un col roulé
un

L1 mais non le truc blanc là
qu’ils ont ceux qui se sont cassé là
L2 ah oui
L1 là
la
la
tache
L2 la minerve
L1 voilà la minerve

What is interesting to notice is that if the syntactic status of the fragments is the same (object of faire), their semantico pragmatic status is different. The material which is added to the NP - disfluences, metalinguistic remarks (comme on dit, je ne sais plus) discourse markers (viola) - helps the participants to evaluate the information status of the fragments - approximation, invalid lexical search (non), successful filling (oui, voilà). This material is not to part of the abstract syntactic structure. It comments on the process of lexicalization which belongs to the utterance building level.

3.2 The pseudo-cleft case

In the former example the lexicalization process involves paradigmatic listing of one grammatical category (NP) with added items not integrated in the grammatical structure (oui, voilà…).

In other cases Indirect lexicalization involves a grammatical device: the combination by means of the pseudo-cleft construction of two possible fillers of a syntactic slot between which stands a semantic relationship of “progressive specification” (Roubaud, 2000). In the following examples the two possible fillers of the object of faire are ce que (what) and the NP le saut en extension:

(2) ce que je sais faire c’est le saut en extension (oral, privé)
[what I can do is the extension jump]

Part 1 (what I can do) is semantically underspecified what = [gr. function : P1], [-pers], [-verbal]
I can do = head verb and other dependants
Part 2 (is the extension jump) is semantically specified
the extension jump = lexical features : [movement of body] [+ extension of body]

Notice that when a full pseudo cleft pattern is used semantic progressive specification must obey grammatical constraints as both fillers must meet subcategorisation rules coming from the lexical structure of the “main” verb as well as lexical restrictions; with the verb say the filler introduced by c’est must be something that can be said:

(3) ce que je peux dire c’est que nous ne sommes pas...
inutiles (oral, TV)
[what I can say is that we are not useless]

This pattern of progressive semantic specification (PSS) acts further as a repair device for clausal filling. This has many advantages. For example, it facilitates the processing of a clausal subject, which is almost excluded in spontaneous speech as filler in Direct filling:

(4) ce qui rendait les choses particulièrement difficiles c'est que la variation est double (oral, public)
* que la variation est double rendait les choses particulièrement difficiles
[what made the things particularly difficult was that this variation was double]

(5) ce qui me choquait un petit peu c'est qu'il s'agissait toujours d'orgie (oral, privé)
? qu'il s'agissait /s'agit/ s'agisse toujours d'orgie me choquait un petit peu
[what shocked me a little bit was that it was always the case of an orgy]

As a consequence of his discourse nature, PSS allows lexicalization even to go beyond grammatical constraints of subcategorisation:

- the specifying part of the utterance may contain direct discourse

(6) ce qui m'a paru bizarre c'est que quand je lui ai dit je vous mets à l'ordre quel ordre Monsieur il m'a dit non non laissez laissez j'ai l'habitude je le ferais moi-même (oral, privé)
[what looked strange to me was that when I said to him I put (on this check) payable to whom sir he answered no no don't bother I know how to manage I will do this by myself]

or a kind of rhetorical self addressed question

(7) ce qui est embêtant c'est que c'est que quelle est l'opération la plus simple en général c'est l'addition (oral, privé)
[what is annoying is (that) is (that) what is generally the simplest operation : it is addition]

- or allows freer contrastive patterns than in direct licensing:

(8) et maintenant dans l'imprimerie ce qu'on demande à un imprimeur c'est non pas d'être un artiste c'est d'être un gestionnaire (oral, professionnel)
[now in printing what you require from a printer is not to be an artist is to be a manager]

In direct lexicalization, mais (but) is needed:

(8’) on demande à un imprimeur non pas d'être un artiste mais d'être un gestionnaire

- allows category mismatch in lists

(9) moi ce que je proposerais au comité de quartier + c'est que nous fassions une commission malgré euh ce qu'on a pu nous dire que il n'était + le projet était pas encore bouclé de s'emparer des des données que l'on a déjà + et de voir nous en tant qu'habitants + ce qu'on souhaiterait qui + enfin ce qui nous inquiète et que le le cabinet qui est en train de donc de plancher sur le projet on lui amène nous aussi des éléments de réflexion + (oral, public)
[as for me, what I would propose to the district assembly is that we set up a committee - in spite of the fact that they said that the project was not completed – in order to consider the data that we already have and to see as neighbors what we would like well what bothers us and the consulting office who is working on the project to bring him elements to think about]

In direct lexicalization, complementizers preferably match:

(9’) ? je proposerais… que nous fassions une commission …
- s'emparer des … données …
- et de voir … ce qu'on souhaiterait …
- et que … on lui amène … des éléments …

- allowing filling by paratactic constructions

(10) ce que je peux rajouter même mieux que ça c'est qu'en fait + elle était la première soliste à l'orchestre moi j'étais le second flûtiste + (oral, public)
[what I can add better than this is that in fact she was the first soloist of the orchestra (and) me I was second flautist]

In direct lexicalization, complementizer que is needed:

(10’) je peux rajouter même mieux que ça qu'en fait + elle était la première soliste à l'orchestre et que moi j'étais le second flûtiste

The discursive and on line nature of PSS can even result in specific strategies based on paratactic syntactic patterns without c'est, in which “the semantic underspecification of the first member let the hearer expect the second” (Blanche-Benveniste, 2010-a):
(11) ce qui m’est arrivé au début + j’ai décollé dans
le vent un peu trop fort (oral, privé)
[what happened to me in the beginning + I
landed off against the wind somewhat too hard]

(12) il y avait une chose chez maman euh elle était
illiterate (oral, privé)
[there was one thing with ma well she was
illiterate]

Claire Blanche Benveniste noticed that all this
patterns can be ordered in a cline, such that : « La
cohésion la plus forte est fournie par le modèle canonique
de pseudo-clivée qui réunit un faisceau de propriétés
grammaticales favorisant la cohésion. D’autres modèles
n’utilisent qu’une partie de ce faisceau de propriétés, la
cohésion la moins forte étant celle des organisations par
parataxe. » (Blanche-Benveniste, (2010-a)

4. From processing repairs to discourse patterns

PSS has to face processing constraints, due to what can be
called the “efficient communication paradox”. On one
side, the indirect lexical specification by extended pseudo
clefts allows the speaker to accurately make his point in
spite of lack of “right word” by means of “periphrasis”.
But a “long” lexical specification puts the main verb
licensing the lexical part out of short time memory and
even introduces irrelevant grammatical material blurring
coherent transition with following discourse units. There
seems to be a way out of the paradox: a “reformulation”
step, using constructions with c’est, clitic “doublings”, etc.
For example, an indirect lexicalization in which the
speaker wants to explain what bothers him (ce qui me
gêne un peu) and which evolves to a long piece of speech
(In square brackets below) becoming more and more
autonomous is “recapitulated” by the word choses
allowing to reintroduce through the verb m’inquiéter
(synonym of gêner) at the end of the discourse unit the
semantic role of the lexicalization (source of bothering for
the speaker):

(13) enfin moi ce qui ce qui me gêne un peu c’est
[aujourd'hui on a + on a un projet hein vous
l'avez l' vous l'avez lu comme moi j'ai entendu
des choses qui m'ont quand même beaucoup
inquiété moi quand ici en réunion publique on
m'a dit deux fois une voie que j'en sois
j'entends parler deux fois deux voies après
j'entends + au niveau logement quand je fais et
et tout tout est acté hein puisqu'il y a euh phase
un il y a euh les logements qui vont être
construits par exemple cette école maternelle
qu'on nous dit qu'elle sera pas euh construite
tout de suite elle est phasée en phase deux +
c'est-à-dire qu'elle est phasée elle est euh +
c'est phasé le budget est là tout tout est là hein
euh je sais pas si vous l'avez lu comme moi si
vous pouvez confirmer je pense (...) hein donc

For Blanche-Benveniste (2010-a), the reformulation
appears as the conclusion of a discourse unit. But it is not
always the case. The reformulation can be a specific move
within larger discourse patterns and be a step for further
clarification. In the following example of discourse pattern,
we can see this scheme: explanation, summary,
synthetic reformulation and clarification

(14) L2: ah oui ah oui + fidéiser le le client c'est
important + surtout les gens âgés ils
aiment bien qu'on s'occupe d'eux + ils
arrivent ici faut faut même si ils doivent
se servir ils aiment bien que qu'on les
serve quand même + ils prêtendent toujours quelque chose pour qu'on qu'on
aille se s- aider et + voilà + des fois les
X il faut les ramener chez eux parce qu'ils ont pris trop de marchandise(s) + donc il
faut les ramener chez eux parce qu'ils sont
ils en ont trop ils peuvent pas marcher +
quand il y a trop de vent quand il pleut +
c'est vraiment à part + c'est vraiment à
part en grande(s) surface(s) c'est sûr qu’on
leur fait pas + ça ils arrivent ils se
débrouillent et + ils rentrent par leurs
propres moyens

L1 : donc là vous pouvez faire la différence
L2 : ouais + c'est ce qu'ils rechercher + les
gens très âgés qui peuvent pas se
déplacer ce qu'ils recherchent c'est la
proximité + puis le la façon de + les petits
commerçants c'est vrai on a le temps de
s'occuper de des gens + en grande surface
ils ont pas le temps + les employés sont
pas là pour ça de toute façon + (oral,
professionnel)

The item proximité synthesizes summarizes the
former long explanation, opening an opportunity for
further clarification (the superiority of small shops over
supermarkets for attending old costumers).

As puts Apotheloz (2008: 91), “we can conclude
from these observations that the identificative
constructions [our PSS] are a central device for the
sequential organization of some discursive patterns. From
this point of view they appear as building the interface
between grammar and discourse”.

It is nevertheless to notice that the reformulation
devices play a complementary part in maintaining
discourse cohesion and coherence when PSS is used. In
the last example the speaker wants to emphasize that it is
important to point out the scandalous attitude of some
occupational doctors:

(15) enfin il y a quand même quelque chose à signaler qui est important c’est que la médecine du travail + lui a demandé de ne jamais parler de son diabète à son employeur + donc ça c’est quand même quelque chose d’assez grave + qu’il faut encore noter parce qu’on rentre en l’an deux mille quand même hein (oral, privé) [well there is something to be pointed out which is important it is that occupational medicine staff asked him not to mention his diabetes to their boss so this is something quite serious that has to be noticed because we are now on the way to the years 2000 anyway, aren’t we]

The clause introduced by parce que qualifies as even more scandalous their attitude as we are entering the years 2000. The reformulation part in italics appears as a necessary part of the whole pattern.

Indeed a coherence gap appears if we erase the reformulation step. In the following example, “parce que” has default scope on the preceding clause and not on the main clause of the PSS:

(15’) enfin il y a quand même quelque chose à signaler qui est important c’est que la médecine du travail + lui a demandé de ne jamais parler de son diabète à son employeur parce qu’on rentre en l’an deux mille quand même hein

5. Conclusion

Beyond PSS patterns, the progressive specification semantic relationship plays a crucial part both in structuring the interface between syntax and semantics and, as a “projection device”, in smoothening “online” building and processing of utterances.

Our next step is a corpus based typology of discourse patterns involving fragments and progressive specification with more registers and comparison with similar facts in other languages (Cresti & Moneglia, 2005).

On methodological grounds, this study shows that it is important to take in consideration larger contexts than one sentence even complex to investigate properly the links between grammar, lexicon and discourse. Such wide scope “useful contexts” (Blanche-Benveniste, 1988) further allow us to sort out in what kind of discourse contexts such complex constructions appear, beyond the argumentative ones pointed out for PSS by Roubaud (2000) and Apothéloz (2008).

6. References


“Subordinate” clauses and syntactic annotation of spoken French

Frédéric SABIO
Aix-Marseille University, Laboratoire Parole et Langage
5 avenue Pasteur, 13100 Aix-en-Provence, France
frederic.sabio@orange.fr

Abstract

The grammatical analysis of clauses introduced by a “subordinating conjunction” has always been a challenge for linguists because, on the one hand, spontaneous spoken data exhibits highly variable syntactic and discursive organizations which have never been properly described through the sentence-based framework of traditional grammar; and on the other hand because continuous reference to the notion of “subordination” tends to unify in an artificial way several types of syntactical configurations that it would be advisable to distinguish carefully. Within the Rhapsodie project (directed by Anne Lacheret, Univ. of Paris Ouest), which is devoted to the syntactic and prosodic tagging of spoken French, we have been directly confronted to such difficulties, and we have had to make some methodological choices which will be the theme of our paper. The tagging system which has been developed both annotates the microsyntactic dependences and the macrosyntactic groupings. Taking those two levels of analysis into account allows us to describe most of the attested uses of conjunctions, including the most problematic ones. The annotation system will be illustrated with a selection of corpus-drawn utterances.

Keywords: syntax; spoken language; French; tagging; subordination; macrosyntax.

1. Introduction

This study has been conducted within the Rhapsodie project (headed by Anne Lacheret, Univ. Paris-Ouest) which is a four-year program (2008-2012) which aimed at annotating a 36,000 words spoken French corpus on both syntactic and prosodic grounds (cf. http://www.projet-rhapsodie.fr.). The ultimate goal of the project was to model the interface between syntax and prosody and to identify the existing correlations between prosodic and syntactic boundaries. The present paper is not meant to give a detailed account of the Rhapsodie framework; it will not even address the diverse aspects of the syntactic annotation system (see Benzitoun et al., 2009, 2010); it will merely illustrate some specific issues regarding the analysis and annotation of “subordinate” clauses. Spontaneous speech seems to be a particularly valuable type of data for the description of sequences which are introduced by so called “subordinating conjunctions”, since it offers a large and somewhat puzzling variety of forms which would not be properly described by the sentence-based framework of traditional grammar.

Before we introduce our annotation system, we will say a few words about the drawbacks of the traditional concept of subordination.

2. Subordination as syntactic dependency

Grammatical tradition quite commonly assumes that any clause which is introduced by a conjunction such as when, because, since or other morphemes of the same kind be automatically regarded as a “subordinate clause” (Riegel et al., 1994). In our view, continuous reference to the notion of subordination tends to unify in an artificial way several types of syntactical and discursive configurations that it would be advisable to distinguish carefully. If we wish to make a reasonable - and somewhat more restricted - use of the concept of subordination, we must stop considering that the conjunctional status of the initial morpheme is a per se, and keep the notion to sequences that share a real dependency relation to the verb of the construction (Debaisieux, 2006a; Deulofeu, 2011).

Obviously, what can be regarded to be a “real dependency relationship” is no simple matter and crucially depends on some theoretical choices. We will refer to the theoretical frame of “Pronominal approach” (Blanche-Benveniste, 1980; Blanche-Benveniste et al., 1984, 1990; Deulofeu, 1991) which postulates that syntactic dependency (“rectional relations”) must necessarily correlate with a set of paradigmatic properties, such as the equivalence with a pronoun, the possibility to be cleft, and a few other features that will be detailed below. The application of these criteria is useful since it enables us to distinguish between clearly dependent sequences, that pertain to the strict domain of syntax and can readily be analyzed as subordinate clauses; and other configurations that do not possess any paradigmatic property, and thus appear to only be linked to the neighboring constructions, sharing with them mere “association”, or paratactic, relations.

The following example will serve to characterize dependant subordinate clauses:

il viendra [quand on lui demandera]
he will come [when he will be asked to]

Here is a set of criteria that show that the When-clause is syntactically dependant on (or governed by) the verb venir (to come), and could therefore be considered as a genuine subordinate clause. The temporal sequence:

(a) could be replaced by a pronominal form such as the interrogative pronoun when or a quasi-pronominal expression like at that moment:

quand est-ce qu’il viendra? [when will he come?]
il viendra à ce moment-là [he will come at that moment]
(b) can occupy a focus position within some sentence-types like cleft constructions, among others:

\[
\text{c'est quand on lui demandera qu'il viendra [it is when he will be asked to that he will come]}
\]

(c) is liable to develop a contrast between positive and negative modality:

\[
\text{il viendra quand on lui demandera et pas quand il le decidera [he will come when he will be asked to, and not when he will decide to]}
\]

\[
\text{il viendra non pas quand il le decidera mais quand on lui demandera [he will not come when he will decide to but when he will be asked to]}
\]

(d) can be modified by a paradigmatic adverbial like seulement, uniquement, surtout (only, mostly):

\[
\text{il viendra seulement quand on lui demandera [he will come only when he will be asked to]}
\]

Here are three corpus-drawn oral utterances in which the clause between brackets is syntactically dependent on the main verb of the construction:

\[
\text{le métier de fleuriste était pas drôle [parce que il fallait avoir les mains dans l’eau] lit: working as a florist wasn’t funny [because you always had to keep your hands in the water]}
\]

\[
\text{nous avons vu une euh euh un crépuscule euh [alors que nous étions d- au au sommet de la mosquée]}
\]

\[
\text{lit: we saw a er – er a twilight [while er we were i- at the at the top of the mosque]}
\]

\[
\text{il chantait à Saint Laurent à la cathédrale [quand il y avait des fêtes] lit: he used to sing at Saint-Laurent in the cathedral [when there were parties]}
\]

In contrast with such canonical examples, the following conjunctural clauses (in brackets) would react in a negative way to the paradigmatic criteria listed above: they have no equivalence with a pronoun, cannot be cleft, and so on.

\[
\text{vos clients euh pourront euh à cet endroit admirer la vue sur le lac et le barrage - [parce que n’oubliez pas que le le Muséoscope surplombe le lac de Serre Ponçon hein]}
\]

\[
\text{lit: your customers er can er in this place admire the sight on the lake and the dam - [because don’t forget that the Muséoscope overhangs the lake of Serre Ponçon]}
\]

\[
\text{ici par exemple c'est du corail qu'elle va porter dans sa corne d’abondance - [alors que là-bas ça sera des fruits]}
\]

\[
\text{lit. here for example it is coral that she is going to carry in her horn of plenty - [while over there that will be fruits]}
\]

\[
\text{[quand je vois les les les les élèves qui descendent dans la rue et tout] moi je les soutiens lit. [when I see the the the the pupils who go down in the street and stuff] me I support them}
\]

In the *Rhapsodie* project, it was essential to make a clear distinction between the syntactically dependent conjunctional sequences, and those that have a non-dependent status. But of course, other aspects had to be taken into account, such as distributional and prosodic properties. We have chosen to study such phenomena in the theoretical frame of macrosyntax (Blanche-Benveniste *et al.*, 1990; Deulofeu, 2003; Sabio, 2012).

