SPOKEN CORPORA COMPILATION AND ANNOTATION
Building a corpus for comparative analysis of language attrition

Lúcia de Almeida FERRARI
Universidade Federal de Minas Gerais (UFMG)
Av. Antônio Carlos, 6627 - Belo Horizonte/MG - Brasil
ferrari.lu@gmail.com

Abstract

The aim of this research is the study of first language attrition of Italian L1 in contact with Brazilian Portuguese. Language attrition is the gradual decline or the loss of a first or second language by an individual. This is a corpus-based study: a corpus of oral spontaneous speech was collected using eight different subjects. This corpus, composed of 21298 words, was compared with fourteen different texts from the Italian C-ORAL-ROM (Cresti & Moneglia, 2005). The results were then compared with those of previous studies by Raso and Vale (2007, 2009). The attrition of Italian L1 was confirmed, with a few differences that may deserve further and deeper analysis in future studies. The variation of the percentage of loss between the two researches seems to be mostly due to: 1) differences in typology of texts; 2) different diaphasic varieties; 3) different pragmatic contexts. The greater dissimilarities are noticed between the two reference corpora. Finally, data seems to confirm that attrition is a process that doesn't come to a halt after the first decade, but one that continues in time.

Keywords: attrition; corpus; Italian; clitics.

1. Introduction

This paper discusses the methodology employed to build a corpus for first language attrition study and the results obtained comparing it to previous researches.

The definition of L1 attrition is a "non-pathological decrease in proficiency in a language that has previously been acquired by an individual i.e. intragenerational loss" (Köpke & Schmid, 2004: 5). The process is due to two factors: the influence of L2 system and the lack of use of, and exposure to, the L1. In our case the study is about Italian L1 attrition in contact with Brazilian Portuguese.

Previous researches (Raso & Vale, 2007, 2009) on a group of clitics adopted the corpus methodology to investigate the degree of attrition of a group of Italians living in São Paulo for 20 to 30 years.

The aim of our research was to create a corpus with a greater diaphasic variety, in order to ensure the higher possible degree of spontaneousness.

The object of the study was the same group of clitics analysed by Raso and Vale, that is: ci attualizzante, lessicalizzante and locativo; ne partitivo, argomentale and locativo and the third person accusative clitics lo, la, li, le l’.

2. Corpus design and methods

Raso and Vale researches analysed a corpus extracted from a collection of interviews (Revista de Italianística, 1997), for a total of 18080 words, and compared it with an excerpt of the BADIP corpus (De Mauro et al., 1993) for a total of 18080 words.

To guarantee their complete acquisition of the language and some kind of meta-linguistic remark skills, the participants were all Italians, born and raised in Italy until the coming of age, with a high school degree obtained in Italy and, preferentially, a college degree.

In choosing the informants for our research we followed the same criteria; the required contact period with Brazilian Portuguese was of at least eight to ten years, as recommended by the attrition bibliography.

Eight different participants were selected: we were able to obtain various types of interactions, namely: a conversation between three people watching a soccer match on TV; five dialogues (one between a couple making dinner, one between two sisters, one about sports, one during a meal, and a discussion about doctors); and two monologues in which people spoke about their life experiences. Therefore, the resulting corpus reflects a higher degree of diaphasic variation than the one used by Raso and Vale. This is a key element for our study because it's correlated to a greater spontaneity of speech and can allow us to study the actual degree of attrition in real-world situations.

Our corpus has a total of 21298 words; as a reference corpus we selected fourteen different texts, the most similar to ours, from the Italian C-ORAL-ROM (Cresti & Moneglia, 2005), to a total of 21224 words. The choice of C-ORAL-ROM is due to it being a third generation corpus, highly spontaneous, transcribed in CHAT format (McWhinney, 1994), the same one we used in our corpus, and to the fact that all the digital recordings are available (as they are for our corpus).

The first step was to search our corpus and the Italian C-ORAL-ROM for excerpts containing the clitics we were studying and their collocations. Data were then normalized for comparison purpose. Every clitic was compared in normalized form and as a percentage.

The second step was to compare the results of the above described research with those of the studies by Raso and Vale. Again, all data had to be normalized. Several sets of data, as we will show, were extrapolated and compared, in order to point out similarities and differences between the results of both studies and to formulate hypotheses.

3. Data Collected

In the following section we will present the data we collected and the comparison made between our corpus and the reference one (C-ORAL-ROM), and between our findings and those of the Raso-Vale study. Each clitic will be examined separately and, at the end, we'll offer
our conclusions.

3.1 An overview

In this paper all data will be provided in their normalized form, to facilitate the comprehension of the comparison we made.

Our corpus, named in the below tables Raso-Ferrari corpus, presents a total of 191,09 occurrences of clitics every 10000 words, while the Italian C-ORAL-ROM presents 304,37.

<table>
<thead>
<tr>
<th>CLITICS</th>
<th>Raso-Ferrari Corpus</th>
<th>Raso-Vale Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOTAL</td>
<td>191,09</td>
<td>179,18</td>
</tr>
</tbody>
</table>

Table 1: Normalized values (per 10000 words) in both attrition corpora studied

In percentage this means a 37.31% decrease compared to the reference corpus. Looking at the previous studies, the Raso-Vale corpus presents 179,18 occurrences, while the BADIP corpus presents 270,46 occurrences; in percentage, that's a 33.74% decrease. This difference is relatively small and our study seems to confirm the attrition of our test group.

Our data turn out to be much more interesting when the clitics are split, as seen in table 2: it's possible to observe considerable differences between the two studies. While in the Raso-Vale researches the number of ci attualizzanti increases by nearly 10%, our study shows a decrease of about 50%.