### 3. Macrosyntactic patterning

#### 3.1 Presentation

To put it simply, macrosyntax relates to a level of organization which allows the description of sequences which could not be analyzed on the sole basis of their microsyntactic properties, since they share a somewhat discursive relationship with the surrounding context. At the macrosyntactic level, the utterances can be seen as sequences of successive units making up the following pattern:

**Utterance:** [Pre-Nucleus – Nucleus – Post-Nucleus]

What distinguishes those three units has to do with the modality that they are liable to express, certain prosodic properties, and their linear position:

- **The Nucleus** is the basic macrosyntactic unit. It bears an illocutionary value which can be interpreted as a speech act (declarative, question, exclamation), and is liable to form an autonomous utterance. Prosodically, it is associated with a choice of terminal contours that make up a paradigm of prosodic forms, each of them being related to an illocutionary value.

- **The “ad-Nucleus”** (pre- and post-Nucleus) bear no illocutionary value: they seem to be “deactivated” (Verstraete, 2007) as to their capacity to convey any kind of illocutionary content. As a consequence, they cannot constitute an independent communicative unit. Pre- and post-Nucleus are respectively placed before and after the Nucleus unit.

Regarding the way in which micro- and macrosyntax articulate to form utterances, it must be pointed out that in our approach both levels are largely autonomous one from another. This means that two units sharing the same microsyntactic status (that is, with the same syntactic function) may well be realized as two different
macrosyntactic units. Inversely, two elements which have the same macrosyntactical status can fulfill different syntactic functions. Let us consider the following utterances:

\[
\begin{align*}
\text{il n'a pas vu Paul à Paris (mais à Londres)} & \quad [\text{he didn't see Paul in Paris (but in London)}] \\
\text{à Paris, il n'a pas vu Paul} & \quad [\text{in Paris, he didn't see Paul}]
\end{align*}
\]

The Prepositional Phrase à Paris works in both cases as a syntactic adjunct to the verb voir (to see). But their macrosyntactic integration within the utterance is different: in the first case, the locative sequence is part of the Nucleus; in the second utterance, it forms the pre-Nucleus unit.

Here is a second example: the two following utterances share the same macrosyntactic pattern: a Nucleus unit followed by a post-Nucleus unit.

10 ans il avait (en réponse à “il avait quel âge ?”) [10 years old he was (as an answer to “how old was he?”)]

<table>
<thead>
<tr>
<th>10 ans</th>
<th>il avait</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleus</td>
<td>Post-Nucleus</td>
</tr>
<tr>
<td>il est trop jeune je trouve</td>
<td>[he is too young I think]</td>
</tr>
</tbody>
</table>

Table 1: examples of a Nucleus unit followed by a post-Nucleus unit

But their microsyntactic organization is quite different: 10 ans is an object to the verb avait; whereas there is no direct dependency relationship between il est trop jeune and je trouve.

### 3.2 Macrosyntactic annotation

The Rhapsodie annotation system is organized on several levels. This paper will only mention the first level, which is mainly concerned by major grammatical groupings (such as macrosyntactic ones). The following labels are used, which will be illustrated in 4 below:

| 1 marks the end of a macrosyntactic utterance |
| 2 marks the frontier between pre-Nucleus and Nucleus |
| 3 marks the frontier between Nucleus and post-Nucleus |
| 4 Indicates a (macrosyntactic) dependency relationship between two successive macrosyntactic units |

<table>
<thead>
<tr>
<th>Conjunctional sequence Type</th>
<th>a) Dependent on the verb (microsyntactic level)</th>
<th>b) Forms an autonomous macrosyntactic utterance</th>
<th>b) Located in the same macrosyntactic unit as the main verb</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Dependent sequences inside a macrosyntactic unit</td>
<td>+</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>2 Dependent sequences forming a macrosyntactic unit inside the utterance</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3 Dependent sequences forming a macrosyntactic utterance</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>4 Non-dependent sequences inside a macrosyntactic utterance</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>5 Non-dependent sequences forming an macrosyntactic utterance</td>
<td>-</td>
<td>+</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: The 5 syntactic configurations

4. “Subordinate” clauses: a typology

The micro- and macrosyntactic frame which has briefly been introduced above leads us to distinguish between 5 different configurations involving sequences introduced by a “subordinating conjunction”. This typology constitutes an exhaustive classification of all the types of subordinate and “pseudo-subordinate” clauses that we have found in written or spoken French corpora.

The three following features are needed to distinguish between our 5 types:

(a) The conjunctional sequence is / is not dependent on the verb, on a strictly syntactic base (cf. section 2 above).

(b) The conjunctional sequence constitutes / doesn’t constitute an autonomous macrosyntactic utterance (cf. section 3 above).

(c) The conjunctional sequence is / is not located in the same macrosyntactic unit as the main verb (that is: in the same Nucleus, or the same pre-Nucleus Unit, or the same post-Nucleus Unit).

The following table indicates the three features a-b-c on the X-axis, and the 5 syntactic types on the Y-axis:
Each type is tagged in the following way [“CS” for “conjunctional sequence”]:

<table>
<thead>
<tr>
<th>Type</th>
<th>No tagging</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type 2</td>
<td>// CS + &lt; [if the CS is a pre-Nucleus]</td>
</tr>
<tr>
<td></td>
<td>// CS + &gt; [if the CS is a Nucleus]</td>
</tr>
<tr>
<td>Type 3</td>
<td>// + CS //</td>
</tr>
<tr>
<td>Type 4</td>
<td>// CS &lt; [if the CS is a pre-Nucleus]</td>
</tr>
<tr>
<td></td>
<td>// CS &gt; [if the CS is a Nucleus]</td>
</tr>
<tr>
<td>Type 5</td>
<td>// CS //</td>
</tr>
</tbody>
</table>

Table 4: The tags used in Rhapodie

We will now illustrate each of those grammatical types. Due to lack of space, we will not go in much detail but will only present an overview of our typology.

4.1 Type 1: dependent sequences inside a macrosyntactic unit

With this first type, the conjunctional sequences appear to be grammatically integrated both in terms of microsyntax (since they are dependent on the verb) and in terms of macrosyntax (since they are realized into the same unit as the verb itself, showing no detachment of any kind). For example, the conjunctional sequence and the rest of the construction can be placed in the same Nucleus Unit, as in:

// il est parti plus tôt que prévu parce qu’il avait un rendez-vous //
[he went away earlier than expected because he had an appointment//]
// il ne viendra que si cela est nécessaire //
[he will come only if this is necessary//]

But the whole of the construction can be realized in another macrosyntactic Unit, such as a pre-Nucleus Unit:

// si Pierre a l’intention d’arriver quand la réunion sera terminée < autant qu’il reste chez lui //
[if Pierre intends to turn up once the meeting is over < he’d better stay home//]

Those subordinate clauses are obviously the most canonical and easy to describe and annotate, since the micro- and macro-syntactic levels strictly overlap. At the first level of our annotation system, we do not feel the need to specify that the adjunct has been realized as a conjunctional sequence (rather than a Prepositional Phrase or any other category). This is why we only annotate the beginning and end of the macrosyntactic Unit with no internal delimitation.

4.2 Type 2: dependent sequences forming a macrosyntactical unit inside the utterance

This type deals with the conjunctional phrases that are dependent on the verb (as in type 1 above, or type 3 below), but are realized as a specific macrosyntactic unit placed at the initial position of the construction. The conjunctional phrase can either be a pre-Nucleus Unit or a Nucleus Unit.

Here is an example in which the subordinate clause constitutes a pre-Nucleus unit:

// quand ils vont rentrer dans la vie active + ça va être dur pour eux // [oral, corpaix]
[lit. // when they will enter the labour market + it will be hard for them //]

Notice that two labels are used for the annotation: a) the left angle bracket “<”, which signals the frontier between pre-Nucleus and Nucleus Units ; b) the “+” sign, which indicates that there is a dependency relationship between the initial temporal clause and the verb located into the Nucleus Unit.

Here is an example where the subordinate clause has the value of a Nucleus Unit:

Loc.1: // vous allez aller vous promener ? //
Loc.2: // seulement s’il fait beau + > on ira //
[Speaker 1: // will you be going for a walk?//]
[Speaker 2: // only if the weather is fine + > we will be going //]

The last mentioned examples are quite frequent in everyday conversation (Sabio, 2006). The initial clause constitutes the Nucleus Unit, that is, the macrosyntactic element which bears the illocutionary value of an assertion.

As in the preceding example, 2 labels will be useful here: a) the right angle bracket “>”, signaling the limit between the Nucleus and the post-Nucleus. 2) the “+” sign indicating that the two macrosyntactic units are linked by a dependency relationship.

4.3 Type 3: dependent sequences forming a macrosyntactic utterance

Here, the conjunctional clause is once again syntactically dependent on a verb, but it appears to be completely detached from the rest of the construction, in such a way that it forms a completely independent macrosyntactic utterance; thus the construction appears to be realized as a sequence of two successive utterances. Such examples have sometimes been analyzed as “delayed complements” or “supplements” (Debaisieux, 2006b); it appears that the subordinate clause can be detached from the preceding sequence in several different ways:

- In dialogues, it can take the form of a “supplement” which is given by one of the speakers:

   Baga: Et si je ne faisais que dormir comme toi < qui est-ce qui lèverait les impôts ? // Tu dépenses tout pour bouffer. //
   Le roi: //+ Parce que je n’ai rien d’autre à faire.//
   (Architurc, R. Pinget, Ed. Minuit, 16-17)

   Baga: // If I spent my time sleeping as you do < who would levy the taxes? // You spend all the money to buy food. //
   The king: //+ Because I have nothing else to do.//
In the Rhapsodie annotation system, these detached sequences are isolated between a double-slash symbol, based on the fact that they are utterances on their own. In addition, the “+” sign is here to indicate that there is a rectional link between the verb of the first utterance and the subordinate clause of the second utterance.

- Prosodic or graphic cues indicate that such a “detachment effect” can be found in monologues as well:

// quand je sors de la consultation + < je suis euphorique //+ parce que j’ai aimé être avec les gens //

// when the medical examinations are over + < I am thrilled //+ because I like being with people //

With that kind of delayed clauses in the public sector, the conjunction appears to be frequently preceded by a variety of elements; like:

(a) A connective morpheme like et (and) or mais (but):

moi < je préfère une édition originale //+ mais pas parce qu’elle est originale //

me < I prefer an original version //+ but not because it is original //

(b) A negation mark:

// c’est un métier pénible d’accord //+ mais pas parce que c’est un métier privé ou parce que c’est un métier public //

// it is a hard work indeed //+ but not because it is in the private sector or in the public sector //

(c) A paradigmatic adverbial like seulement (only) or surtout (mostly, especially):

// j’aimais pas du tout les cours de français //+ surtout quand on faisait des dictées //

[\[I didn’t like French classes //+ especially when we made dictations //\]

les jeunes en Angleterre < euh quand ils parlent < c’est fou // faut s’accrocher pour comprendre //+ surtout quand tu es pas anglais //

[the young people in England < er when they speak < it’s amazing // it is necessary to hang on to understand //+ especially when you are not English //]

(d) The conjunction can be preceded by a pre-Nucleus Unit like pour moi (for me) or à mon avis (in my view):

// il y allait souvent //+ mais d’après ce qu’on m’a dit < beaucoup moins régulièrement quand l’hiver arrivait // [invented ex.]

[\#/he went there often //+ but as far as I know < much less regularly when wintertime came//\]

A very special pre-Nucleus type we can find in those specific configurations is the expression et ce or et cela (lit. and this), for example:

// il répondait par l’affirmative, //+ et ce parce qu’il en avait toujours été ainsi.// [written ex.]

[lit. // he gave a positive answer //+ and this because he had always done so //]

Let us point out once again that we consider the delayed clause as a syntactically dependant clause. That position is easy to justify on the basis of two examples like:

il a parlé // mais pas à Paul

[he spoke // but not to Paul]

il a accepté de se désister // mais pas en faveur de Paul

[he accepted to withdraw // but not in favor of Paul]

The prepositions à (to) and en faveur de (in favor of) clearly show that the delayed sequence has the grammatical form of a canonical complement.

4.4 Type 4: non dependent sequences inside an autonomous macrosyntactic utterance

We will give very few illustrations for this type:

// vu que ça se transmet par les moustiques < c’est quand même relativement dangereux //

[// since it is a mosquito-borne disease < it is quite dangerous //]

// comme on le sait < il y a pas eu d’effusion de sang //

[// as we know < there has been no bloodshed //]

Here, the underlined clauses have the status of a pre-Nucleus Unit. But in contrast with the second type described above, there is absolutely no dependency relationship between that initial sequence and the verb of the following construction (see section 2 above). C. Blanche-Benveniste (1980) describes the link between such clauses and the following verbal construction as a mere “association” relationship.

One hint to the absence of dependency is the impossibility to develop the conjunctive sequence as a cleft:

• it is since it is a mosquito-borne disease that it is quite dangerous;

• it is as we know that there has been no bloodshed.

In our annotation system, the angle bracket indicates the end of the pre-Nucleus sequence, but (in contrast with type 2, we use no “+” sign, in order to show the absence of any syntactic dependency.
We would adopt the same tagging for sequences placed after the Nucleus Unit, with a right bracket instead of a left one, in order to indicate that the structure is organized as a succession of a Nucleus Unit and a post-Nucleus Unit, as in:

// il y a de la bière dans le frigo > si tu as soif //
[// there is beer in the fridge > if you are thirsty //]

4.5 Type 5: non-dependent sequences forming a macrosyntactic utterance

The last configuration we would like to mention is found in examples like:

// ce film n’a pas du tout fonctionné en France tout du moins // parce que en Amérique + < beaucoup de gens sont allés le regarder // [ex. Debaisieux] // that film had no success at all in France anyway // because en America + < many people went to see it //

// généralement < les mâles sont aussi plus beaux et plus colorés dans la plupart des espèces // bien que chez les poissons comme les Trichogaster leeri < ils sont exactement pareils // [ex. Debaisieux] // usually < males are more beautiful and more colorful in most species // although with fishes like Trichogaster leeri < they look exactly the same //

// vos clients euh pourront euh à cet endroit admirer la vue sur le lac et le barrage // parce que n’oubliez pas que le le Muséoscope surplombe le lac de Serre Ponçon hein //
[// your customers er can er in this place admire the sight on the lake and the dam // because don’t forget that the Muséoscope overhangs the lake of Serre Ponçon //]

In such examples, the conjunctual sequences (because..., although,...) are totally distinct from what precedes them both regarding microsyntax, since no dependency relationship can be postulated between the successive sequences, and macrosyntax, since they form utterances bearing their own illocutionary force.

The last example is particularly striking since it shows that the successive constructions are liable to be associated to two different modality values, that is, a declarative in the first one (“your customers can admire the sight on the lake”), and a command in the second utterance (“don’t forget that the Muséoscope overhangs the lake of Serre Ponçon”). Just to give another example, the following sequence presents a declarative in the first utterance, and a question in the second (which is in fact some kind of a “rhetorical” question):

// on est influençable par rapport à l’anglais > finalement // parce que pourquoi emprunter des mots euh à l’anglais et pas à l’espagnol ou à l’allemand //

[// we are influenced by English > in fact // because why should we borrow words from English instead of Spanish or German //]

In our view, it would be extremely misleading to describe those conjunctional sequences as “subordinates”: all things being equal, the conjunctions seem to behave like connective markers that operate at the discursive level.

The only kind of “independence” they lack is discursive independence, not grammatical one: just like a construction starting with but, therefore or anyway could not be considered as “independent” at the discursive level, the structures illustrated here have to be placed after an utterance on the basis of which they can be interpreted.

5. Conclusion

Spoken data shows that French conjunctions seem to be used in two very different ways: as a syntactic tool liable to achieve microsyntactic integration; and as a discursive marker devoted to macrosyntactic organization. In the past, most of the studies have mainly focused on the microsyntactic structures, which appear to be more canonical and easier to deal with. But the description of spoken data makes it urgent to go into the detail of macrosyntactic aspects of the problem. In the Rhapsodie frame, we have adopted a range of 4 labels (<, >, +, //) which make it possible to annotate both the dependency relations of the conjunctional phrases, and some major macrosyntactic characteristics (such as the fact that conjunctional phrases are liable to form utterances on their own, or the fact that they can be used as an “ad-Nucleus”, bearing no illocutionary value).

6. References


“SUBORDINATE” CLAUSES AND SYNTACTIC Annotation OF SPOKEN FRENCH 435


SPEECH AND SOCIOLINGUISTICS
Nominal agreement in the speech of students from urban areas of Sao Tome

Silvia Figueiredo BRANDÃO

Universidade Federal do Rio de Janeiro/CNPq
Rio de Janeiro, Brazil
silvia.brandao@terra.com.br

Abstract

In this study, performed according to the theoretical and methodological assumptions of variational sociolinguistics, we take up the question of non-implementation of number agreement mark in Noun Phrase (NP) in the speech of Sao Tome, considering individuals from 10 to 18 years in various stages of schooling. It has been designed to test, in speaking of these individuals, the role of variables that were salient for not applying the number mark in the noun phrase (SN). Non-implementation of the nominal plural mark in the speech of students of Sao Tome will depend, among other factors, on the domain or partial knowledge of another language(s) spoken in the region, more interaction with speakers of these languages and on the lower level of education. In the urban variety of Sao Tome, level of education is a variable of primary importance to the distribution of polarized variant patterns of agreement. We discuss the claim of Hagemeijer (2009: 19-20) that, given the linguistic situation of Sao Tome and Principe, which is probably the only country in the Portuguese-speaking Africa where the majority of the population now has Portuguese as first language, there would be conditions for the emergence of a new variety.

Keywords: number agreement; Noun Phrase; Portuguese of Sao Tome; urban variety.

1. Introduction

Questions concerning the loss of inflectional morphology and rules of agreement are important parameters for defining the status of varieties emerging from the contact between linguistically and culturally distinct populations. In this sense, studies about nominal and verbal agreement have served as the basis for the formulation of different interpretations about the emergence and development of varieties of Portuguese, as well as to characterize the Portuguese-based creoles.

Unlike what occurs in relation to the Portuguese of Brazil (PB), there are few studies carried under variational sociolinguistic perspective that focus the nominal agreement in African countries where Portuguese is the official language. In general, studies have been focusing on the Portuguese-based Creole and on cases classified as restructured Portuguese that are observed in rural areas (Baxter, 2009; Figueiredo, 2010). Only recently was awarded the speech of individuals who have Portuguese as L1 and live in urban areas, as found in Brandão (2011a, 2011b), who dealt with this variable in the capital of Sao Tome and Prince, national state which has marked multilingualism.

Brandão (2011a) argues that, among educated speakers, the agreement rule is rated semicategorical, approaching what is seen in European Portuguese, while among those with high school and/or fundamental education, it has variable character, conditioned by linguistic and social factors.

2. Goals

In the current study, we take up the question of non-implementation of number agreement mark in Noun Phrase (NP) in the speech of urban areas of Sao Tome, this time also considering individuals from 10 to 19 years in various stages of schooling. It has been designed to test, in the speech of these students, the role of variables that were salient for not applying the number mark in the noun phrase (NP) according to Brandão (2011b). It starts with the hypothesis that non-implementation of the nominal plural mark in the speech of students of Sao Tome will depend, among other factors, on the domain or partial knowledge of another language(s) spoken in the region, on more interaction with speakers of these languages, on the level of education and particularly on the socio-economic conditions of individuals.

3. The linguistic situation of Sao Tome

In the archipelago of Sao Tome and Prince, located in the Gulf of Guinea, several languages coexist due to a series of historical contingencies related to its colonization process: the Forro (or Santome) and the angolar on the island of Sao Tome, the Lung’ie on the island of Prince, as well as the Creole of Cape Verde, the Portuguese of Tonga and remnants languages from the Bantu group Bantu -, these ones used by a smaller contingent of the population.
The overall index of the absence of plural mark is very low (.05) and is subject to contraints relating to the performance of the individual (Table 1) and the linear and relative position of the constituent in the NP (Table 2).

### INDIVIDUAL PERFORMANCE

<table>
<thead>
<tr>
<th>Informant/ Number of NPs</th>
<th>N</th>
<th>R.W.</th>
<th>Informant/ Number of NPs</th>
<th>N</th>
<th>R.W.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ST-E1-E6m (10 NPs)</td>
<td>8/19</td>
<td>42.1</td>
<td>ST-E6-FDh (36 NPs)</td>
<td>8/8</td>
<td>9.1</td>
</tr>
<tr>
<td>ST-E2E6m (15 NPs)</td>
<td>0/28</td>
<td>0</td>
<td>ST-E7-FDh (17 NPs)</td>
<td>3/3</td>
<td>7.3</td>
</tr>
<tr>
<td>ST-E3-F8h (36 NPs)</td>
<td>0/62</td>
<td>0</td>
<td>ST-E8-FDh (41 NPs)</td>
<td>3/8</td>
<td>3.6</td>
</tr>
<tr>
<td>ST-E4-F8m (98 NPs)</td>
<td>0/26</td>
<td>1</td>
<td>ST-E9-FDhm (44 NPs)</td>
<td>0/9</td>
<td>0</td>
</tr>
<tr>
<td>ST-E5-F8m (15 NPs)</td>
<td>9/26</td>
<td>34.6</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Input: .05  Significance: .000

### LINEAR AND RELATIVE POSITION OF THE CONSTITUENT IN THE NP

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>%</th>
<th>R.W.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-nuclear</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st position</td>
<td>4/262</td>
<td>1.5%</td>
<td>.25</td>
</tr>
<tr>
<td>2nd/3rd positions</td>
<td>2/26</td>
<td>7.7%</td>
<td>.77</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nuclear</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st position</td>
<td>0/12</td>
<td>0%</td>
<td>---</td>
</tr>
<tr>
<td>2nd position</td>
<td>16/268</td>
<td>6%</td>
<td>.62</td>
</tr>
<tr>
<td>3rd/4th positions</td>
<td>3/35</td>
<td>8.6%</td>
<td>.79</td>
</tr>
<tr>
<td>Post-Nuclear</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2nd/3rd/4th positions</td>
<td>6/30</td>
<td>20%</td>
<td>.90</td>
</tr>
</tbody>
</table>

Input: .05  Significance: .000

Table 1: Individual Performance

Table 2: Linear and relative position of the constituent in the NP

Of the nine informants, four categorically applied the rule of canonical agreement. Among the five informants to which the rule is variable, two girls showed a greater tendency not to apply the rule: one of the 6th, another of the 9th grade (R.W. .91 in both cases). The remaining three, all male and attending the 10th or 11th
grade, remained below the rate of .50.

Despite the low input of the rule and the small number of data, this analysis confirmed what has been observed in other studies on nominal agreement in both the Brazilian Portuguese and the Portuguese of Sao Tome: linear and relative position of the constituent in the NP is the most relevant linguistic variable to the presence or absence of number marker So, as shown in Table 2, (a) the marks are concentrated (W. R. .25 and .77) in the area to the left, the pre-nuclear area; (b) in the nucleus and from there marks will be less frequent: (i) the nucleus in the second position: R. W. 62, in the third or fourth, R.W. .79; (ii) constituents on the right, R.W. .90.

All nuclei in the first position (located therefore far left) presented plural mark, a trend also observed in the aforementioned analyses. It is, however, one observation on the behavior of the pre-nuclear constituent in second or third position: the R. W. obtained for the non-implementaion of the plural mark is often far above the reported rate, usually not more than 20 points higher than that observed in the first position.

6. Final remarks
Although we have not done a classical variacionist analysis, since it was based on the speech of a small number of informants and not filling with the same number of informants all social cells, the indication of individual performance as the most important variable for the absence/presence of the plural marker in NP suggests that the agreement, in Sao Tome society, has strong socio-economic-cultural implications. Regardless of the level they are in school, while, in the speech of four students, the rule is categorical, in five others, has variable character in a greater or lesser degree. This, of course, is linked to aspects not controlled in this study and which relate to their family environment, to their greater or lesser exposure to cultural goods, to languages spoken in the region, and to the type of school they attend. It is worth noting the remarks of two of the students who use categorically the rule: one claimed that his father gives him all the means for his intellectual development, and another said that their parents prefer her to study at the Portuguese School because they think that in this school the teachers are better prepared, which, consequently, would provide a better quality of teaching.

are related to the linear and relative position of constituents in NP, which obeys the scale represented in Figure 3 and shows that the marks are concentrated to the left of the nucleus or in the nucleus in first position, decreasing in constituents in the right.

This study, as well as those mentioned here, that is based on corpora of spontaneous speech, and that focus nominal agreement in Portuguese of Sao Tome, have confirmed the observations of Hagemeijer (op. cit) regarding the existence of different "registers" (or standards) dependent on the actuation of socio-economic and cultural factors.

This confirms also the tendencies indicated by Brandão (2011a, 2011b), which outlined, for the urban area, a framework of strong sociolinguistic polarization, despite the low overall rate of absence of plural mark in constituents of the NP.

7. References

<table>
<thead>
<tr>
<th>[+ marks]</th>
<th>[- marks]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-nucleus</td>
<td>Nucleus</td>
</tr>
<tr>
<td>1</td>
<td>2/3</td>
</tr>
</tbody>
</table>

Figure 3: Continuum of marking plurality in the NP constituents in non-European varieties of Portuguese
Abstract

This paper focuses on four linguistic processes in Brazilian Portuguese: (i) the use of subjunctive versus indicative mood in embedded clauses; (ii) the replacement of morphological simple future by periphrastic future; (iii) R-deletion and (iv) vowel harmony. The data are extracted from a corpus of informal interviews with university graduates (standard dialect), stratified for age groups (25-35; 36-55; 56 on), gender and geographical region. The analysis makes use of sociolinguistic methodology (Labov, 1994) and the theory of prosodic hierarchy (Selkirk, 1984; Nespor & Vogel, 1986). We conclude that (i) the use of subjunctive in embedded clauses is related to the semantic/lexical component of the main clause and not all verbs license variable use; (ii) in spoken language the morphological simple future has been replaced by periphrastic forms and the hypothesis is that children incorporate the simple morphological future only in school; (iii) there is a gradual process of R-deletion and even the IP and PhP boundaries no longer inhibit deletion of the segment; (iv) vowel harmony process shows stability in Brazilian Portuguese and similar behaviour in all cities. In order to have a clear picture of all processes it is necessary to understand the interplay of grammatical, prosodic and social constraints.

Keywords: variation; subjunctive mood; periphrastic future; R-deletion; vowel harmony.

1. Introduction

The aim of this paper is to discuss four variable linguistic processes in standard dialects of Brazilian Portuguese: (i) the use of subjunctive versus indicative mood in embedded clauses (eu não acho que seja/é ‘I do not think that it be/is’); (ii) the ongoing replacement of the morphological simple future by the periphrastic future (cantarei ‘I will/shall sing’ ~ vou cantar ‘I am going to sing’); (iii) R-deletion (cantaØ ~ cantar ‘to sing’) and (iv) vowel harmony (pirigo ~ perigo ‘danger’).

All analyses are based on spoken corpora -- informal interviews --, collected in the 70’s and in the 90’s, with University graduates (standard dialects), in urban centers of Brazil, Salvador, Recife (Northeastern region), Rio de Janeiro, São Paulo (Southeastern region), and Porto Alegre (Southern region). The samples are stratified for age (1= 25; 2 = 36-55; 3 = 56 on) and gender. These speech samples have been built within the Project “Estudo da norma linguística urbana culta (NURC)” and more than 1500 hours of standard dialect are available for research. The analysis makes use of sociolinguistic methodology (Labov, 1994) and VARBRUL/GOLDVARB computational programs.

2. Subjunctive versus indicative

The usual explanation for the variable use of subjunctive versus indicative mood in Brazilian Portuguese is that there is a difference in meaning between the two constructions: the indicative mood expresses factual reality and the subjunctive mood -- considered by traditional grammar the prototypical mood of subordination -- expresses eventuality and potentiality (the irreals hypothesis).

This variable use is not restricted to Portuguese and has been also attested in other Romance languages such as French (Poplack, 1992) and Spanish (Rivero, 1971; Bosque & Demonte, 1999). Mattos e Silva (1989: 741) points out that this alternation has been in use since the 13th century.

The subjunctive/indicative mood variation occurs not only in adverbial (1), but also in embedded clauses (2), although with different rates.

(1) Embora o homem diga/diz que está pobre
Although the man says that (he) is poor

(2) A mãe de Maria não quer que ela vá/veai
Mary’s mother does not want that she go(es)

The use of subjunctive in embedded clauses -- around 20% -- is related to the semantic/lexical component of the main clause (the matrix verb). Not all verbs present variable use of the subjunctive.

<table>
<thead>
<tr>
<th>Verbs of ‘opinion’</th>
<th>Oco/total</th>
<th>% Subj.</th>
<th>% Ind.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acreditar/crêer (believe)</td>
<td>34/50</td>
<td>68%</td>
<td>32%</td>
</tr>
<tr>
<td>Supor (suppose)</td>
<td>04/04</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>Achar (think)</td>
<td>123/1046</td>
<td>12%</td>
<td>88%</td>
</tr>
<tr>
<td>Pensar (think)</td>
<td>05/16</td>
<td>31%</td>
<td>69%</td>
</tr>
<tr>
<td>Parecer (seem)</td>
<td>01/54</td>
<td>4%</td>
<td>96%</td>
</tr>
</tbody>
</table>

Table 1: Frequency of subjunctive/indicative mood, according to each verb

Comparing dialects (Figure 1 below), we can see that there is a more significant difference of use between the three cities with two verbs: ‘acreditar’ (believe) and ‘pensar’ (think).
Figure 1: Frequency of use of each verb in each city

Three significant factor groups were pointed out in all dialects. The subjunctive mood (23% - input .24) is more frequent when the verb is in the first person rather than in the third person; there is a negative particle in the matrix clause; and the matrix verb is in the past tense, as in example (3), from Callou & Almeida (2009).

Table 2: Person of the matrix verb

(3) eu pensei que fosse alguma coisa que ele tivesse roubado ... 
I thought that it was something that he had stolen ...

Table 3: Negation effect

(4) eu não acho que casar e ter filhos seja uma coisa natural, da vida
I do not think that getting married and having children be a natural thing, of life

The embedded clause analysis reveals age-group differentiation, when the verb believe ‘acreditar’ is pointed out (Figure 2): older -- rather than younger -- speakers use the subjunctive more often. Regional and time variables also play a role in mood choice: the use of subjunctive forms is less frequent in Rio than in Salvador (Figure 3), once more, with the verb ‘acreditar’ (believe); from the 70’s to the 90’s, the use of subjunctive mood is related to the lexical item (Figure 4).

3. Periphrastic future versus simple morphological future

In Portuguese, future tense is mainly expressed by two simple forms (morphological simple future; simple present tense + obligatory time marker) or by periphrastic forms (present/future tense of modal auxiliary verb ir (‘to go’) + main verb infinitive). In contemporary spoken Brazilian Portuguese the morphological simple future has been replaced by periphrastic forms, except when the auxiliary and the main verbs are the same, as in example (5) below.
(5) eu vou ir ao cinema
‘I will go to the movies’

Nowadays, the use of haver+de+infinitive is very rare and put emphasis on the action.

(6) Hei de trazer o livro amanhã
‘I will bring the book tomorrow for sure’

<table>
<thead>
<tr>
<th>spoken language</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>morphological simple future</td>
<td>7%</td>
</tr>
<tr>
<td>periphrastic form (ir+inf.)</td>
<td>77%</td>
</tr>
<tr>
<td>simple present tense</td>
<td>16%</td>
</tr>
<tr>
<td>Tokens</td>
<td>393</td>
</tr>
</tbody>
</table>

Table 4: Future constructions in contemporary Brazilian Portuguese

Nevertheless, the grammaticization process in Portuguese is still in progress, and a complete merger of adjacent elements has not yet occurred (Oliveira, 2006) and the two elements maintain a certain degree of independence, allowing insertion of adverbs between the auxiliary and the main verb:

(7) ela vai simplesmente escrever…/ * she will simply write…).

We conclude that variation between simple and periphrastic forms is a reflection of competition between two grammars, following Kroch’s proposal (1994), the same way as variation of ter/haver-existential constructions. Language acquisition researches have shown that children incorporate the simple morphological future to their lexical inventory only on exposure to a wider range of written language in school.