<table>
<thead>
<tr>
<th>CLITICS</th>
<th>Raso-Ferrari Corpus/Italian C-ORAL-ROM</th>
<th>Raso-Vale Corpus/BADIP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ci attualizzanti</td>
<td>-50.91</td>
<td>9.34</td>
</tr>
<tr>
<td>Ci lessicalizzanti</td>
<td>-54.72</td>
<td>-70.16</td>
</tr>
<tr>
<td>Ci locativo</td>
<td>-84.22</td>
<td>-38.47</td>
</tr>
<tr>
<td>Io, la, li, le, l'</td>
<td>-23.81</td>
<td>-45.39</td>
</tr>
<tr>
<td>Ne (total)</td>
<td>-25.86</td>
<td>-51.71</td>
</tr>
<tr>
<td>TOTAL</td>
<td>-37.21</td>
<td>-33.74</td>
</tr>
</tbody>
</table>

Table 2: Percentage variation between attrition studies

This is the most evident discrepancy, but there are others: in the case of the ci lessicalizzanti we can see a 70.16% decrease in the Raso-Vale corpus, greater than the 54.72% decrease registered in ours; the ci locativo, on the other side, shows a decrease of about 84% in our study, while in the Raso and Vale research is about 38%; third person accusative clitics show a decrease of nearly 24% in our corpus and about 45% in the Raso and Vale studies; finally, the total ne clitics show a decrease of about 26% in our studies and nearly 52% in the previous ones.

In an attempt to explain such remarkable differences, table 3 shows the normalized data of all corpora used in both researches to see how much weight each references corpora have in the total values.

<table>
<thead>
<tr>
<th>CLITICS</th>
<th>Raso-Ferrari Corpus</th>
<th>Raso-Vale Corpus</th>
<th>Italian C-ORAL-ROM</th>
<th>BADIP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ci attualizzanti</td>
<td>63.38</td>
<td>64.71</td>
<td>129.09</td>
<td>59.18</td>
</tr>
<tr>
<td>Ci lessicalizzanti</td>
<td>4.69</td>
<td>1.65</td>
<td>10.36</td>
<td>5.53</td>
</tr>
<tr>
<td>Ci locativo</td>
<td>1.4</td>
<td>13.27</td>
<td>8.95</td>
<td>21.57</td>
</tr>
<tr>
<td>Io, la, li, le, l'</td>
<td>111.27</td>
<td>91.81</td>
<td>146.06</td>
<td>168.14</td>
</tr>
<tr>
<td>Ne (total)</td>
<td>15.02</td>
<td>7.74</td>
<td>20.26</td>
<td>16.03</td>
</tr>
<tr>
<td>TOTAL</td>
<td>191.09</td>
<td>179.18</td>
<td>304.37</td>
<td>270.46</td>
</tr>
</tbody>
</table>

Table 3: Normalized values (per 10000 words) in all analysed corpora

As it can be easily seen, the two attrition corpora don't show such a huge difference as percentages could induce to believe. In fact, the values of the ci attualizzanti are mostly the same, while percentage data between the two studies suggested a considerable divergence. Also, third person accusative clitics don't show such a big difference in normalized values. The most significant differences are the ci locativo and total ne clitics, but, as we'll see, those discrepancies can be explained quite easily.

What is surprising is the strong difference perceptible between the two reference corpora and the two attrition corpora, and between the two reference corpora themselves. Table 3 shows clearly that the ci attualizzanti found in the Italian C-ORAL-ROM are more than twice than those found in the BADIP corpus: 129.09 versus 59.18, respectively. The other clitics, excluding third person accusative clitics, also present two- or three-fold differences. We can assert, than, that the differences between the results of the studies can be due to the differences between reference corpora; but this isn't the only explanation, as we'll see analysing some particular cases.

3.2 The ci locativo clitic

As we can observe in table 4 below, a big difference could be seen in the use of the ci locativo both in the attrition corpora and in the reference corpora.

Data suggests that the Italian C-ORAL-ROM corpus has a much smaller number of occurrences of this clitic than the BADIP corpus. The same happens with the Raso-Ferrari corpus in comparison with the Raso-Vale.

<table>
<thead>
<tr>
<th>CLITICS</th>
<th>Raso-Ferrari Corpus</th>
<th>Raso-Vale Corpus</th>
<th>Italian C-ORAL-ROM</th>
<th>BADIP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ci locativo</td>
<td>1.4</td>
<td>13.27</td>
<td>8.95</td>
<td>21.57</td>
</tr>
</tbody>
</table>

Table 4: Normalized ci locativo (per 10000 words) in all analysed corpora
What can explain this behaviour? Our hypothesis is that both the Italian C-ORAL-ROM and the Raso-Ferrari corpora contain texts much more spontaneous than the other two corpora. The Raso-Vale corpus is mostly composed by interviews, where the speaker was asked about his migration and travels, so he would use this clitic much more than in a normal conversation. The same happens in BADIP, a corpus based on much less spontaneous types of interactions than C-ORAL-ROM.

So, the divergent data has their explanation in the different kind of texts and interactions that compose the corpora analysed.

3.3 The ne clitics

Another clitic that registers divergent values between corpora is ne. To better understand this behaviour it's necessary to split this clitic into its various functions and see the resulting figures as shown in table 5 below.

<table>
<thead>
<tr>
<th>CLITICS</th>
<th>Raso-Ferrari Corpus</th>
<th>Raso-Vale Corpus</th>
<th>BADIP</th>
<th>Italian C-ORAL-ROM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ne partitivo</td>
<td>8,92</td>
<td>3,31</td>
<td>8,84</td>
<td>13,19</td>
</tr>
<tr>
<td>Ne argomentale</td>
<td>6,1</td>
<td>4,42</td>
<td>5,53</td>
<td>7,06</td>
</tr>
<tr>
<td>Ne locativo</td>
<td>0</td>
<td>0</td>
<td>1,65</td>
<td>0</td>
</tr>
<tr>
<td>TOTAL Ne</td>
<td>15,02</td>
<td>7,74</td>
<td>16,03</td>
<td>20,26</td>
</tr>
</tbody>
</table>

Table 5: Normalized values (per 10000 words) in all corpora compared

We can see that the values of the Raso-Ferrari corpus and the Italian C-ORAL-ROM are higher than the other two corpora, both the attrition one and the reference one.