4. R deletion

Regarding R, our hypothesis is that, besides linguistic and social factors, such as morphological class – non-verbs (mar ‘sea’) versus verbs (cantar ‘to sing’) -- age group and region, the prosodic structure also plays a role in the loss of the segment in final coda position. We postulate that the domain of deletion is not the syllable but rather a prosodic boundary, i.e., this phenomenon is also prosodically motivated.

Similar to other segmental phenomena, as external sândi, for instance, which takes into consideration prosodic constituent boundaries (Bisol, 1996, 2002; Tenani, 2002), the hypothesis is that R-deletion is also conditioned by the position of the syllable as regards the edge of the prosodic domain:

prosodic word (Pw) -- A prosodic word has one and only one primary accent and a PW max has one and only one prominent element (Vigário, 2003).

A prosodic word is, for instance, the domain of dactylic lowering and neutralization in the direction of a high vowel in Brazilian Portuguese (Battisti & Vieira, 1996);

phonological phrase (PhP) -- A phonological phrase should contain more material than one prosodic word (Frota, 2000; Tenani, 2002) and the domain of ϕ-formation is defined by the configuration […] Lex XP […]Lexmax (where Lex stands for the head of a lexical category, and Lexmax for the maximal projection of a lexical category). In Brazilian Portuguese, ϕ characterizes itself by regular occurrence of a pitch accent in its more prominent element (Frota & Vigário, 2000; Tenani, 2002; Fernandes, 2007); or

intonational phrase (IP) -- The domain of IP may consist of all the ϕs in a string that is not structurally attached to the sentence tree or any remaining sequence of adjacent ϕ in a root sentence (Nespor & Vogel, 1986). Long phrases (in number of syllables and/or prosodic words) tend to be divided in the same way as small phrases tend to form a unique IP with an adjacent IP, i.e, balanced phrases are preferred (Frota, 2000; Serra, 2009). In Brazilian Portuguese, the domain of IP is indicated by a nuclear contour (pitch accent + boundary tone) and a potential pause in its right boundary. There is also a preferential occurrence of L+H* associated to the first stressed syllable of IP, no matter this syllable is the most prominent of (ϕ (Tenani, 2002; Moraes, 2007; Serra, 2009; Silva, 2011).

Taking into consideration these three domains, R deletion would be more frequent at lower levels rather than at higher levels, as we can see in example (8):

(8) [(pra sair)pw Lhp IP [[[teO]pw Lhp [lque ficaO]pw (quietinho)pw Lhp]IP / to go out (to) have to keep quiet

Data from Votre (1978) and from Gomes (2006) – adult and child speech, respectively, have shown that the presence of a pause -- durational trace frequently associated with the right edge of IP --licenses R realization. This reasoning represents another argument in favor of our hypothesis.

In recent research about coda acquisition, in European Portuguese, Jordão (2009) asseverates that the final position of IP clearly favors not only the reconstruction strategies but also the realization of coda.

Moreover, this interpretation could be able to explain the higher frequency of deletion in final coda position (46%) and lower frequency in internal coda position (3%) – Callou et al., 1998.

This analysis is restricted to age group from 25 to 35 years old, male and female, confronting Rio de Janeiro and Salvador data, in order to explain the
trajectory of the phenomenon from initiation to completion, as far as R-deletion was strongly concentrated on speakers of this age group (72%), at least, at the beginning of the process. We make use of sociolinguistic methodology (Labov, 1994) and the theory of prosodic hierarchy (Selkirk, 1984; Nespor & Vogel, 1986).

In Rio de Janeiro, R-deletion may be considered a midrange change, and in Salvador a change nearing completion, affecting almost every word in which the given sound appears, no matter whether a verb (97%) or non-verb (78%), as we can see in Figure 5.

![Figure 5: R deletion in final coda position, in standard dialect, in Rio de Janeiro and Salvador, in the 70's, according to morphological class](image)

This analysis confirms previous studies with several different samples which have always pointed to morphological class (verbs / non-verbs) as the predominant conditioning factor of this sound change: R-deletion is much more frequent in verbs, although it conveys semantically relevant information, for it is a marker of the infinitive and of the subjunctive mood (querer ‘to want’; se eu quiser ‘if I want’).

If we compare Rio de Janeiro dialect in real time, in the 70’s and in the 90’s, we will be able to say that R-deletion has continued to advance (Figure 6) and is always conditioned by morphological class.

![Figure 6: R deletion in final coda position, in standard Rio de Janeiro dialect, in the two decades](image)

In Salvador, it is possible to affirm that among young speakers, in the 90’s, R-deletion process is completed, no matter the word in which the segment is inserted is a verb (100%) or a non-verb (99%).

According to the hierarchy prosodic hypothesis, R deletion would be more frequent at lower levels rather than at higher levels. The multivariate analysis of 232 tokens allows to conclude that IP and PhP boundaries favor the preservation of the segment while PW favors R-deletion, in the 70’s.

The opposition between verbs and non-verbs remains significant and must be taken into consideration, since it is only if we analyze each boundary separately that it is possible to have a wider vision of the process. At least, at the 70’s, in Rio de Janeiro dialect, R-deletion in non-verbs is restricted to word boundary (PW).

There is a gradual process of deletion and from the 1970’s to the 1990’s even the IP and PhP boundaries no longer inhibit deletion of the segment (Figure 7).

![Figure 7: R deletion in final coda position, in standard dialect, Rio de Janeiro dialect, in the two decades, according to prosodic boundary](image)

To sum up, we are still trying to understand the interplay of grammatical, prosodic and social constraints which governs R-deletion in Brazilian Portuguese.

5. Vowel harmony

Traditionally, vowel harmony is defined as the raising of pre-stressed mid vowels e and o due to high vowels i or u in the stressed syllable (perigo → pirigo ‘danger’; coruja→curuja ‘owl’). It can also apply to the lowering of pre-stressed mid vowels in the environment of a low vowel in the stressed syllable, as in bolota ~ b[ɔ] ‘ball’; Pele~ P[ɛ] ‘Brazilian soccer player’.

Vowel harmony process shows stability in Brazilian Portuguese, although it is a process almost completed in European Portuguese since the 15th century. The analysis has shown that the target vowels /e/ and /o/ behave differently in Brazilian Portuguese. We observe that vowel harmony is a split phenomena as far as raising of pre-stress mid vowels can be obtained either by the quality of adjacent syllable high vowel or due to the articulatory or acoustic assimilation of neighboring adjacent consonants: moqueca → [m][a]queca “kind of food”; boneca → [b][u]neca “doll”; pomada → [p][u][m]ada “cream”; colher → [k][a]lher “spoon”.

The comparison of mid vowel raising in five Brazilian cities -- São Paulo (SP), Rio de Janeiro (RJ), Salvador (SSA) and Recife (RE) -- shows a similar
behavior: almost the same general input and conditioning environments, as related above.

The trapezoid form of the mouth cavity allows a larger vertical space for the production of front vowels than the vertical space for the production of back vowels. Within this hypothesis [i] is higher than [u] (Bisol 1989) and this would explain why [i] is a better trigger than [u]. Bisol’s results are based on Porto Alegre data.

Acoustic studies of Brazilian stressed vowels (Moraes, Callou & Leite, 1996) shows, however, that the articulatory explanation does not work in all Brazilian dialects. In Recife, Salvador, São Paulo for instance [i] and [u] have the same F1 value. So F1, related to vowel height, can not be the explanation for the asymmetric behavior of i / u.

An alternative hypothesis is that the distinctive feature for back vowels is not degree of openness but degree of labialization (lip rounding). Figure 1 shows that the acoustic space of [o] and [u], based on F1 and F2 plotation, is practically the same, reinforcing this hypothesis. If it is rounding that is the distinctive feature for back vowels, Brazilian vowel system is asymmetrical, as far as for front vowels the distinctive feature is height while for back vowels it is roundness.

6. References


Banco de dados sociolinguísticos do Norte do Brasil

Regina CRUZ1,2, Carlos NEDSON1, Raquel COSTA1, Josivane SOUSA3, Socorro CAMPOS1, Orlando CASSIQUE1, Doriedson RODRIGUES1, Mara COSTA2,4

1UFPA; 2CNPq; 3PMPA; 4Bolsista PIBIC
Av. Augusto Correa, s/n – Campus do Guamá – Belém (PA) – 66075-900
E-mail: regina@ufpa.br, nedson@ufpa.br, raqmaria@ufpa.br, josivane@yahoo.com.br, bmscampos@ufpa.br, cassique@ufpa.br, doriedson@ufpa.br, mara.costa@il.com.br

Abstract

This paper presents how they formed corpora for study of the unstressed mid vowel of the linguistic varieties of Brazilian Portuguese (PB) spoken in Amazon are being organized, processed and annotated. The NORTE VOGIAS Project aims to verify the variations of unstressed mid vowel in Amazon PB to provide a sociolinguistic configuration of the phenomena like vocalic harmony or rising in Pará state, for example. So far the formed corpora are from the following cities: Belém (Sousa, 2010; Cruz et al., 2008); Cametá (Rodrigues & Araújo, 2007; Rodrigues & Reis, 2012; Costa, 2010); Mocajuba (Campos, 2008); Breves (Cassique et al., 2009; Dias et al., 2007) and Breu Branco (Marques, 2008). The NORTE VOGIAS project's team has been investigating three vowel processes in variation: a) unstressed (pretonic) vowel mid rising; (Cruz, 2012, 2010; Sousa, 2010; Rodrigues & Araújo, 2007; Campos, 2008; Cassique et al. 2009; Dias et al., 2007; Marques, 2008); neutralization of non-final post-tonic vowel (Costa, 2010) and allophonic nasalization (Rodrigues & Reis, 2012). The NORTE VOGIAS project has speech samples of 342 PB speakers from Amazon in its database and it is linked to PROBRAVO team.

Key words: sociolinguistic corpora; Amazon Brazilian Portuguese; PROBRAVO project; pretonic mid vowel; linguistic variation.

1. Introdução

Desde 2007, quando passou a integrar o grupo PROBRAVO, o projeto Norte Vogais já efetuou estudos do processo de variação das vogais médias pretônicas do português falado em cinco localidades do Estado do Pará, a saber: i) Cametá (Rodrigues & Araújo, 2007; Rodrigues & Reis, 2012; Costa, 2010); ii) Mocajuba (Campos, 2008); iii) Breves (Cassique et al., 2009; Dias et al., 2007); iv) Belém (Sousa, 2010; Cruz et al., 2008) e; v) Breu Branco (Marques, 2008; Coelho, 2008; Campelo, 2008). Todas são descrições sociolinguísticas de cunho variaçãoista e apresentam um tratamento quantitativo dos dados, que possibilitam uma comparação de seus resultados quanto ao fenômeno estudado, no caso as vogais átonas. São justamente estes procedimentos que passaremos a detalhar no presente trabalho.

2. Projeto Norte Vogais

O projeto Norte Vogais está diretamente ligado ao Diretório nacional de pesquisa do CNPq PROBRAVO, coordenado por Dr. Marco Antônio de Oliveira (PUCMG) e Dr. Seung-Hwa Lee (UFMG). O grupo de investigadores do PROBRAVO realiza uma investigação multidisciplinar – sócio-histórica e linguística – para descrever as realizações fonéticas das vogais nos dialetos do Sul ao Norte do Brasil. Até o presente momento cinco regiões foram investigadas no Estado do Pará: Belém, Breves, Cametá, Mocajuba e Breu Branco, tanto nas suas zonas rurais quanto urbanas.

De maneira geral, a equipe da UFPA pretende ao mesmo tempo caracterizar o sistema vocalico átono e suas variantes, com base em amostra estratificada e em termos variaçãoistas, assim como analisar e explicar o processo de variação das vogais médias pretônicas e postônicas nãofinais no português falado no Norte do Brasil interna e qualitativamente.

3. Fenômenos investigados

As descrições sociolinguísticas empreendidas pela equipe da UFPA priorizam a investigação de três aspectos fonéticos em particular: a) a variação das vogais médias pretônicas; b) a variação das vogais médias postônicas médias e; c) a nasalidade alofónica, cujos detalhes são fornecidos nesta seção.

3.1 Vogais médias pretônicas

Muitos estudos já foram realizados sobre as vogais médias em posição pretônica no Brasil. Elencamos aqui, a partir de uma sucessão temporal, aqueles realizados na região Norte: Rodrigues (2005) sobre o alteamento /o>/ [u] no português falado em Cametá (PA); Dias et al. (2007) sobre a alteração na falá rural de Breves (PA); Oliveira (2007) sobre a harmonização vocalica no português urbano de Breves (PA); Araújo & Rodrigues (2007) sobre as vogais médias /e/ e /o/ no português falado no município de Cametá (PA); Cruz et al. (2008) sobre a harmonização das médias pretônicas no português falado nas ilhas de Belém (PA); Campos (2008) sobre o alteamento vocalico em posição pretônica no português falado no município de Mocajuba (PA); Marques (2008) sobre o alteamento das vogais médias pretônicas no português falado no município de Breu Branco (PA) e; Sousa (2010) sobre a variação das vogais médias pretônicas no português falado na área urbana do município de Belém (PA).
Em sua maioria as descrições sociolinguísticas realizadas pelo projeto Norte Vogais investigaram as vogais médias pretônicas na perspectiva do alteamento (Rodrigues & Araújo, 2007; Oliveira, 2007; Campos, 2008; Marques, 2008; Cassique et al., 2009; Sousa, 2010). Apenas Dias et al. (2007) e Cruz et al. (2008) analisaram o fenômeno de variação das médias pretônicas na óptica da harmonização vocalica. De forma generalizada, os dados demonstraram uma tendência ao não alteamento nos dialetos paraenses. Os resultados sobre o alteamento confirmaram a afirmativa de Bisol (1981) de serem as vogais altas na sílaba seguinte um contexto altamente favorecedor (Rodrigues & Araújo, 2007; Dias et al., 2007; Campos, 2008; Cruz et al., 2008; Cassique et al., 2009). Outro resultado convergente diz respeito ao fato de os dados de fala de informantes de mais baixa escolaridade e de maior faixa etária apresentarem maior probabilidade de alteamento.

Como se pode constatar avançou-se bastante nas descrições sociolinguísticas das vogais médias pretônicas no português falado na Amazônia Paraense, os procedimentos metodológicos adotados foram comuns, principalmente no que diz respeito a formação dos corpora e tratamento dos dados.

### 3.2 Vogais postônicas não-finais

O único trabalho sobre postônicas medias realizados no seio do PROBRAVO pela equipe da UFPA é o de Costa (2010). A autora verifica o comportamento das vogais médias /e/ e /o/ em posição postônica não-final de itens lexicais no português falado nas áreas urbana e rural do município de Cametá. O corpus foi constituído com amostras de fala de 96 informantes estratificados em sexo, faixa etária, nível de escolaridade e procedência. A coleta dos dados foi realizada através de dois tipos de entrevista: a livre (48 informantes); e o teste ou nomeação de figuras (48 informantes).

O corpus apresenta 2.177 dados, sobre o qual se observou a partir de uma análise estatística, no programa computacional Varbrul, considerando variáveis linguísticas e não linguísticas, que o fenômeno de alteamento com peso relativo de .46 apresenta probabilidade menor de ocorrência do que o não alteamento com peso relativo de  .54.

Este trabalho apresenta igualmente uma análise qualitativa do comportamento das vogais médias - /e/ e /o/ - postônicas não-finais, as quais apresentam quatro variantes possíveis: manutenção /e[/l[o] , alteamento [i][u] , apagamento [ø] e abaxamento [E][O].


### 3.3 Nasalidade alofônica


Gráfico 1: Tendência da nasalidade alofônica do norte ao sul do Brasil. Fonte: Cruz (2010: 253)

Consta-se, portanto, que há um declínio da nasalidade do norte ao sul do Brasil. O índice baixo da variedade de Breves parece não contrariar tal tendência, uma vez que Breves tem indícios de apresentar uma situação sociolinguística particular que será comentada na secção 6.

### 4. Procedimentos metodológicos adotados por projetos

Os dados foram coletados em trabalho de campo, com gravações em áudio. Para a coleta destes, priorizaram-se as narrativas de experiência pessoal nos moldes da teoria da variação (Tarallo, 1988). Utilizou-se para cada variedade investigada uma amostra estratificada em sexo, faixa etária (15 a 25 anos; 26 a 45 anos e acima de 46 anos) e escolaridade (analfabeto, fundamental, médio e superior).

Uma vez as gravações concluídas, os dados obtidos foram transcritos grafematicamente observando os parâmetros da Análise da Conversação (Castilho, 2003).

Um arquivo contendo a triagem dos dados, tomando como unidade de análise o grupo de força
como estabelecido por Câmara Jr. (1969), foi criado, por informante. Uma cópia do mesmo foi feita, para nela se proceder à transcrição fonética do vocábulo contendo o fenômeno estudado. Utilizou-se para a transcrição fonética o alfabeto SAMPA\(^2\).

Uma vez a transcrição fonética concluída, procedeu-se à codificação dos dados. Para os estudos sobre vogais médias pretônicas, utilizou-se o mesmo arquivo de especificação do PROBRAVO, de autoria de Orlando Cassique e Doriedson Rodrigues. Costa (2010) e Rodrigues & Reis (2012), por conta da especificidade de seus estudos, utilizaram arquivos de especificação mais adequados a seus objetos de estudo.

De maneira geral, os arquivos de especificação contêm fatores de diversas naturezas: a) fonéticos b) morfológicos; c) sintático entre outros, além dos fatores sociais. Por último, realizou-se o tratamento estatístico dos dados pelo programa VARBRUL.

5. Caracterização dos corpora formados

Os corpora do projeto Norte Vogais possui um número total de informantes variando de 24 (vinte e quatro) a 72 (setenta e dois), como podemos visualizar no Quadro 1 abaixo.

<table>
<thead>
<tr>
<th>Localidade</th>
<th>Total de informantes</th>
<th>Fonte</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breves (urbano)</td>
<td>42</td>
<td>Oliveira (2007)</td>
</tr>
<tr>
<td>Breves (rural)</td>
<td>36</td>
<td>Dias et al (2007)</td>
</tr>
<tr>
<td>Belém (urbano)</td>
<td>48</td>
<td>Sousa (2010)</td>
</tr>
<tr>
<td>Belém (rural)</td>
<td>24</td>
<td>Cruz et al (2008)</td>
</tr>
<tr>
<td>Cametá</td>
<td>48</td>
<td>Costa (2010)</td>
</tr>
<tr>
<td>Mocajuba</td>
<td>48</td>
<td>Rodrigues (2005)</td>
</tr>
</tbody>
</table>

Quadro 1: Número total de informantes do Projeto Norte Vogais por variedade investigada com a indicação da fonte de cada estudo realizado. Fonte: Atualizado de Cruz (2012: 200)

<table>
<thead>
<tr>
<th>Localidade</th>
<th>Total de informantes</th>
<th>Duração total das gravações</th>
<th>Fonte</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breves</td>
<td>42</td>
<td>10 h 35 min</td>
<td>Oliveira (2007)</td>
</tr>
<tr>
<td>Belém</td>
<td>48</td>
<td>15 h 28 min</td>
<td>Dias et al (2007)</td>
</tr>
<tr>
<td>Cametá</td>
<td>120</td>
<td>45 h 21 min</td>
<td>Sousa (2010)</td>
</tr>
<tr>
<td>Mocajuba</td>
<td>48</td>
<td>24 h 21 min</td>
<td>Costa (2009)</td>
</tr>
<tr>
<td>Breu Branco</td>
<td>24</td>
<td>4 h 24 min</td>
<td>Rodrigues (2005)</td>
</tr>
</tbody>
</table>

Quadro 2: Tamanho do corpus do Projeto Norte Vogais em horas de gravação

O Projeto Norte Vogais do Brasil conta com um banco de dados de amostra de fala de 342 (trezentos e quarenta e dois) informantes nativos da Amazônia Paraense, originários de cinco variedades locais: Belém, Cametá, Breves, Breu Branco e Mocajuba, em suas zonas rural e urbana.

Além das transcrições, o corpus contém o áudio das gravações realizadas em trabalho de campo. O Quadro 2 contém uma descrição do corpus em horas gravadas.

6. Tendência do Português da Amazônia Paraense

De forma geral, as descrições sociolinguísticas realizadas sobre o português falado na Amazônia Paraense tem demonstrado uma tendência à não aplicação da regra de alteamento das vogais médias em posição pretônica, como podemos constatar no Quadro 3 abaixo.


Outro resultado relevante compreende a inexpressiva ocorrência de vogais médias baixas nas  

\(^2\)http://www.phon.ucl.ac.uk/home/sampa/index.html
posição átona. Tais resultados contrariam um estudo a divisão dialética de Nascente que caracteriza o dialecto do Norte do Brasil como apresentando uma tendência à realização das vogais médias abertas nas posições átonas, em oposição aos dialectos do Sul do Brasil que prefeririam as vogais médias fechadas. Por outro lado os resultados reforçam a hipótese de Silva Neto (1957) de que o Pará compreenderia uma ilha dialética na classificação de Antenor Nascente entre os dialectos do Norte do Brasil. Silva (1989) menciona nos seus resultados, uma predominância das vogais baixas no seu corpus formado com amostras de fala do dialeto alvo – o de Salvador, – que fora confrontado com amostras de fala de 50 pontos do território baiano e de uma localidade do estado de Sergipe emprestadas, respectivamente, do Atlas Prêvio dos Falares Baiano e de Mota (1979).

Os resultados dos estudos empreendidos pela equipe do Projeto Vozes da Amazônia têm buscado prioritariamente caracterizar o português regional paraense. Nesse sentido, os resultados sobre as vogais médias pretônicas têm demonstrado uma tendência ao uso de suas variantes com probabilidade de maior ocorrência de manutenção das médias pretônicas em decorrência do alentamento das mesmas, inclusive com índices percentuais muito próximos de ocorrência da manutenção das médias pretônicas entre as variedades investigadas (Breves (rural), Belém, Cametá e Mocajuba). Duas, das variedades investigadas (Breves (urbano) e Breu Branco) confirmam a tendência à manutenção, mas apresentam percentuais muito distantes das quatro outras variedades comparadas.

Os resultados do estudo da variação das médias pretônicas no português da Amazônia paraense mostram que os percentuais de alteamento são muito baixos de modo geral nas zonas dialetais do Pará. Os índices mais distantes de Breves (33%) e de Breu Branco (24%), por indicarem a necessidade de uma investigação mais aprofundada sobre a situação sociolingüística destes dois municípios em particular, levaram a equipe da UFPA vinculada ao PROBRAVO a lançar uma nova edição do *Vozes da Amazônia* destinada a investigar o português falado nas zonas de migração do Pará. Breves e Breu Branco apresentam em comum o fato de terem sido justamente regiões que receberam um fluxo migratório considerável em decorrência de projetos econômicos da região.

O município de Breves sozinho apresenta um terço da população de todo arquipélago marajoara. O inchaço populacional sofrido por Breves se deu no segundo ciclo da borracha, durante a segunda guerra mundial, quando o governo apostando em um crescimento econômico oriundo da borracha, fez vir nordestinos para trabalharem na exploração da borracha na Amazônia, os ditos soldados da borracha. Uma vez terminada a guerra e o declínio do segundo ciclo da borracha, os imigrantes nordestinos não tiveram como voltar para a sua terra de origem e fixaram residência obrigatoriamente na Amazônia, uma boa parte deles ficou justamente na cidade de Breves.

Breu Branco é um dos municípios de criação recente no Pará, seus moradores, em sua maioria, são brasileiros originários de diferentes regiões do Brasil – mineiros, paulistas, gaucho, paranaenses, maranhenses, cearense, piauiense, tocantinenses – que migraram para o Pará para trabalhar na construção da hidrelétrica de Tucuruí na década de oitenta. Com a conclusão da primeira etapa dos trabalhos de implantação da Hidrelétrica de Tucuruí, a maioria desses trabalhadores fixou residência nos municípios da região. Desta forma a população atual de Breu Branco se assemelha a de Brasília. Breu Branco, portanto, apresenta a mesma situação linguística atestada em Brasília (DF) e no sul do Pará onde por questões econômicas – no caso de Breu Branco (PA) tal situação foi ocasionada pela construção da hidrelétrica de Tucuruí – vários dialectos do português brasileiro convivem em uma mesma localidade, oscilando de tal contato dialético uma nova norma linguística.

Os resultados dos estudos sobre as vogais médias das variedades da Amazônia Paraense demonstram que estas duas variedades investigadas fomem completamente a uma característica comum das variedades da Amazônia paraense que é a quase neutralização da variação entre as médias. As variedades de Breu Branco (próximo a Tucuruí) e da zona urbana de Breves (no Marajó) têm como pontos em comum o fato de serem localidades que receberem uma forte migração de falantes do português de outras regiões do Brasil por conta de projetos econômicos. Essas regiões não possuem marcas de identidades (e aí em todos os sentidos) com a Amazônia paraense, e tudo indica inclusive na variedade linguística.

Nossa hipótese é a de que os fatores externos são relevantes no condicionamento da realização das variantes das médias pretônicas e fazem com que tais variedades sejam muito diferentes das demais da Amazônia Paraense. Para corroborar tal hipótese procederemos a uma nova coleta de dados, controlando como principal fator a origem ou ascendência do falante, como fez Bortoni-Ricardo (1985). Acreditamos ser talvez o fator que esteja controlando a realização dessas variantes. Verificaremos também além da variável origem do falante, o fator faixa etária, em especial a faixa dos mais jovens, a fim de se verificar se se trata de uma variação estável ou mudança em progresso.

Como última hipótese, acreditamos que nas regiões em questão – Breu Branco e Breves - ainda não se cristalizou uma nova norma resultado do contato intervariedades, como ocorrido em Brasília, e o fato desta nova norma ainda não ter sido estabelecida resulta em contraste muito acentuados da realização das variantes atestadas.

Os resultados sobre a nasalidade vêm justamente fortalecer nossa hipótese de sustentação de uma
investigação diferenciada para o português falado nas zonas de migração, uma vez que os dados de Breves (Cassique, 2002) contrariam a tendência da nasalidade do português falado no Norte que seria de ocorrência de alto índice de nasalidade.

7. Conclusão

O presente trabalho apresenta os corpora formados pela equipe do Projeto Norte Vogais vinculado ao PROBRAVO que estuda prioritariamente o vocalismo átono no Norte do Brasil, mais especificamente na Amazônia Paraense.

O projeto conta com corpora formados da variedade do português falada nas localidades de: Cametá, Mocajuba, Breves, Breu Branco e de Belém.

Ao todo o banco de dados do referido projeto contém amostras de fala de 342 informantes nativos do Pará e um total de mais de 100 horas de gravação.

Este banco de dados já subsidiou a investigação de três fenômenos relacionados diretamente ao vocalismo átono: o alteamento das vogais médiaticas pretônicas; a neutralização das vogais postônicas mediais e a nasalização aolfônica.

8. Referências


And I’d say “This week, we’re not going to clean the windows”: direct reported speech within a domestic labor workplace context

Kellie GONÇALVES
University of Bern, English Department
Länggassstrasse 49, 3000 Bern 9, Switzerland
goncalves@ens.unibe.ch

Abstract
This study investigates the meta-discursive accounts of successful and unsuccessful communication within a domestic labor workplace context of a multilingual cleaning company in New Jersey, USA. 41 semi-structured interviews were carried out with Portuguese-speaking domestics, language brokers and their Anglophone clients in order to understand how meaning is negotiated within this particular language contact situation. The analysis indicates that the main linguistic feature employed by participants was that of direct reported speech (DRS). Using DRS functioned to dramatize the effect of their speech events, represented the development of their accounts among interlocutors at the time of the actual conversation as well as claiming authenticity about their actual language practices within their daily interactions. The specific linguistic features investigated include personal, spatial and temporal deictic markers, marked changes in prosody, and speech verbs.

Keywords: reported speech; deictic markers; domestic labor workplace; discourse analysis.

1. Introduction
This study is about a specific language contact situation among Portuguese-speaking domestics and English-speaking clients in New Jersey, USA. It is part of a larger project on communication among domestics and their Anglophone clients, where meta-discursive strategies and the significance of dense, tightly-knit social networks (Milroy, 1980; Milroy & Milroy, 1992; Wei, 1993; Stoessel, 2002) are investigated as well as the linguistic landscapes of the neighborhood in which domestics reside. Preliminary results indicate that domestics’ use of English in the workplace consists of meta-linguistic strategies such as ‘basic’ English, gestures, as well as communicating through ‘language brokers’ (Tse, 1996; Weisskirch & Alva, 2002; Weisskirch, 2005; Del Torto, 2008). As a result of living in a Portuguese-speaking community, most of these women do not require English on a daily basis since most of their interactions can be carried out in Portuguese only. In meta-discursively reconstructing their interactions with one another, direct reported speech (DRS) (Volosinov, 1971; Bakhtin, 1981; Goffman, 1981; Coulmas, 1986; Li, 1986; Tannen, 1986, 1989; Clark & Gerrig, 1990; Butny, 1997; Biber et al., 1999; Holt, 1996, 2000, 2009; Myers, 1999; Carter & McCarthy, 2006; Sams, 2007, 2010) is employed, which functions to convey authenticity of the actual speech event (Coulmas, 1986; Li, 1986; Mayes, 1990; Holt, 1996, 2000, 2009), as well as representing the development of the conversation between parties and the interlocutors’ respective stances (Holt, 1996; Niemelä, 2005). Moreover, the use of DRS within this context functions to depict the story’s climax (Drew, 1998; Clift, 2000, Golato, 2000) and dramatize (Mayes, 1990; Myers, 1999) the effect of achieving both successful and unsuccessful communication within the reported interaction between domestics, clients and language brokers. The features of DRS that are scrutinized in this study include personal, spatial and temporal deictic markers, marked changes in prosody, and speech verbs (Holt, 1996). More specifically, the personal pronouns investigated include (I, you, she, we, they) while the spatial and temporal markers include those tense (present, continuous, past, etc.) and time adverbials (then, now), while the speech verbs consists of the reporting clause, namely a pronoun or name followed by a reported verb such as “said” or the quotative “like”. For Carter and McCarthy indexical markers or deictic words “are especially common in situations where joint actions are undertaken and where people and things referred to can be seen by the participants” (2006: 178). Deictic markers index the various ways individuals orient themselves and their interlocutors in interaction and function to make reference to physical, psychological and emotional closeness and distance as well as expressing contrast and difference (ibid.). A discourse analytic approach is employed within this study in order to reveal how the use of DRS within the context of spoken discourse functions and deems communication among Portuguese-speaking domestics and their Anglophone clients as successful or unsuccessful. The research questions driving this study are:

1) What linguistic strategies are used by participants to meta-discursively describe communication in their workplace?
2) What linguistic features are employed in their descriptions and what functions do they serve?

2. Data Collection
Obtaining data for a project among domestics and their employers can be extremely challenging and has been well documented by several researchers (Rollins, 1985; Anderson, 2000; Chang, 2000; Parreñas, 2001; Romero, 2002; Lan, 2006 and Parreñas, 2008). While Romero
(2002) worked as a domestic herself, Rollins (1985: 9) “worked for a month as a domestic to submerge [herself] in the situation prior to designing the research in order to sensitize [herself] to the experience of domestic work and of relating to a female employer”. I was fortunate that I had direct access to a cleaning company in New Jersey through familial ties and was able to conduct interviews with both employees and clients.

The data for this study consists of 41 semi-structured interviews, 18 with domestics, 19 with clients and 4 with language brokers. The interviews were recorded and lasted between 16 minutes – 1 hour and 30 minutes producing a total of 21.5 hours of recordings. Due to the data-driven nature of this study, hypotheses were not addressed in an *a priori* fashion. Rather, several thematic categories emerged from the transcripts and corpus, which are indicated in Table 1.0.

<table>
<thead>
<tr>
<th>Categories</th>
<th>Domestics</th>
<th>Clients</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Language use &amp; practices at work</em></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Language attitudes</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>English skills among domestics</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Social networks</td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Thematic categories

For the purposes of this study, I looked at language use and practices at work among domestics and clients. Below I scrutinize three excerpts, one from a Luso-Brazilian Portuguese-speaking domestic, one from an Anglophone language broker and the last one from an Anglophone client. In investigating how communication is achieved in the workplace context, I analyze how meaning is negotiated by interviewees’ evaluations and the DRS employed to reconstruct their conversations, which are deemed successful or not.

In extract 1 below, Livia, a Brazilian domestic, who has been residing in the U.S. for seven years discusses her difficulties of speaking English, but describes her ability to understand English at work when it is in written form. In order to exemplify what she means, Livia employs DRS to reconstruct a telephone conversation she had with Dona Magda, the company owner and language broker, concerning the content of a note left for Livia by an English-speaking client:

Extract 1 A domestic’s interpretation

1. L: mas olha eu não consigo soltar a língua (.)
2. não sei se é vergonha também (. ) sabe (.)
3. não sei
4. K: e com os clientes?=
5. L: =ah=?
6. K: =e com os clientes (.) por exemplo?
7. L: entendo que é XXX (.) igual quando elas
8. escreve alguma coisa eu sempre entendo (.)