Again, what in our opinion may explain the divergent behaviour of these data is the different kind of texts that compose the corpora and the diaphasic variation in texts. The Italian C-ORAL-ROM is a much more modern corpus than BADIP and is representative of the actual spoken language in Italy. Proof of this is the higher number of ne partitivi in comparison with ne argomentali, less used in modern Italian, and the total absence of ne locativi, the latter being, as Russi (2008) supports, totally set aside nowadays. As data indicates, the Raso-Ferrari corpus also depicts this situation, with a minor degree of attrition in relation to the Raso-Vale corpus. This last consideration induces us to think that, as our corpus is composed by interaction of people that have been living in Brazil for less longer than the ones who were interviewed for the Raso-Vale corpus, this may be a possible evidence of the fact that attrition continues in time and does not, as theorized by many scholars (for example Kopke and Schmidt, 2004), come to a halt after the first decade.

3.4 Ci clitic in the verbs esserci and averci

The ci clitic can have various functions in Italian. We saw above that it can have a locative use but, as we'll explain, it can also be a particle lexicalizing a verb connected to it. In this paper we call ci attualizzanti the forms esserci and averci, where the grammaticalization is complete, and ci lessicalizzanti all the other forms, like andarcì (going to a place) or starci (to agree to do something), independently of the degree of grammaticalization.

This distinction is important to understand our analysis. In the first place, as shown in table 3 above, in the Italian C-ORAL-ROM the number of ci lessicalizzanti is double respect to BADIP, the other reference corpus. This indicates, once again, the recency of the first corpus. In both attrition corpora the values decrease quite a lot, much more than in the Raso-Vale one, confirming our assumption that attrition increases over time.

If we observe ci attualizzanti, we can notice that both attrition corpora exhibit very similar values: the Raso-Ferrari corpus has 63,38 occurrences every 10000 words while the Raso-Vale has 64,71 occurrences. What is quite surprising is the strong difference between the two reference corpora. This time we expected a smaller number of occurrences in the Italian C-ORAL-ROM: again, the explanation lies in the broader diaphasic variation of the texts and in the spontaneity of them, as this form is a pretty comprehensive verb form.

We won't linger over the ci lessicalizzanti as the values are very small, to a point where it isn’t possible to go further in our investigation.

On the other side, we will analyse in a little more depth both esserci and averci.

<table>
<thead>
<tr>
<th>CLITICS</th>
<th>Raso-Ferrari Corpus</th>
<th>Raso-Vale Corpus</th>
<th>BADIP</th>
<th>Italian C-ORAL-ROM</th>
</tr>
</thead>
<tbody>
<tr>
<td>esserci esistenziale</td>
<td>48,36</td>
<td>49,77</td>
<td>24,33</td>
<td>78,68</td>
</tr>
<tr>
<td>esserci presentativo</td>
<td>7,51</td>
<td>9,04</td>
<td>7,19</td>
<td>7,53</td>
</tr>
<tr>
<td>TOTAL esserci</td>
<td>55,87</td>
<td>59,18</td>
<td>31,52</td>
<td>86,22 / 66,9*</td>
</tr>
<tr>
<td>averci</td>
<td>7,51</td>
<td>5,53</td>
<td>27,65</td>
<td>42,87</td>
</tr>
</tbody>
</table>

Table 6: Normalized values (per 10000 words) of the ci attualizzanti in all corpora analysed (from Panunzi, 2010)

Again, we had to split the data, dividing the esserci form into esistenziale, when it can replace other verbal forms of existence; and presentativo, constituted by the form esserci+SN+che pseudo-relativo, a conformation that de-emphasize, from the cognitive point of view, the

1 A discussion about the functions of ci with verbs can be found in Sabatini (1985, 1986) and Russi (2008).
structure of a totally new and rhematic phrase. Previous *esser* and *avere* data were reviewed by Vale (2009) and we'll present them together with those by Panunzi (2010) for comparison purpose.

It's easy to notice that in both attrition corpora the *esser* existenziale values don't present significant differences. On the other way, the reference corpora exhibit a large difference in number of occurrences: 24,33 every 10000 words in the BADIP corpus and 78,68 in the Italian C-ORAL-ROM. In the case of the *esser* presentativo, both the attrition corpora and the reference corpora don't show considerable differences, corroborating the fact that the informative function this form carries doesn't depend on textual variety.

We now have to explain the great difference in the *esser* existenziale between the reference corpora. In their studies Raso and Vale suggested, referring to the attrition corpus, that a high presence of this form would mean a lack of lexical variability. We can agree with this theory, but we can also assume that the main reason of a more than triple value of *esser* existenziale in the Italian C-ORAL-ROM in comparison to BADIP is due, once again, to the greater diaphasic variety of texts and, most of all, to their spontaneity. To support this hypothesis, table 6 presents the values of total *esser* found in Panunzi (2010) who analysed the entire corpus of 300000 words. With a more general view it's possible to see that the differences between the two reference corpora continue to be quite noticeable, but smaller than the ones presented previously.

The case of *avere* is quite different. Raso and Vale suggested that future studies would show a smaller degree of attrition of this form, as it became widespread in Italy only after the migration of their informants. Instead of this, our research shows a quite similar level of attrition. As a possible explanation, we can propose that in this case the phenomenon may be due to the subjects of the research being mostly Italian teachers or translator, or individuals otherwise working in an Italian-speaking environment. As their professions require a high degree of proficiency, we can suppose that they tend to practice a higher level of self-control when speaking, especially when it comes to using a form that, while nowadays quite accepted in Italy, they perceive as incorrect or inaccurate.

### 3.5 The third person accusatives

We will now analyse the attrition of the third person accusatives *lo, la, li, le, l’*. As table 3 above shows, normalized data of all corpora don't seem to demonstrate a great difference of values between these clitics, and the degree of attrition seems quite low. But once again we have to split the data to obtain a more complete overview. In table 7 we can observe third person clitics divided by function and dislocation in the phrase.