9. eu sempre ligo pra dona magda e falo
10. “dona magda olha eu (1.0) tá assim assim
11. assim” “ah (.) mas é isso?” “tá ok” é o que
12. eu falei era aquilo mesmo (.) ela falou (.)
13. “não (.) tá tudo certo”

Livia begins this extract by explaining her challenges of speaking English when she employs the metaphor “soltar a língua” (line 1). She continues and states that she is not sure why, but confesses that it could be her embarrassment “vergonha também” (line 2) at actually speaking. When asked about her communication with clients, Livia states that she always understands when they write her notes “quando elas escreve alguma coisa eu sempre entendo” (line 8). Her use of the adverb of frequency always “sempre” is repeated in line 6 when she claims to always call her boss in order to confirm that she has understood the client’s note of instruction through written text. Livia reconstructs this conversation by using several features of DRS such as personal and temporal deixis markers, reported verbs as well as a shift in prosody. First, Livia uses the personal pronouns I “eu” and she “ela” to refer to herself and Dona Magda (lines 9 & 12) as the speakers of the conversation. Second, Livia employs the reported verb say in “falo” (line 9) to introduce her reported utterance and the pronoun-plus-speech-verb “ela falou” (line 12) to reintroduce Dona Magda into the conversation. This reintroduction of Dona Magda occurs in line 10 subsequent to the adjacency pair of a question and answer sequence that has been exchanged by Livia and her boss through the changes in prosody, represented in the extract by the underlined words, to mark both speakers (lines 10 & 11). Finally, Livia’s use of the verb tenses within this conversation are the present tense of the verb to be in “tá assim”, “é isso” and “tá ok” and are considered “appropriate to the reported speaker/context rather than the current one” (Holt 1996: 222). The exchange between Livia and Dona Magda presented in this extract is one that occurs on a regular basis in order to confirm Livia’s comprehension of the English instructions left for her by her English-speaking client. The DRS within this exchange indexes Livia as somebody who understands English well, but may be just embarrassed to speak it while simultaneously depicting Dona Magda as the language broker who provides encouragement and confirmation of Livia’s English comprehension skills “tá tudo certo” (line 7). As a result, this sequence depicts the communication between Livia, Dona Magda and the client as successful.

In the next extract, Janet, the English-speaking driver, who also functions as the main language broker when the company owner is unavailable, discusses and assesses Bella’s (a Portuguese domestic) English skills. Janet claims that because of Bella’s language insecurity, communication is stymied, which has previously led to prolonged and unnecessary problems:
Extract 2) A language broker’s view

1. Janet: bella’s problem is (.) is her insecur*ity
2. about her english and i tell her that (.) i
3. said (.) “bella (.) i understand everything
4. you:’re saying to me’ and you know like
5. over christmas (.) one of her insecurities (.)
6. i felt (1.0) if she wouldn’t have felt so
7. insecure (.) we could’ve resolved some
8. problems faster

In this extract, Janet reconstructs the conversation she had with Bella by using DRS, which functions to replicate the actual conversation as well as dramatize the hardships concerning their communication. This is done through Janet’s use of the speech verb “I said” (line 3) as well as the personal pronouns “I”, “you” and “me”. The personal pronouns “I” and “me” are co-referential with Janet who is doing the reporting. Similar to the co-referential functions of the pronouns used, are the temporal references of the present tense and present continuous tense of the verb forms in “I understand” and “you’re saying” (lines 3 & 4). The shift in prosody used within the reported utterance (underlined segment in lines 3 & 4) functions to dramatize the speech event and emphasize Janet’s comprehension and Bella’s intelligible English-speaking skills. The main problem of communication between Bella and Janet, however, lies in Bella’s apparent insecurity of speaking English (lines 1 & 5), which has led to delays of problem solving among domestics and clients. As a result, the utterance analyzed using DRS functions to dramatize communication between one particular domestic and language broker as often unsuccessful due to Bella’s linguistic insecurity.

In the final extract, Mrs. Malloy, an English-speaking client, discusses how she communicates with Patricia, her Portuguese-speaking domestic, by using both verbal communication as well as gestures. In exemplifying a typical situation, Mrs. Malloy uses DRS to offer evidence for the reported speech event as it actually happened:

Extract 3) A client’s perspective

1. M: i’d say erm (.) “patricia this week we’re
2. not going to clean the windows” and i’ll
3. point to the window and i’ll say (.) “i have
4. had them a:ll cleaned they’re fine (.) you
5. don’t need to touch them (.) so they’re a:ll
6. fine” like @@@ and we do hand signals
7. so and i say (.) “do you under- ok?” and
8. she’s like (.) “ok” and i don’t know if that
9. means “yes (.) i understand you” or “ok,
10. (.) you’ve said something” you know? i
11. (1.0) that (.) there is no like (.) there is no
12. real verbal communication back

In this extract Mrs. Malloy begins with the reported verb “say” and then continues her account of the conversation by addressing Patricia directly (line 1), which functions to convey that these were the actual words uttered during the initiation of the conversation. Second, she uses the inclusive personal pronoun “we”, the present continuous verb tense “going”, as well as the spatial deictic marker this week (line 1), all of which function to signal Mrs. Malloy’s point of view at that particular time. Her next DRS utterance (line 3) includes features such as temporal reference in the past perfect tense “I have had them all cleaned” as well as the present tense and personal pronoun “you don’t need to touch them” (lines 4 & 5), which function to indicate the time of speaking during the actual conversation with her interlocutor. Her claim of pointing to the window and their joint use of hand signals (line 6) suggest that Mrs. Malloy and Patricia use both linguistic and non-linguistic strategies in order for communication to be achieved which prove to work for both Mrs. Malloy and Patricia. In order to confirm Patricia’s understanding of Mrs. Malloy’s instructions, however, she inquires directly. This is seen in (line 7) when Mrs. Malloy uses the reported verb “I say”, which precedes the direct question “do you under- ok?” What is interesting about this question is Mrs. Malloy’s initial report about comprehension. She begins her utterance by asking if Patricia understands her instructions, but then resorts to simplifying her request by asking “ok?”, which is marked by a shift in prosody and rising intonation. In this context, Mrs. Malloy employs basic English skills in order for the communication between her and Patricia to be regarded as successful. Mrs. Malloy further states that Patricia confirms her request by her response when Mrs. Malloy makes use of the quotative in “she’s like “ok” (line 8). She then employs DRS to report a hypothetical account of her thought process and how the exchange developed (Sams 2007; 2010). This is done when Mrs. Malloy confesses to not knowing how she should socio-pragmatically understand Patricia’s use of “ok” by giving two possible options of its potential meaning. The first meaning could be a preferred response in positively responding back to Mrs. Malloy’s question while the second option “ok, you said something” (line 10), acknowledges Mrs. Malloy’s utterance. Despite the fact that Mrs. Malloy employs DRS to reconstruct this conversation and hypothetical thought process, which has the effect of dramatizing her account, she states that “no real communication” has taken place because the socio-pragmatic meaning of Patricia’s “ok” in response to Mrs. Malloy’s question remains ambiguous. Nevertheless, the reconstructed conversation reveals that the communicative event of giving directions between Mrs. Malloy and Patricia using gestures and basic English is ultimately deemed successful.

3. Conclusion

According to Coulmas (1986: 2) the use of DRS “evokes the original speech situation and conveys, or claims to convey, the exact words of the original speaker” in the interaction. The effect of employing DRS within storytelling or narratives also functions to dramatize the
unfolding events of interlocutors’ interactions at the time and place of the actual speech event. In my analysis, I showed how the use of DRS among domestics, language brokers and clients was employed as a prominent linguistic strategy, which functioned to convey authenticity of the actual speech event between domestic and language broker or domestic and client. This was shown in all three extracts analyzed above. The second function DRS had within the analysis was to represent the development of the conversation between interlocutors’ as well as their particular stances concerning their joint communication of the speech event. The final function that DRS had within this study was to depict the story’s climax and dramatize the effect of achieving both successful or unsuccessful communication within a specific language contact situation within a domestic labor and workplace context. In presenting the analysis, I focused on typical DRS features, which included personal pronouns, spatial and temporal markers, shifts in prosody as well as speech verbs. In her work on workplace discourse, Holmes states that “few researchers have ventured into blue collar worksites; they tend to be noisy and dirty and often rather uncomfortable places for academics undertaking research” but asserts that “this is undoubtedly another direction in which it is important to expand workplace discourse research (forthcoming: 15). The aim of this study was to “venture” into an area of research that is not always easily accessible to researchers and as a result, a dearth of linguistic studies exists within the context of domestic labor. The intention of my study was to expand the direction of workplace studies in general and thus shed light on how meaning is negotiated between Portuguese-speaking domestics and their Anglophone clients. Research on workplace studies outside of white-collar contexts is indeed challenging yet, I hope to have shown that communicative strategies within a domestic labor context yields fruitful insight into how meaning is achieved and reported on between interlocutors of different language backgrounds.

4. References


University Press.


## 5. Appendix

Transcription Conventions:

@@ = signals laughter

wo:rd = perceptible lengthening

(,) = pause shorter than one second

(1.0) = pause lengths in seconds

? = rising intonation, often signals questions

= = latched talk

___ = underlined text is marked for changes in prosody
Mapping Paulistano Portuguese: the SP2010 project

Ronald Beline MENDES, Livia OUSHIRO
University of Sao Paulo
Av. Prof. Luciano Gualberto, 403 - Sala 16, Cidade Universitária, 05508-010 - São Paulo - SP
rbeline@usp.br, livia.oushiro@usp.br

Abstract
This paper reports on the objectives, methods, and results from the project SP-2010 (Mendes, 2011), currently under the execution by the Grupo de Estudos e Pesquisa em Sociolinguística (GESOL-USP). Its main objectives are (i) to build a contemporary and representative sample of São Paulo Portuguese; (ii) to develop studies of sociolinguistic variation in the city, an understudied speech community (Mendes, 2009; Rodrigues, 2009); and (iii) to make the corpus of recordings and transcripts available online for a wider group of researchers. The first phase of the project aims at collecting 60 sociolinguistic interviews with speakers stratified by sex/gender, age, and level of education by 2013. In view of the highly heterogeneous sociodemographic make-up of the city of São Paulo, fieldworkers also observe distinctions in informants’ social class, family generation in the city, and area of residence. Interview recordings follow Variationist Sociolinguistics premises (Labov, 1984, 2006; Tagliamonte, 2006) and data transcription norms are designed as to facilitate automatic data handling in softwares such as R.

Keywords: spoken corpus; Paulistano Portuguese; variationist sociolinguistics; data collection; transcripts.

1. Introduction
Although São Paulo Portuguese has already been documented and analyzed through broad and significant research projects such as Projeto NURC-SP (Castilho & Preti, 1986, 1987; Preti & Urbano, 1988, 1990) and Projeto Para a História do Português Paulista (Castilho 2007), most works within these projects aim at analyzing “Brazilian Portuguese,” either in contrast with European Portuguese (e.g., studies on parametric variation), or in relation to its internal processes of change (e.g., studies on grammaticalization).

Among the very few works about Paulistano Portuguese in its social context, Rodrigues (1987) analyzed variable subject-verb agreement (e.g., nós vamos vs. nós vai ‘we go’) in the speech of 40 (semi-)illiterate speakers in two favelas, and Coelho (2006) analyzed the variable use of 1PP pronouns (nós vs. a gente ‘we’) in the speech of 24 speakers living in a working class community. Yet, to date, little is known about the linguistic production and perception of many other (supposedly) typical Paulistano variants (e.g., the realization of coda /r/ as a tap in words such as porta ‘door,’ the diphthongization of nasal /e/ in words such as fazenda ‘farm’) and other variants in the city, as well as their social distribution and evaluation in the speech community at large.

This may be due to the difficulties of building a representative speech corpus of a heterogeneous and multicultural city with more than 11 million people, highly diverse in terms of their geographical origin, socioeconomic class, and cultural background. According to a recent survey by the Instituto de Pesquisa Econômica Aplicada (IPEA, 2011), 46% of the adult working population (between 30 and 60 years old) living the the São Paulo Metropolitan Area were not born in the state of São Paulo (see Figure 1). Although the survey does not refer exclusively to the city itself, it gives an idea of the intense presence of non-native inhabitants in this region. One can consider that the number of non-Paulistanos living in the city may be even greater, since the 54% of Paulistas include all people born in the state of São Paulo and not only the capital city.

Figure 1: Adult population living in the São Paulo Metropolitan Area. Source: IPEA 2011

This fact raises a number of questions: which social parameters are most relevant for linguistic differentiation and stratification and how to reach speakers of varied social networks? How to gather detailed ethnographic information from each informant (Poplack, 1989), acknowledging a persistent point made by the “third-wave” of sociolinguistic studies (Eckert, 2005) on the importance of observing individuals’ social practices? Which methodologies are best for handling a large amount of spoken linguistic data?

In this paper, we report on the objectives, methods, and results from the Project SP-2010 (Mendes, 2011), currently under execution by the Grupo de Estudos e Pesquisa em Sociolinguística da USP (GESOL-USP), which aims at: (i) building a contemporary and representative sample of Paulistano Portuguese; (ii) fostering the development of sociolinguistic studies in the city; and (iii) making the corpus of recordings and transcripts available online for a wider group of researchers.

1 http://linguistica.fflch.usp.br/gesol.
2. Methods and Results

In 2009-2010, GESOL-USP collected 82 sociolinguistic interviews with residents of the city of São Paulo, native or not to the city, of both sexes and different sexual orientations, from 15 to 89 years of age, with different levels of education, of varied socioeconomic statuses, living in 59 different neighborhoods in the city. In view of São Paulo's great sociodemographic complexity, these exploratory recordings had the objective of defining the most relevant social variables for the sociolinguistic description of Paulistano Portuguese; elaborating an interview schedule; developing best practices in approaching possible informants; identifying possible technical and methodological problems that may occur during the recordings (e.g. avoiding noise, making the informant comfortable) and coming up with solutions for them; and elaborating criteria for transcribing the interviews.

From this experience, we observed that certain sociolinguistic profiles are hard to locate – for instance, younger native Paulistanos who have not concluded at least high school, especially women living in more central areas, or people over 70 who were actually born in the city, especially in more suburban areas. In addition, in spite of our initial aim of locating prototypical speakers from certain neighborhoods (e.g. Mooca, Bexiga, Pinheiros), geographic and socioeconomic mobility seems to be characteristic of the city and its inhabitants, many of whom prefer not to settle in a single place for life. Further, a technical but not to be ignored challenge is the presence of noise (traffic, constructions, people), even in residential areas of the city. The methods designed for this project try to address some of these issues.

In the present phase, to be concluded by 2013, the social parameters for constituting the sample are sex/gender (men and women), three age groups (20-34 y.o.; 35-59 y.o.; 60+ y.o.), and two levels of education (up to high school; college). As our focus is on the social meaning of variation (Chambers, 1995), these variables have been chosen primarily because of their potential to shed light on the relationship between variable linguistic uses and social identities, as well as to enable cross-comparisons with other linguistic corpora of Brazilian Portuguese – e.g. VARSUL (Bisol et al., s/d), VALPB (Hora, 2004), PEUL (Paiva & Scherre, 1999), ALIP (Gonçalves, 2003).

Sex/Gender and Age have been broadly analyzed in sociolinguistic studies and have been shown to be correlated with variables whose variants are differently evaluated in terms of prestige: a number of works have observed that the prestigious forms in the community tend to be employed by women (Chambers, 1995; Labov, 2001; Cheshire, 2004), and that unprestigious forms tend to be avoided by speakers in the intermediary age group, who mostly suffer pressures of the linguistic market (Bourdieu, 1991; Labov, 2001). Correlations with Age can also point to possible changes in progress in the linguistic system through apparent time analyses (Labov, 2001). The three age groups are mostly based on their relative position in the job market, but also take into account each group's general lifestyles in a big city. The younger speakers, those between 20 and 34 years old, comprise young adults who tend to be relatively less stable than people in the other two age groups; in São Paulo, it is not rare to find people up to 34 years old who are not married, who do not own their own place, who go to college or who lead life more similarly to people in their early 20s. The group aged between 35 and 59 years old, in turn, is intended to comprise people more fully inserted in the job market and relatively more stable. Finally, the group over 60 years old refers to people in or close to retirement.

Level of education is also directly associated with stigmatization and prestige. The general hypothesis is that more educated speakers will tend to avoid unprestigious forms in the community, or otherwise that the forms they employ will be considered more "correct." In Brazilian sociolinguistic studies, the division between "educated" and "uneducated" speakers is normally taken as an index of socioeconomic status (Rodrigues, 2009: 151). This situation seems to be changing in São Paulo as well as in many other urban centers through extensive public policies of improved access to primary, secondary, and higher education (for instance, Progressão Continuada in the state of São Paulo and ProUni in a national scope); the division between only two levels of education is a consequence of these changes. However, general increase in average levels of education is not always followed by a direct ascension in individual socioeconomic status, which means that the equation between level of education and social class should not be overestimated. We suggest that level of education should be treated as constitutive of speakers' social class, but not as its substitute.

The combination of these social parameters yields 12 sociolinguistic profiles (e.g. men between 20-34 y.o. without a college degree), each of which is to be filled by 5 speakers, in a total of 60 sociolinguistic interviews. Each of these 5 speakers per cell should reside in a different zone of the city (North, South, East, West, Central), and each cell should contain at least one speaker of three city areas (Downtown, Extended Central Area, Suburbs), as a way to ensure a broad coverage of the city. The speakers' place of residence is defined as the place where he/she has lived for the most part in the past 10 years.