In both attrition corpora, a first glance to the non-phoric dislocated constituents, informatively neutral, confirms the general opinion: the Raso-Ferrari corpus demonstrate an attrition process, albeit much lower than the one shown by the Raso-Vale corpus. This could confirm our theory that attrition continues to grow even after the first decade of contact with the L2.

If we look at the phoric dislocated constituents we can see that the situation is much more complicated. In their researches Raso and Vale found that left anaphoric constituents have an increase in values comparing to BADIP, in contrast with the decrease of the total dislocated constituents and, to an even greater extent, of the right dislocated constituents.

<table>
<thead>
<tr>
<th>CLITICS</th>
<th>Raso-Ferrari Corpus/Italian C-ORAL-ROM</th>
<th>Raso-Vale Corpus/BADIP</th>
<th>Raso-Ferrari Corpus/BADIP</th>
<th>Raso-Vale Corpus/Italian C-ORAL-ROM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonphoric dislocated constituents</td>
<td>-18,17</td>
<td>-50,98</td>
<td>-32,22</td>
<td>-40,82</td>
</tr>
<tr>
<td>Left anaphoric dislocated constituents</td>
<td>-21,01</td>
<td>26,02</td>
<td>-15,17</td>
<td>17,34</td>
</tr>
<tr>
<td>Right cataphoric constituents</td>
<td>-65,85</td>
<td>-53,55</td>
<td>-63,63</td>
<td>-56,39</td>
</tr>
<tr>
<td><strong>TOTAL lo, la, li, le, l’</strong></td>
<td>-23,81</td>
<td>-45,39</td>
<td>-33,82</td>
<td>-37,14</td>
</tr>
</tbody>
</table>

Table 7: Percentage variation between third person accusatives in a cross analysis of all corpora studied

Our research confirms the decrease of non-phoric dislocated constituents but exhibits a decrease in the left anaphoric constituents and a greater reduction of the right cataphoric constituents.

A cross-analysis of all corpora values can give us an answer about those incongruous results. First of all, it is quite evident that when both attrition corpora are compared with BADIP the results of left anaphoric constituents grow. Once more it seems that we have to investigate the kind of texts every corpus presents and the context of appearance of the object clitic.

In fact, the use of an anaphoric pronoun in Italian in thematized phrases is mandatory, in order to constitute the cognitive semantic bound of an illocution. If the semantic referent is clear to the listener, it’s not necessary to constitute this cognitive semantic bound through a thematization, that requests the use of an anaphoric pronoun. To be clear, either in the Italian C-ORAL-ROM or in the Raso-Ferrari corpus, the texts are dialogical and very spontaneous: people know what are they talking about. The Raso-Vale and BADIP corpora, on the other hand, are more formal, with interviews or guided interactions, so people seems to be compelled to thematize the referents they are talking about, hence using the anaphoric pronouns much more.

In the case of the right anaphoric dislocations, it seems that the communicative situation effect plays a much smaller function, and a similar construction isn't
found in Brazilian Portuguese so, as it can be seen, the degree of attrition is higher.

4. Conclusion

This paper presented an L1 attrition corpus-based research. This study had the purpose to delve into this topic deeper than previous ones, building a new corpus with more up-to-date criteria.

As in previous investigations, attrition of Italian L1 in contact with Brazilian Portuguese is confirmed, with a few distinctions that we tried to explain.

The variation in percentage of loss between the two researches seems mostly be due to three reasons:

- differences in typology of texts;
- different diaphasic varieties;
- different pragmatic contexts.

The most relevant divergences can be noticed between the two reference corpora.

The facts above described can explain some seeming incongruous data, like the increase of the number of generic forms like esserci in the Italian C-ORAL-ROM or the absence of the ne locative clitic in our corpus.

Finally, smaller signs of attrition in our corpus in the case of third person accusative clitics can be a signal that the process doesn’t come to a halt after the first decade but continues over time.

We are aware of the fact that the set of data we collected is still too small for a general overview of the L1 attrition discussion, so we hope that this subject and the questions that remain open could be answered by future studies.

5. References


Annotating a corpus of spoken English: the Engineering Lecture Corpus (ELC)

Siân ALSOP, Hilary NESI
Coventry University
Priory Street, Coventry CV1 5FB, UK
alsops@uni.coventry.ac.uk, h.nesi@coventry.ac.uk

Abstract
This paper describes an approach to what we are calling the ‘pragmatic’ annotation of the Engineering Lecture Corpus (ELC). The ELC contains 70 English-medium engineering lectures from across the world, currently including Malaysia, New Zealand, the United Kingdom and Italy (www.coventry.ac.uk/elc). The lectures are in the form of videos, raw text transcripts and XML files encoded using traditional TEI methods, but also marked for a limited number of features which shed light on the specific nature of lecture discourse. These functions will be discussed in terms of: how the current working list was reached, markup and annotation processes, and possible uses of the complete corpus.

Keywords: lecture; engineering; annotation; corpora; pragmatic.

1. Concept behind the corpus
Academic staff and students are increasingly moving from country to country to receive and deliver academic lectures. However, although English is often used as a lingua franca in higher education, and although lecture topics and syllabuses for disciplines such as engineering and medicine tend to be similar around the world, it is likely that different cultural norms and expectations will result in different lecture styles and structures in different local academic contexts. This suggests that staff and students may need to adjust the way they deliver and receive lectures in unfamiliar academic contexts, and that they may benefit from corpus linguistic insights when making these adjustments.

The corpus annotation of features other than syntax and part of speech is extremely time-consuming and encumbered by questions of subjectivity (Meyer, 2002; Leech, 2005; Smith, 2008). Some spoken corpora such as the London-Lund Corpus (LLC) (Garside et al., 1997) and the spoken component of the HKCSE business corpus (Warren, 2004; Cheng, 2004) have been manually encoded for prosodic features such as tone units, pitch and stress, but very few corpora have been annotated from a functional perspective, because of the labour intensive nature of such work, and because of the degree of interpretation it requires.

A number of small written corpora have been marked up in terms of generic moves and steps (see, for example, Durrant & Mathews-Aydinli, 2011), and classroom interaction in the *Singapore Corpus of Research in Education* (SCoRE) has been marked for pragmatic and pedagogical features (Peréz-Paredes and Alcaraz-Calero, 2009), but as far as academic lectures are concerned, progress with pragmatic mark-up has been very slow. Young (1994) identified a sort of generic move structure in academic lectures, consisting of various ‘phases’, each with a different communicative function, and Maynard and Leicher (2007) experimentally tagged a small subcorpus of 50 MICASE transcripts by identifying pragmatic features such as ‘advice’ and ‘disagreement’ in header metadata, but there does not seem to have been any prior attempt to mark up an entire corpus of lectures to reflect their structure or purpose.

The largest British lecture corpus, the British Academic Spoken English (BASE) corpus (Nesi, 2001), is only encoded for part of speech, pausing, and contextual information. The BASE corpus annotation follows TEI (Text Encoding Initiative, www.tei-c.org) conventions so that it can be compared with other similarly encoded corpora, but TEI has not traditionally been used to signal the function of larger stretches of discourse, and appropriate coding strategies are still under development.

By annotating what we are calling ‘pragmatic’ features, we are able to identify and describe features that are typical of the discourse; in this case, engineering lectures. It will also allow us to compare the styles of English-medium engineering lecturers in different parts of the world, and explore what role English-medium instruction currently has in the discipline of engineering.

2. The corpus
We have annotated six functions of the lecture within a cross-cultural corpus of 70 English-medium university level lectures across five areas of engineering (see Table 1). The ELC currently contains four subcorpora of lectures from: the United Kingdom (UK, four digit id. series: 1…), Malaysia (MS, id. series: 2…), New Zealand (NZ, id. series: 3…), and Italy (IT, id series: 4…).

<table>
<thead>
<tr>
<th>area of engineering</th>
<th>MS</th>
<th>NZ</th>
<th>UK</th>
<th>IT</th>
</tr>
</thead>
<tbody>
<tr>
<td>civil</td>
<td>4</td>
<td>27</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>mechanical</td>
<td>11</td>
<td>8</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>electrical</td>
<td>17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>graphics</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>telecoms.</td>
<td></td>
<td></td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>total lectures</td>
<td>15</td>
<td>28</td>
<td>30</td>
<td>7</td>
</tr>
<tr>
<td>total lecturers</td>
<td>9</td>
<td>4</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 1: ELC holdings

3. Categories annotated
The current set of six pragmatic features was arrived at through a three-stage process. The initial working list

Heliana Mello, Massimo Pettorino, Tommaso Raso (edited by), Proceedings of the VIIth GSCP International Conference : Speech and Corpora
was based on Nesi and Ahmed’s (2009) set of 15 features (outlined in Table 2). For the first pass at annotation, the lead annotator, in collaboration with local experts, worked through samples from each of the four subcorpora, cycling between the original working list and the functions that actually occur in the corpus.

Using this data-driven approach to refine the pragmatic categories annotated resulted in the first adjustment to the working list. At this stage, it became clear that some of the functions identified in the original working list (or elements) needed to be expanded to include subcategories (or attributes), and some should be hierarchically demoted and subsumed under a more general umbrella category (see Table 2). Such changes included incorporating: ‘review lecture content’ and ‘preview lecture content’ as attributes of the umbrella element ‘summary’; ‘personal narratives’ under ‘storytelling’, with the addition of the attribute ‘professional narratives’; and the six independent types of humour that were originally identified were subsumed under a single unified element, which was expanded to include five more attributes and ‘word play’. Two other elements from the original working list (‘reference to students’ future profession’ and ‘greetings’) and one partial element (‘register’ from ‘register and wordplay’) were not evident in sufficient quantity to justify their inclusion in the adjusted list when considered against the original criteria of identifying and describing typical engineering lecture discourse features.

The second pass at refining the clipboard was undertaken by a single researcher overviewing the entire corpus with the aim of ensuring consistency across all identified features. In this second adjustment, attributes of the ‘summary’ element were further expanded to identify reviews of previous and current lecture content, and previews of current and future lecture content. Attributes of the storytelling element were replaced; the distinction in genres of anecdote, exemplum, narrative and recount (cf. Plum, 1988; Martin, 2008; also see Alsop et al. forthcoming) were considered to be more useful than the former limited description of narrative type (as ‘personal’ or ‘professional’).

<table>
<thead>
<tr>
<th>Nesi and Ahmed (2009)</th>
<th>1st adjustment</th>
<th>2nd adjustment</th>
</tr>
</thead>
<tbody>
<tr>
<td>element</td>
<td>attribute</td>
<td>element</td>
</tr>
<tr>
<td>prayer</td>
<td>prayer</td>
<td>prayer</td>
</tr>
<tr>
<td>housekeeping</td>
<td>housekeeping</td>
<td>housekeeping</td>
</tr>
<tr>
<td>defining term</td>
<td>defining term</td>
<td>defining term</td>
</tr>
<tr>
<td>review lecture content</td>
<td>summary</td>
<td>review lecture content</td>
</tr>
<tr>
<td>preview lecture content</td>
<td></td>
<td></td>
</tr>
<tr>
<td>personal narratives</td>
<td>storytelling</td>
<td>personal narrative</td>
</tr>
<tr>
<td>teasing</td>
<td>humour</td>
<td>bawdy humour</td>
</tr>
<tr>
<td></td>
<td></td>
<td>disparagement</td>
</tr>
<tr>
<td></td>
<td></td>
<td>mock threats</td>
</tr>
<tr>
<td></td>
<td></td>
<td>playful humour</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sarcasm</td>
</tr>
<tr>
<td>register and word play</td>
<td></td>
<td></td>
</tr>
<tr>
<td>greetings</td>
<td></td>
<td></td>
</tr>
<tr>
<td>reference to students’ future profession</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Refining the clipboard