In a second stage, we will focus on social class, a social factor generally overlooked in Brazilian sociolinguistic studies due to lack of reliable criteria for categorizing speakers in different socioeconomic groups (Rodrigues, 2009; Mendes, 2011). In the city of São Paulo, speakers' socioeconomic status possibly should take into account, in addition to their income and level of education, their type of residence, occupation, and access to cultural goods. The corpus will also be stratified according to speakers' generation in the city, in order to examine the contribution of different groups of migrants and immigrants in the community, and speakers' area of residence, which is also an index of socioeconomic status.
During this first phase of the project, information on these variables is collected through the sociolinguistic interview and post-recording questionnaires, which will enable preliminary analyses of their role in the sociolinguistic stratification in São Paulo.

Speakers to be recorded have been contacted through the “friend of a friend” method (Milroy, 2004). Our experience has shown that speakers in the city are very resistant to talking to a “stranger” (the researcher); however, when introduced by a common acquaintance, speakers tend to be much more receptive and solicitous, a fact that also has consequences for naturalness of speech. After a speaker has been recorded, the researcher asks her/him to suggest another speaker. As a means to ensure that informants do not belong to the same or few social networks, the new suggested speaker can only be recorded if he or she is not acquainted with the person who indicated the current informant. For instance, in the example in Figure 2, B has indicated two new speakers, C and D, but only the latter can be selected as a new informant.

![Figure 2: Selection of informants](image)

The interview schedule has the twofold objective of obtaining samples of spontaneous speech by Paulistanos of varied sociolinguistic profiles and more information on these speakers’ living conditions, sociolinguistic evaluations and perceptions (Labov, 2006). It is divided into two parts. The first one is more personal and covers topics such as the speakers’ neighborhood, childhood, parents and family, education, current occupation, social network, and leisure activities. It aims at obtaining narratives in the past (e.g. “What was your childhood like in neighborhood X?”), in the present (e.g. “In your leisure time, what do you and your family like to do?”) and in the future (e.g. “What would you do if you won the lottery?”), as well as opinion accounts (e.g. “What do you think of the new law for gay marriage?”). The second part contains more specific questions about the speakers’ relation to the city and their perceptions on Paulistano identities (e.g. “When you were in (another city), did people recognize you as a Paulistano? If so, how?”). In the last part of the interview, speakers are asked to read a word list, a news report, and a ‘statement’ (a text with strong marks of oral language). Although the interview schedule is divided into two parts, it enables easy transition between topics and has yielded natural sounding conversations.

After the interview is recorded, the fieldworker fills out a form with detailed speaker’s sociolinguistic information (date of birth, occupation, family’s place origin and first generation that migrated to São Paulo, schools, place(s) of residence etc.), and makes note of any relevant additional information in the fieldwork journal. The informant is also asked to fill out a socioeconomic form, if he/she feels comfortable to do so, containing seven multiple-choice questions about their monthly income and living conditions. Our experience has shown that the multiple-choice form greatly improves the chance of obtaining these data (instead of having the informant orally answer these questions directly to the fieldworker).

Each sociolinguistic interview is about 60-70 minute long and has been stored in .wav (stereo, 44,100 Hz) format. The recordings have been made with TASCAM DR-100 recorders and two Sennheiser HMD26 microphones (one for the fieldworker and one for the informant). Although it could be argued that the presence of these technical paraphernalia possibly enhances the Observer’s Paradox (Labov, 2006), we find that speakers’ occasional uneasiness tends to decrease considerably after some 15 minutes of recording and, more importantly, that the improved audio quality is worth the trouble, especially in a city as noisy as São Paulo.

All interviews are then evaluated by four members of the research group not involved in the field recordings, according to the speakers’ fitness to the sociolinguistic profile, audio quality, naturalness of conversation, and conformity to the interview schedule. The 82 previously collected interviews during the pilot experience have also been evaluated according to the same parameters, and some of them may be included in the final corpus to be made available online, in addition to the 60 recordings of the present data collection phase, as long as they meet the high-quality requirements.

The criteria for transcribing the recordings follow a simplified semioptographic approach in order to make the material more easily available in a written medium. The following criteria aim at facilitating the manipulation of text files in softwares such as R (Gries, 2009; Hornik, 2011) to automatically identify and extract tokens of a variable into a spreadsheet program (Oushiro, 2012).

Transcripts do not contain any special formatting such as boldface, italics, tab stops, columns, and are saved in plain text (.txt) with UTF-8 encoding. Orthographic rules of Brazilian Portuguese are followed in every case, even if speakers produce variants that differ from the written standard. The idea here is that a transcriber is unable to pay attention to all variable phenomena simultaneously – e.g. monophthongization of /ow, /ej/, diphthongization of nasal /e/, postvocalic /t/ deletion, nasal assimilation of /ndo/, vowel raising of unstressed /e, o/, to name a few. In addition to creating unintelligible texts, this would probably cause transcripts to be unstandardized; further, the fact that the recordings will be made available lessens the need for a highly detailed transcript. On the other hand, grammatical variables should not be “corrected” by the transcriber (e.g. lack of
nominal agreement). Punctuation is limited to ellipses (to signal pauses), and question and exclamation marks (to indicate intonation of certain phrases). Capital letters are only employed in proper names (e.g. cities, institutions), abbreviations (e.g. USP, and identifying speakers (e.g. S1, D1).

GESOL-USP has also been developing parallel data collection projects, in addition to gathering a sample from the community at large. These parallel projects and studies are centered on specific groups of speakers and/or social variables within the city: residents of the upper class neighborhood Itaim Bibi (Ciancio, 2012); social class (Faria, 2012); gay men and gender (Soriano, 2012); different groups of migrants – Paraibanos (Mendes, forth) and Alagoanos (Silva, 2012). These studies aim at describing and contrasting general sociolinguistic patterns of the community and their uses within certain social groups residing in the city.

Based on the corpus collected so far, the research group has been developing studies of sociolinguistic variation in Paulistano Portuguese: the variable realization of coda /r/ as a tap or a retroflex, in words such as porta ‘door’ and mulher ‘woman’ (Mendes, 2009, 2010; Mendes & Oushiro forth); variable nasal (e) as a monophthong or a diphthong, in words such as fazenda ‘farm’ (Mendes, 2010; Oushiro, 2011); verbal negative structures (e.g. Não vou vs. Não vou não ‘I won’t go’) (Rocha, 2012); nominal and verbal agreement (Silva, 2012; Oushiro, 2011).

3. Conclusion

The SP-2010 Project has been collecting a contemporary corpus of Paulistano Portuguese and fostering the development of sociolinguistic studies focusing on the correlations between variable linguistic uses and social identities. By 2013, more than 60 sociolinguistic interviews (audio and transcriptions) will be made available online to the linguistic community. Parallel to this data collection project, a number of studies have also been analyzing specific social networks and communities of practice in the city, in contrast with larger community variational patterns, as to provide a broader and more detailed description of linguistic uses in São Paulo.

4. Acknowledgements

This research is funded by FAPESP (Grant 2011/09278-6).

5. References


Documentação da Língua Indígena Brasileira Yaathe (Fulni-ô)

Januacele da COSTA, Miguel OLIVEIRA Jr., Fábia SILVA
Universidade Federal de Alagoas
Maceió – AL, Brasil
januafe@bol.com.br, miguel@fale.ufal.br, fabia_fulnio@hotmail.com

Resumo
Este artigo tem por objetivo descrever o Projeto de Documentação Linguística da Língua Indígena Brasileira Yaathe, falada pelo povo Fulni-ô. O povo Fulni-ô, que vive no município de Águas Belas, interior de Pernambuco, é o único povo indígena do nordeste brasileiro que preservou sua língua depois do processo colonizador. A despeito do uso sistemático que os Fulni-ô fazem de sua língua, sobretudo em situações privadas, ela tem sido considerada por órgãos internacionais uma língua em extremo risco de extinção. Justifica-se, desta maneira, a urgência de um projeto de documentação como o que se descreve a seguir. O artigo apresenta um breve histórico do povo Fulni-ô, situando-o socio-historicamente, descreve a situação atual de sua língua, lista os objetivos do projeto a ser desenvolvido, justificando a sua relevância, e detalha metodologia específica a ser adotada na coleta e no tratamento de dados, metodologia que segue padrões hoje adotado por bancos de dados de línguas em perigo de extinção.

Palavras-chave: Yaathe; Fulni-ô; documentação linguística.

1. Introdução

Os índios Fulni-ô vivem no município de Águas Belas, no oeste-sudoeste de Pernambuco, a cerca de 300 quilômetros de Recife, a capital do Estado de Pernambuco. A reserva indígena Fulni-ô está localizada a pouca distância da margem esquerda do Rio Ipanema, um dos afluentes, também da margem esquerda, do Rio São Francisco.

Um dos aspectos mais interessantes da situação dos índios Fulni-ô é a sobrevivência da língua, uma vez que todas as outras línguas indígenas faladas nessa parte do país já desapareceram. Embora se possa afirmar a vitalidade da língua neste momento, divergências internas e outros problemas, como o empobrecimento cada vez mais crescente da região e o descaso das autoridades regionais, poderiam vir a mudar esse quadro em poucos anos. As pessoas mais jovens da comunidade foram encarajadas, por um período de cerca de 40 anos, a não falar sua língua ou viver de acordo com os costumes de seu povo. Esse direcionamento e as atitudes dele decorrentes, que preservou sua língua depois do processo colonizador. A despeito do uso sistemático que os Fulni-ô fazem de sua língua, sobretudo em situações privadas, ela tem sido considerada por órgãos internacionais uma língua em extremo risco de extinção. Justifica-se, desta maneira, a urgência de um projeto de documentação como o que se descreve a seguir. O artigo apresenta um breve histórico do povo Fulni-ô, situando-o socio-historicamente, descreve a situação atual de sua língua, lista os objetivos do projeto a ser desenvolvido, justificando a sua relevância, e detalha metodologia específica a ser adotada na coleta e no tratamento de dados, metodologia que segue padrões hoje adotado por bancos de dados de línguas em perigo de extinção.

2. Justificativa
Recentemente, a UNESCO divulgou um relatório sobre línguas em risco de extinção e, de acordo com os critérios
utilizados pela pesquisa, o Yaathe é uma língua que se encontra em “extremo perigo de extinção”.\(^1\)

Apesar de os números indicarem uma alta porcentagem de falantes de Yaathe entre os Fulni-ô (cerca de 3.000 pessoas, o que corresponde a mais de 90% da população total), o uso da língua está restrito a situações bastante específicas. Raramente os Fulni-ô fazem uso de sua língua nativa em situações públicas; há, no entanto, evidências de que quase todos eles a utilizam em situações privadas. Nas famílias, por exemplo, os pais, em geral, dão ordens ou fazem perguntas aos filhos em Yaathe, a despeito de estes invariavelmente responderem em português. Estudos recentes indicam que crianças muito pequenas dominam aspectos particulares do uso da língua, como, por exemplo, a caracterização de gênero.

A despeito do uso sistemático que os Fulni-ô fazem de sua língua em situações privadas, e do esforço que o povo tem demonstrado em manter vivas a sua língua e a sua cultura, através de iniciativas educacionais, há ainda muito pouco registro do Yaathe, o que dificulta bastante quaisquer atividades relacionadas à preservação de suas manifestações linguísticas e culturais.

Atualmente, o material usado nas escolas como recurso de ensino-aprendizagem da língua na reserva indígena Fulni-ô é bastante escasso e de qualidade questionável.\(^2\) Os professores fazem o que podem: escrevem seus próprios textos, preparam aulas e planos de aula, conforme exigido pelas instâncias oficiais, falam sobre cultura e religião, incentivam o uso da língua e o respeito pela cultura como um todo, tudo feito de maneira muito pouco sistemática e sem amparo em usos reais, documentados, da língua. Além de uma cartilha elaborada nos anos 90 do século passado, não há outro material oficial para o ensino da língua.\(^3\) Há, por outro lado, muito material criado e produzido pelos professores, e um esforço cada vez mais constante no sentido de se padronizar a escrita de modo a ser aceita pela comunidade.\(^4\) Parece evidente que o acesso a um banco de dados da língua será de vital importância para a elaboração de materiais didáticos mais adequados, bem como para auxiliar no processo de sistematização da grafia da língua.


Mais recentemente, três trabalhos foram efetuados sobre a língua. Costa (1993) procurou investigar a atual situação linguística dos Fulni-ô, dada a sua peculiaridade de última língua nativa no Nordeste do Brasil, a fim de verificar tendências à substituição ou ao deslocamento em relação ao Português. Esta investigação serviu como pano de fundo para a observação de fenômenos de atitudes linguísticas de professores não-índios, face à variedade de Português falada pelas crianças índias que chegam à escola da cidade, e de interferências de uma língua na outra, mais precisamente da influência do Yaathe – que consideramos língua materna – sobre o Português – segunda língua. Neste caso, tratava-se da variedade de Português falada pelas crianças índias. Os resultados de tal trabalho podem, por um lado, ajudar a clarear e a melhorar a compreensão dos professores de língua portuguesa das variedades linguísticas que são utilizadas pelos alunos de procedências diversas. Por outro lado, devem contribuir para o conhecimento e o autoconhecimento das nações indígenas. Costa (1999) detém-se sobre a estrutura do Yaathe, procurando descrever e explicar o sistema (fonologia e gramática) e o seu funcionamento. Cabral (2009) enfocou o sistema prosódico da língua, buscando descrever o acento no nível da palavra, experimentalmente.

Atualmente, há estudos em andamento dentro do projeto Gramática descritiva (de usos) do Yaathe (Fulni-ô), desenvolvido no PPGLL/UFAL, sendo duas monografias de iniciação científica (uma sobre gênero e outra sobre nasalidade em Yaathe) e uma dissertação de mestrado (sobre a estrutura da sílaba em Yaathe). A disponibilização de um banco de dados etiquetado, transcrito e devidamente anotado em muito auxiliará a boa execução destes e de futuros estudos acerca da língua.

### 3. Objetivos

Em vista do trabalho que vem sendo efetuado há algum tempo na aldeia e com a língua, já se dispõe de uma quantidade razoável de material coletado – listas de palavras, textos variados (letras de músicas, narrativas, cânticos religiosos) e respostas a questionários variados. Parte desse material foi gravado em formato digital. Entretanto, é preciso que se faça um tratamento mais consistente em termos de digitalização e organização para arquivamento e disponibilização pública, de modo a que esse material possa efetivamente vir a constituir um banco de dados da língua.

O objetivo central deste projeto é compor um banco de dados do mais abrangente possível acerca da língua Yaathe, constituído de materiais já coletados e de materiais por coletar. O banco de dados seguirá os modelos hoje adotados por bancos de dados de línguas em perigo de extinção\(^5\), contendo materiais transcritos, anotados e acessíveis à comunidade. Os dados já coletados serão organizados, etiquetados, transcritos e anotados.

Também o projeto tem por objetivo coletar materiais complementares para constituir o banco de dados. Assim, e de acordo com as necessidades estabelecidas a partir da sistematização dos dados já existentes, objetiva-se fazer

---

2. A escola da aldeia oferece educação básica, do maternal ao ensino médio, incluindo educação de adultos, recebendo, aproximadamente, 1.000 alunos em condições precárias.
3. Neste ano de 2010, a língua foi incluída na matriz curricular da escola regular da aldeia, sendo assim uma das poucas línguas indígenas brasileiras a ser oficialmente incluída no ensino regular, reconhecida pelo MEC e pela Secretaria de Educação do Estado de Pernambuco.
4. Cabe observar que a equipe que se propõe a desenvolver este projeto participa deste movimento, apoiando-o, fornecendo assessoria linguística e propondo descrições mais minuciosas da língua, que contribuirá para a elaboração de materiais didáticos mais adequados.
5. Utilizaremos, para este fim, as recomendações feitas pela E-MELD School of Best Practice ([http://www.emeld.org/school/](http://www.emeld.org/school/)).
coleta de dados acústicos de alta qualidade, contendo não apenas material proveniente de listas (como as clássicas Swadesh, Linguas Descriptive Questionaire, e aquelas propostas por Healey, em seu Manual de trabalho de campo), mas, sobretudo, exemplos discursivos, entre os quais narrativas de experiência pessoal, mitos, narrativas procedimentais e conversas espontâneas. Muito desse material também será gravado em vídeo, uma vez que informações visuais têm sabidamente importância fundamental para a compreensão de determinados fenômenos linguísticos.

Esse banco de dados é, como já se apontou, o produto principal deste projeto. Entretanto, espera-se que a constituição do banco de dados sirva como ponto de partida para novas pesquisas acerca da língua, para a implementação de estudos já em andamento, para o aprofundamento das discussões acerca de um sistema gráfico aprovado pela comunidade e para a elaboração de materiais didáticos para o ensino da língua. O projeto que aqui se propõe tem por objetivo envolver e formar pesquisadores em diferentes níveis – da IC ao doutorado – e professores pesquisadores, na tarefa de descrição e estudo dos diferentes aspectos da estrutura do Yaathe.

4. Metodologia

O material já existente será selecionado, levando-se em conta a qualidade da gravação e a potencial utilidade do mesmo. Os exemplares escolhidos serão tratados (digitalizados e editados, em alguns casos), etiquetados e organizados dentro de uma estrutura computacional hierárquica a ser definida.

Uma vez que se tenha uma ideia do material aproveitável dentro do corpus não-catalogado já existente, uma coleta de dados em campo será organizada, sendo portanto complementar o material já disponível para compor o banco de dados.