The ELC is a growing corpus and we are constantly seeking new contributions from around the world. Because the pragmatic categories annotated are largely data-driven, we anticipate that further adjustments to the working list of functions may be made as the corpus expands and a larger data set becomes available. An early example of the need to encode an unexpected category is that of ‘prayer’,
which only occurs in the Malaysian subcorpus. Given the highly technical content of large stretches of the language that currently constitute the ELC, we predict that further emphasis may need to be given to the way in which specialized vocabulary is conveyed. ‘Defining’, for example, could be subsumed under a new ‘explaining’ umbrella element, and further attributes (for example, ‘categorising’, ‘equating’, ‘naming’ and ‘translating’) added. Similarly, if storytelling emerges as a more prominent function as the corpus grows, it may be useful to revisit the original significance of describing ‘personal’ involvement and attribute another layer of annotation to the current categories by specifying whether the instance of storytelling is based on the lecturer’s own experience or the experience of others.

4. Examples of categories annotated
When identifying the boundaries of pragmatic categories, we have worked on the principle of including enough data so that the chunk of text annotated makes sense in isolation from its immediate context. Where boundaries were unclear, the widest scope was incorporated.

Some of the ELC categories are self-explanatory, such as ‘prayer’, or most usefully clarified by the subcategories attributed to them, such as ‘humour’ or ‘story’. Some require further explanation. ‘Housekeeping’, in this context, refers to instances where lecturers talk about academic commitments and events external to the lecture. Also, ‘defining’ refers to the specific explanation of the meanings of technical terms in the ELC.

Given the inevitably somewhat subjective nature of the annotation process, we do not consider rigidly prescriptive definitions of the categories described to be either possible or desirable. Table 3, however, gives some examples from the current corpus.

<table>
<thead>
<tr>
<th>Element</th>
<th>Attribute</th>
<th>Example of discourse</th>
</tr>
</thead>
<tbody>
<tr>
<td>defining</td>
<td></td>
<td>so mathematically if we define the force the magnitude of the force as F and the angle that defines its direction to the horizontal is theta then simple trigonometry of triangles our horizontal component will be F cosine theta and our vertical component will be F sine theta simple enough (1001)</td>
</tr>
<tr>
<td>housekeeping</td>
<td></td>
<td>okay so there will be no class this Thursday and Friday because has been replaced here today (2010) how far have those certificates got well bring what’s left down to the front and anybody else who wants their certificate come down to the front (1012)</td>
</tr>
<tr>
<td>humour</td>
<td>mock threat</td>
<td>I will open it up again for another two weeks except for the person whose phone's going off cause they’re not gonna be able to sit down for about a month (1004)</td>
</tr>
<tr>
<td></td>
<td>teasing</td>
<td>after a good lunch I’m sure you can answer what's the purpose of the horizontal curve (2009)</td>
</tr>
<tr>
<td></td>
<td>irony</td>
<td>now today is a great day because we’re going to allow the charge to move we’re going to have current so don’t get too excited (3005)</td>
</tr>
<tr>
<td>story</td>
<td>narrative</td>
<td>I hate to admit to this one but one site I was on we had cube failures and the reason was that when I’d been sending the cubes off I’d been having to break the ice on the top of the tank before I could get them out and um the tank had a heater in we just hadn’t bothered to get the spark to wire it in and ah fairly obviously by the time the area manager appeared to ah come and have a look and see what had gone wrong it was all wired in and working fine and we said oh no no problem with that would we do a thing like that and ah but okay sort of nevertheless it caused endless hassle the fact that we’d had these cube failures if you keep them too cold they’ll go down a low strength (1012)</td>
</tr>
<tr>
<td>summary</td>
<td>review previous lecture content</td>
<td>let’s just review back what we did yesterday we talked about the refrigerator yeah we talked about the refrigerator and you were introduced to refrigerators and the heat pump (2017)</td>
</tr>
<tr>
<td></td>
<td>preview previous lecture content</td>
<td>so what are we going to do today is we are going to wrap up chapter five the second law of thermodynamics yeah so today we should be able to determine finally the thermo efficiencies and the coefficient of performance for our ideal our reversible or our Carnot cycle (2019)</td>
</tr>
<tr>
<td></td>
<td>review current lecture content</td>
<td>main three things that have come out of here though out of these tests is yield stress ultimate stress and modulus of elasticity (3026)</td>
</tr>
<tr>
<td></td>
<td>preview future lecture content</td>
<td>in the next two lectures we’re actually going to delve a little bit into material properties and then we’re going to get back into the solid mechanics (3024)</td>
</tr>
</tbody>
</table>

Table 3: Examples of ELC pragmatic categories
5. Markup and annotation processes

The ELC files have been created by merging two separately stored sets of information: the main body of raw transcribed lecture discourse and the header metadata. The spoken lecture content, varying in duration between 41-104 minutes, was videoed then transcribed as plain text by a local expert\(^1\). TEI compliant header information - such as title, recording equipment, main speaker information, etc. - was generated from a master spreadsheet and outputted in XML format to create a skeletal file, including empty ‘body’ tags. The transcribed plain body text was then merged into the body tags, and TEI compliant markup - including container elements to mark utterances according to speaker identifier, empty elements to mark pauses, gaps (for example, marking inaudible speech), and limited kinesic and vocal descriptions that were essential to context (for example, ‘writes on board’ and ‘laughter’) – was manually added.

We have distinguished this type of ‘structural’ markup from the annotation (c.f. Garretson, 2011) of pragmatic categories because the process by which the boundaries of the pragmatic categories are identified involves a subjective linguistic analysis. We think, therefore, that it should not be described in the same way as the identification of the objective structural components of a text, such as utterances.

In terms of the storage of these annotations, the boundaries of pragmatic categories were initially annotated inline alongside the structural markup. This posed a problem of validity for the XML metadata because the language of lectures often serves more than one function; a story, for example, may also be humorous, in full or in part, causing XML elements to overlap. Similarly, pragmatic categories can span various utterances – a lecturer delivering housekeeping information may be interrupted by a student asking a question, for example – which also results in malformed XML syntax. In addition to the methodological questions linked to storing annotation inline alongside markup, we did not consider using a system of workarounds to force the annotation into a well-formed state to be a desirable option.

Instead, we have decided to convert our current inline annotations into stand-off form and store them in separate XML files. The advantages of this system are that the subjective analysis is stored separately and multiple other layers of annotation can be done on the same text. In addition to the current pragmatic annotation, detailed kinesic or prosodic analyses could be applied, for example. One consideration that may be seen as a disadvantage, particularly in a corpus of spoken language, is that the raw text must be static in order that the indices of the annotations in the stand-off files are correct. This means that the original transcripts must be completely accurate before stand-off files can be created, and the transcripts cannot be edited post-annotation.

We intend to use the Dexter suite of software (http://www.dextercoder.org/index.html) for further coding and analysis once the current inline annotation has been converted into stand-off form. To achieve the conversion, the current annotation (but not the TEI-compliant structural mark-up) will be stripped out and an XSLT stylesheet will be used to convert these ‘pure’ versions of the marked up texts into XML files that are readable by stand-off annotation software (in this case, DexML). Next in the conversion chain, a code file will be created by looping through the original text and, for each inline annotation found, locating the exact stretch of text, and then identifying the indices for that stretch of text and creating a code instance for it in the code file. The result will be one file containing the ‘pure’ text and a code file. The codes that used to be inline annotations will then be in the form of editable stand-off annotation.

6. Possible uses of the corpus when complete

This data-driven process of pragmatic annotation will, we hope, eventually lead to the identification of linguistic features that typically realise the various purposes of lecture discourse. By encoding and then visualising these features we will be able to compare their location, duration and relative frequency in lectures delivered by local lecturers in different cultural contexts.

Looking at such data patterns allows one of two potential conclusions to be drawn. If significant consistency is identified in the way in which the annotated functions of language occur and are used across the subcorpora, we can conclude that key language functions are fundamental to the English-medium engineering lecture regardless of cultural context. We can then begin to build a model of the fundamental purposes of these lectures. If, on the other hand, significant variation in the uses of language functions is identified, we can begin to examine the role played by cultural difference in the delivery of the English-medium engineering lecture, regardless of consistency of language medium (English), discipline (engineering), and education level (undergraduate).

Our annotation system will be of interest to other corpus developers who intend to apply

---

\(^1\) Further information on transcribing conventions can be found here: <www.coventry.ac.uk/elc>.
pragmatic mark-up, and our comparative findings will be of interest to EAP and ESP practitioners, staff developers, and all academics and students on the move.

7. References


A multilingual speech corpus of North-Germanic languages

Janne Bondi JOHANNESSEN, Joel PRIESTLEY, Kristin HAGEN
The Text Laboratory, Department of Linguistics and Nordic Studies, University of Oslo
P.O. Box 1102 Blindern, 0317 Oslo, Norway
jannebj@iln.uio.no, joeljp@gmail.com, kristiha@iln.uio.no

Abstract

The Nordic Dialect Corpus project was initiated by the Scandinavian Dialect Syntax Network (ScanDiaSyn). In order to be able to study the North Germanic (i.e., Nordic) dialects, proper documentation of the dialect was needed. A corpus consisting of natural speech by dialect speakers was designed, in order to systematically map and study syntactic variations across the Scandinavian dialect continuum. The corpus was to be comprised of transcribed and tagged speech material linked to audio and video recordings. Further, it was decided that a user-friendly interface should be developed for the corpus, and that it should be available on-line. The corpus is now ready for use, and is described here.

Keywords: North Germanic languages; speech corpus; dialects; transcription; tagging; maps.

1. Introduction

The Nordic Dialect Corpus project was initiated by the Scandinavian Dialect Syntax Network (ScanDiaSyn). Documentation of the dialects was required, and it was decided that a corpus of natural, spontaneous speech was needed in order to systematically map and study syntactic variations across the Scandinavian dialect continuum. The corpus was to be comprised of transcribed and tagged speech material linked to audio and video recordings. Further, it was decided that a user-friendly interface should be developed for the corpus, and that it should be available on-line. The corpus is now ready for use and described in this paper.

The ScanDiaSyn network is a project umbrella where ten Scandinavian research groups collaborate.

Table 1: The Nordic Dialect Corpus in numbers

<table>
<thead>
<tr>
<th>Country</th>
<th>Informants</th>
<th>Places</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Denmark</td>
<td>81</td>
<td>15</td>
<td>242,885</td>
</tr>
<tr>
<td>Faroe Is.</td>
<td>20</td>
<td>5</td>
<td>62,411</td>
</tr>
<tr>
<td>Iceland</td>
<td>10</td>
<td>2</td>
<td>23,610</td>
</tr>
<tr>
<td>Norway</td>
<td>508</td>
<td>143</td>
<td>2,014,637</td>
</tr>
<tr>
<td>Sweden</td>
<td>126</td>
<td>39</td>
<td>307,861</td>
</tr>
<tr>
<td>Total</td>
<td>745</td>
<td>204</td>
<td>2,651,404</td>
</tr>
</tbody>
</table>

The corpus is now installed in the Glossa corpus system for user-friendly search and results handling (Johannessen et al., 2008; Johannessen, 2012).

There are a number of challenges that have had to be addressed, that we shall focus on in this paper:

- data collection should be carried out in several different countries
- the recordings should be transcribed, with different transcription standards and types for the individual languages;
- the corpus, consisting of different languages should be tagged;
- different tags should refer to the same entities for uniform search possibilities;
- informant metadata (gender, age, sex etc.) should be used as filters for search;
- different geographical divisions should be specifiable (e.g. country, county, town);
- all text from all languages should be accessible in same search;
- transcriptions should be linked to audio and video;
- results should be available in a number of different ways, including different export formats;
- informant data should be plotted on map.