Entretanto, os dados que se planeja coletar incluem-se listas de palavras e frases, tendo como modelo as já clássicas listas Swadesh (Swadesh, 1955), LDQ (Comrie & Smith, 1977), e aquelas propostas por Healey (1975), em seu Manual de trabalho de campo, e uma série de exemplares discursivos, entre os quais narrativas de experiência pessoal, mitos, narrativas procedimentais e conversas espontâneas. Um dos objetivos principais desta coleta de dados é incluir dados de vídeo, uma vez que informações visuais têm reconhecida importância para a compreensão de determinados fenômenos linguísticos. Portanto, objetiva-se gravar também em vídeo a maior parte das sessões de coleta de dados em campo.

Os dados de áudio e vídeo serão gravados e arquivados respeitando todas as medidas e indicações propostas pela E-MELD School of Best Practice⁶, que vem sendo adotadas em projetos de documentação de línguas indígenas internacionalmente, pelo Open Archival Information System (OAIS)⁷, que é um modelo de referência, com padrão ISO (14721:2003), adotado pelos bancos de dados linguísticos mais recentes, e anotados seguindo os preceitos do Metadata Encoding and Transmission Standard (METS)⁸, também adotados por bancos de dados internacionais.

Após essa fase de organização e coleta de dados, proceder-se-á à etapa seguinte: a transcrição, tradução e anotação dos dados. Essa é uma fase que demandar um tempo considerável de trabalho, pelo que estima-se que apenas um percentual do material será transcrito e anotado. Por conta disso, uma cuidadosa seleção será feita do material a ser transcrito e anotado, levando-se em conta a representatividade e potencial utilidade do mesmo.

A transcrição e tradução serão feitas com o auxílio dos professores de Yaathe, o que resultará em um produto acurado e proporcionará uma discussão acerca de um modelo adequado de grafia a ser adotado, com aprovação da comunidade. As transcrições serão feitas no programa Praat (Boersma & Weenik, 2007), uma vez que este programa dá acesso a detalhes acústicos dos dados, o que não apenas facilita a transcrição, nos mais diferentes níveis, mas também auxilia a leitura de estudos acústicos os mais diversos. É importante ressaltar que um dos objetivos deste projeto é elaborar um banco de dados que seja disponibilizado para a comunidade acadêmica, com o objetivo de propiciar estudos linguísticos os mais diversos.

Portanto, é preciso levar em consideração o arcabouço tecnológico utilizado na construção do banco de dados. Os aplicativos computacionais que se pretende utilizar na execução do presente projeto têm sido sistematicamente utilizados por diversos projetos internacionais de documentação de línguas, por possuírem uma estrutura de fonte aberta, por funcionarem em diversas plataformas operacionais e por estarem em constante desenvolvimento.

Os dados transcritos em Praat serão exportados para o programa ELAN (Hellwig & Uytvenck, 2007), que permite uma maior liberdade de uso para anotação, possibilitando inclusive o alinhamento da transcrição e anotação com arquivos de vídeo. As tecnologias tanto do Praat quanto do ELAN possibilitam que os dados transcritos sejam disponibilizados online para consulta, através do programa open source Spock⁹, que permite efetuar buscas no corpus transcreto devolvendo transcrição e som correspondente.

Além de disponibilizar os dados localmente, nos servidores da Universidade Federal de Alagoas, para livre consulta pela comunidade, os dados serão depositados em bancos internacionais, tais como o do LAT (Language Archiving Technology)¹⁰, garantindo assim a sua preservação.

5. Considerações Finais

Entende-se, de acordo com Himmelmann (2006), que documentação de línguas é um campo de investigação e de prática linguística cujas preocupações básicas são a

---

⁶ E-MELD School of Best Practice (http://www.emeld.org/school/).
⁹ Cumpre notar que o projeto conta com a participação de uma falante nativa do Yaathe, Fábia Pereira da Silva.
¹¹ http://corpus1.mpi.nl.
compilação e a preservação de dados linguísticos primários e interfaces entre esses dados e vários tipos de análises neles baseadas. Além disso, embora preocupação com línguas em risco de extinção seja uma boa razão para que se desenvolvam projetos de documentação de línguas, não é a única. Documentações de línguas fornecem subsídios para as bases empíricas da linguística e de disciplinas afins, tais como tipologia linguística, antropologia cognitiva, etc., que dependem muito de dados de comunidades de fala pouco conhecidas para verificação das suas hipóteses, economizando, assim, recursos de pesquisas.

A principal contribuição do presente projeto de pesquisa é, assim, auxiliar a preservação de uma língua nativa brasileira em estado de iminente extinção, oferecendo uma documentação linguística abrangente e representativa, que poderá ser utilizada não apenas para estudos acadêmicos, mas também para a elaboração de materiais didáticos utilizados no ensino da língua na comunidade indígena.

É importante salientar que o esforço para a preservação de línguas em estado de extinção tem sido considerável, por meio, sobretudo, de agências de fomento internacionais (como a UNESCO e a VolkswagenStiftung, por exemplo). O Yaíthê não está incluído em nenhum desses programas, o que torna o financiamento deste projeto ainda mais urgente e relevante. Como apontado acima, o Yaíthê é a única língua indígena brasileira ainda sobrevivente no Nordeste do Brasil, o que torna qualquer esforço no sentido de sua preservação extremamente importante, no sentido de valorizar e preservar a identidade da cultura nativa dessa região do país.

6. References

Tupinambá Nheenga: considerações sobre um dicionário escolar do Tupinambá de Olivença, BA

Clara Carolina SANTOS, Consuelo COSTA
Universidade Estadual do Sudoeste da Bahia
Estrada do Bem Querer, km 3, Vitória da Conquista, Bahia
claracarlolina@gmail.com

Abstract
A intenção é elaborar um vocabulário bilíngue que compreenda um acervo lexical representativo da língua Tupinambá com informações fonéticas correspondentes a cada entrada. Este vocabulário deverá ser de utilidade nas atividades escolares voltadas para o ensino e fortalecimento da língua Tupinambá e pode constituir-se como uma importante referência da língua e de aspectos da cultura Tupinambá. Os resultados deste estudo deverão servir como material de apoio à escola e nuclear das Tupinambá, mas também para o ensino do português, pois atualmente os Tupinambá buscam uma aprendizagem escolar nas duas línguas. O vocabulário escolar bilíngue Tupinambá – Português terá a inovação, em relação aos dicionários escolares em línguas indígenas em Tupinambá, de apresentar a transcrição fonética dos verbetes o que - em conjunto com as oficinas de fonética e fonologia oferecidas aos professores indígenas - proporcionará um suporte material que auxiliará de modo seguro o uso da língua na escola e sua retomada pela comunidade. Além disso, este vocabulário diferenciar-se-á dos demais dicionários do Tupinambá (língua da qual o Tupinambá é uma variedade) por considerar a convenção ortográfica dos índios de Olivença.

Keywords: Tupinambá; línguas indígenas; fonologia.

1. Paper
Quando é impressa em 1595 uma Gramática de José Anchieta1 para uso na Companhia de Jesus à variedade de língua ali descrita não é atribuído nenhum nome (Rodrigues, 2010: 28). É apenas no decurso da empresa lusitana que a língua mais usada na costa do Brasil é denominada língua brasileira ou língua do Brasil2. Nos primeiros livros sobre o Brasil, língua da costa, língua brasileira ou apenas língua é alusão à língua nativa das nações habitantes da quase totalidade da costa brasileira (Rodrigues, 1994), foi uma variedade empregada na missão jesuítica nos séculos XVI e XVII (Câmara Jr., 1979: 99) e, a partir do século XIX, é considerada uma língua das origens do Brasil (Dietrich, 2010: 10). Em estudos mais recentes, a delimitação da língua da costa é referida como “uma realidade linguística complexa (Dietrich, 2010: 9)”.


No caso específico do Tupinambá, julga-se que esta variedade tenha se espalhado “por causa das migrações contínuas dos Tupinambás (Dietrich, 2010: 12)” entre Santa Catarina, Bahia, Maranhão e Amazônia. Neste texto, faremos referência ao uso da variedade Tupinambá.
entre os indígenas em Olivença, Ba. Para este estudo, as línguas da família tupi-guarani formam “um grupo com outras línguas mais distantes na sua diferenciação histórica, mas que, elas também, apresentam correspondências regulares de sons, de palavras e de formas gramaticais (Dietrich, 2010: 10)”.

De um modo geral, escolhemos denominar a língua por Tupinambá, pois este é o uso corrente entre os indígenas em Olivença, embora saibamos que, em seu estudo na escola e uso primeiro, a língua alvo é o Tupi Antigo.

Contrastando diferentes registros seicentistas da língua falada na costa, considerando algumas condições de impressão e escrita destes textos, Rodrigues (2010) constata que há “alguma diversidade (…) entre a fala dos tupis e a dos demais falantes da língua brasileira, diversidade que aparece também nos textos em língua indígena escritos por Anchieta nos primeiros dez anos em que esteve atuando entre os tupis (Rodrigues, 2010: 28⁴)”. Isso não é dado novo.

No contato com as nações da costa brasileira é possível que os jesuítas tenham esbarrado nas cerca de 79 línguas descritas ou meramente referidas na narrativa extensa de Fernão Cardim (1925)⁴. Curiosamente, esta diversidade foi ignorada em seu uso primeiro pois aos jesuítas importava tratar aquelas línguas não travadas, isto é, ignorava-a aquelas línguas “muito difíceis de pronunciar, línguas consideradas anômalas dentro do

---

³ Para esclarecer o lapso na citação, a variação que esta citação faz referência é a pronúncia dos verbos acabados em transitivos iniciados por m que não recebem o prefixo relacional – i após o prefixo do sujeito, tendo nulo em seu lugar (cf. Rodrigues, 2010: 28-29).

⁴ Entre diversas nações, sobre os Tupinambás, cuja variedade é foco neste estudo, assim diz este registro: “Outros há a que chamam Tupinabas: estes habitam do Rio Real até junto dos Ilhéus; estes entre si eram também contrários, os da Bahia com os do Camamu e Tinharé; por uma corda do Rio de São Francisco vivia outra nação a que chamavam Caaeté, e também havia contrários entre estes e os de Pernambuco. Dos Ilhéus, Porto Seguro até Espírito Santo havia outra nação, que chamavam Tupinaquim; estes procederam dos de Pernambuco e se espalharam por uma corda do sertão, multiplicando grandemente, mas já são poucos; estes foram sempre muito inimigos das cousas de Deus, endurecidos em seus erros, porque eram vingativos e queriam vingar-se comendo seus contrários e por serem amigos de muitas mulheres. Já destes há muitos cristãos e são firmes na fé”. (Cardim, F., 1925). O percurso deste livro é curioso. Embora tenha sido recuperado no movimento modernista como um registro fidedigno da “realidade da nação brasileira” sabe-se que a sua primeira impressão é realizada em terras inglesas em 1625, pois o navio de seu autor naufragou e, assim, seus escólios e sobreviventes do naufrágio são capturados pelo capitão James Cook. Escrito entre as décadas de 1580 e 1625, data da primeira publicação do Tratado, este livro é reimpreso pelos lusitanos apenas no século XVIII a mando de D. Manuel, para divulgar a história portuguesa, ilustrando, assim, o seu império. Não sei bem, por isso, se este livro pode ser atualizado como referência aos escritos jesuíticos da Companhia de Jesus. Por outro lado, sua atualização no século XX é bastante proveitosa para o conhecimento da diversidade de línguas indígenas dos seicentos brasileiros e, neste texto, serve a este fim.

---

⁵ A importância do Tupi é divulgada em terras não brasileiras por meio da circulação de livros, em especial, de relatos de viajantes. Conforme Rodrigues (2010): “Um dos primeiros escritores brasileiros a destacar o nome tupi foi o poeta e pesquisador Gonçalves Dias, em sua poesia romântica de grande ressonância. O naturalista Martius (1863-67), no primeiro ensaio de classificação dos povos indígenas do Brasil, distinguiu nove grupos étnicos, ao primeiro dos quais deu o nome de tupis e guaranis; essa classificação foi reorganizada pelo etnólogo von den Steinen (1886), que distinguiu oito grupos e chamou o
Conforme Rodrigues, se por um lado a partir da rememoração Tupi no século XIX como a língua originária brasileira esta variedade ganha destaque entre os estudos, por outro o Tupinambá “foi caíndo em desuso com o quase total extermínio” dos tupinambás na Bahia e a “progressiva catequização e assimilação” (Rodrigues, 2010: 30) dos tupinambás no Maranhão. Esta repercusão pode ser sentida tanto no desenvolvimento de estudos contemporâneos quanto na apropriação das línguas em contato com jeitadas das expedições ultramarinas ao fixarem a gramática da língua indígena.

Há uma controvérsia bastante conhecida sobre a delimitação da língua Tupi Antigo em oposição à Tupinambá e, diz-se, se partirmos do preceito de que estas línguas devem ser comparadas em sua variação histórica, mesmo estudiosos como Aryan Dall'Igna Rodrigues teriam “confundido” os termos Tupinambá e Tupi Antigo, embora tenha levado a termo um trabalho magistral na língua a que nos referimos neste trabalho. Contradições à parte, recusamos esta delimitação arbitrária, bem como os discursos que a aparam, pois que a noção de tempo histórico vinculada a este tipo discussão é aquele progressista, acumulador, no qual exemplos passados podem servir para atualizações presentes. Outro motivo para desconsiderarmos esta discussão histórica e formalista (e talvez o mais contundente) é porque a nós importa a atualização da língua em seu contexto contemporâneo, de revitalização e constituição identitária para as comunidades indígenas em Olivença. Sendo um estudo para revitalização da língua Tupi Antigo como língua estrangeira na comunidade Tupinambá de Olivença os processos linguísticos devem ser respeitados em seu uso contemporâneo.

O efeito desta história é bem conhecido entre os Tupinambás de Olivença e, mesmo lá, em uma comunidade que teve sua língua violentamente apagada, predomina-se uma “noção geral de que o modelo, o verdadeiro exemplo tópico das línguas indígenas do Brasil são os dialetos Tupi da costa” (Câmara Jr., 1979: 100), argumento que Eduardo de Almeida Navarro não se cansa de lançar mão em seu Curso Moderno de Tupi Antigo, chegando ao extremo de escolher como verbo para “chegar” um verbete citado apenas uma única vez na Gramática de Figueira (o iepotar). Chegaram os Portugueses e la nave va...

Anterior à Assessoria Linguística do Projeto Tupinambá, um Curso de Tupi era ministrado na comunidade pelos próprios professores das escolas. O livro de referência para este estudo era o Curso Moderno de Tupi Antigo, de Eduardo Navarro (2005), e, por isso, a primeira lição do livro, “Chegaram os portugueses”, foi estudada durante as oficinas oferecidas em 2011 na escola sede. Este manual, no entanto, é a) destinado a professores que já estejam familiarizados com algum estudo gramatical de alguma língua, o que não é o caso para todos os professores indígenas da escola e b) não cumpre o fim pedagógico de ensinar às crianças da escola estruturas da língua Tupinambá. Espera-se que, com o desenvolvimento de oficinas nas escolas, novos textos dos professores e dos alunos, bem como cantigas e mitos da comunidade, sejam integrados ao ensino da língua Tupinambá nas escolas”.

A permanência das guerras aos indígenas por meios aparentemente pacíficos é história que, infelizmente, conta com grande documentação na historiografia brasileira. Isso não significa, entretanto, que os Tupinambás não tenham resistido (como é comum esta nação ser referida nas histórias desde os seiscentos). Uma das tentativas de revitalização de sua cultura e da língua dos seus ascendentes partiu da própria comunidade indígena que, tendo participado do encontro C-Indy na Universidade Estadual da Bahia, organizado pela professora Consuelo Costa, requisitaram um Curso de Tupi, a princípio na escola Sapucaeira, em Olivença, na intenção de implantar uma escola bilíngue.

2. References


primeiro deles simplesmente tupis. Já dez anos antes Couto de Magalhães, autor brasileiro de grande prestígio, tinha publicado, sob patrocínio do governo imperial, o seu curso de língua geral amazônica...” (p.30).

6 Ao leitor atento que se interesse pelas questões de variação e sobredeterminações acerca a língua Tupinambá, Tupi Antigo, possíveis divergências entre os modos de nomear as línguas deixamos como sugestão a bibliografia seguinte: Freire, J.R.B. & Rosa, M.C. (2003); Câmara Jr, J.M. (2003).

7 Para aqueles curiosos, é interessante compreender que este estudo do Tupi na escola indígena de Olivença é amparado por um conjunto de leis da Bahia, a saber, a Lei no. 18.629/2010 (que institui o plano de carreira para o professor indígena na Bahia); pelo Decreto n. 8.741 de 12 de março de 2013 que cria a estrutura de ensino indígena da Bahia; pelo Decreto 2.010/2004 que estabelece diretrizes e procedimentos para a organização e oferta da Educação escolar indígena no sistema Estadual de Ensino da Bahia.
Leite.