Figure 1: The countries involved in the ScanDiaSyn project

The ten groups are spread across all of the five Nordic countries and one self-governed area. Three non-Nordic groups and a group working on Finnish dialect syntax liaise with the project through a NordForsk network. The core groups are from universities in Denmark, Faroe Islands, Finland, Iceland, Norway and Sweden.
2. Methodology for collecting speech

The corpus comprises recordings made in the five constituent countries of the Northern Germanic language area. From each country a number of sample points were selected specifically to capture dialectic variations.

There is some variation as to the combination of speakers in the corpus, given that the recordings were mostly done on national research funding and national research management.

In Norway, the Norwegian Dialect Syntax Project was funded by the Norwegian Research Council, a savings bank in North Norway and the University of Oslo. This ensured full funding of the recordings in a way that satisfied the criteria given by the researchers. From each point, four informants were identified, two men and two women, old and young. The informants were paired and asked to converse freely for approximately 30 minutes. Care was taken to create comfortable, informal surroundings, in order to encourage spontaneous, unaffected speech. Video equipment was set up, but the informants were left to themselves. Due to privacy legislation, a list of topics deemed off-limits was provided. This included subjects such as trade union and political party membership, as well as the naming of third parties, with the exception of public figures. Each informant also partook in a more formal interview, answering a standard set of questions. The Norwegian part also includes a number of old recordings from 1950–1980, provided by the Målforearkiv at the University of Oslo, and funded by the Norwegian Dictionary 2014 project.

The majority of the Swedish recordings (including Finland Swedish) were generously provided for use in the Nordic Dialect Corpus by the SWEDIA 2000 project. This project was originally aimed at collecting data for phonological research, but the data are mostly fully usable for our corpus, since this corpus, too, contains free speech. The Danish recordings were done by the Danish Syntax Project funded by the Danish Research Council, and contains six recordings from each place, but with no young people. The Faroese recordings were done on the ScanDiaSyn network budget (funded by the Nordic Research Council) and contains both young and old speakers. For Icelandic, the recordings have been less systematic, given a combination of funding and chronological synchronisation with the rest of the project. Some recordings have been generously provided by the University of Iceland, and some have been done by the network, using somewhat imperfect informants (linguists).

In spite of the diverse ways the recordings have been collected, the corpus is a unique source of spontaneous speech well suited for dialect research in syntax, but also for other linguistic disciplines.

3. Transcription and tagging

All recordings have been transcribed with standard orthography. In addition, all the Norwegian recordings and some of the Swedish ones (those of the Övdalian dialect) have been transcribed in a more phonetic way, following (for Norwegian) the method described in Papazian and Helleland (2005) and (for Övdalian) the orthography standardised by the Övdalian language council Rådjärum.

For each language, transcription software was used that inserts time codes directly into the transcribed text at suitable intervals, enabling the transcription to be presented with its corresponding audio and video. The transcriptions were partly done at a national level, and party in Oslo. Different software were used, but they were all adapted to the Transcriber format, which is the interchange format used in the project.

For the Norwegian and Swedish recordings that have also been phonetically transcribed, the process started with the phonetic transcription. These transcriptions were then translated to standard orthography using a program developed at the Text Laboratory, University of Oslo: an automatic dialect transliterator. The program takes as input a phonetic text and an optional dialect setting. Sets of text manually transliterated to orthography provide a good basis for training the program, enabling it to accurately guess the transliteration for further texts. The training process can be repeated, and the trained version can be used for similar dialects. Transcribing each recording twice therefore does not take as much as twice the time.

It is important that all words from the original phonetic transcription have an equivalent in the orthographic transcription. The two must be totally aligned for the results to be used in the corpus search system. Figures 3–5 below show how the phonetic transcription can be used in search and results presentation.

The languages are tagged individually with taggers for the respective languages. This means that each language has an individual tag-set decided by those who developed the taggers originally. The Danish transcriptions are lemmatised and POS tagged by a Danish Constraint Grammar Tagger developed for written Danish, see Bick (2003). The Faroese transcriptions first were tagged with a Constraint Grammar Tagger for written Faroese, see Trosterud (2009). Since spoken Faroese has a lot of words that are not approved in written standard Faroese, about half of the material was manually corrected after the Constraint Grammar tagging. Finally a TreeTagger was trained on the corrected material, and the rest of the transcriptions were tagged again. The Icelandic transcriptions were first tagged with a tagger for written Icelandic, see Loftsson (2008), and manually corrected afterwards. The orthographic version of the Norwegian corpus was lemmatised and POS tagged by a TreeTagger originally developed for Oslo speech. The Oslo speech tagger was trained on manually corrected output from the the written language Oslo-Bergen tagger, see Naklestad and Søfteland (2008). The Oslo speech tagger was then further adapted to the dialect corpus. The Swedish subcorpus was tagged by a modified version of the TwTagger developed by Kokkinakis (2003). The tagger was trained on the Swedish PAROLE corpus and manually
tagged orthographic Övdalian transcriptions. The tagger was applied to both the Swedish transcriptions and the orthographic versions of the Övdalian transcriptions.

Each language subcorpus has its own tag-set, but the tags have been standardised in the search system, making it possible to search for the same category across all the corpora. The linguist can choose for example all adjectives to be shown, irrespective of language. This is illustrated in Figure 5.

4. Metadata

The corpus has metadata relating to each informant and recording. There is information on the sex, age group, and place of origin; the latter being divided into country, region, area and place. Also, there is information on the year of recording, which is crucial for the Norwegian subcorpus, which contains both modern and old recordings, with 30–60 years between them. Finally, some recordings are distinguished according to genre: either interview or conversation.

The metadata can be used to create search filters for search in the corpus interface, as depicted in Figure 6.

The metadata is simply represented in a MySQL database, from which the corpus interface system Glossa picks the correct data according to the user’s needs.

4. Metadata

The corpus has metadata relating to each informant and recording. There is information on the sex, age group, and place of origin; the latter being divided into country, region, area and place. Also, there is information on the year of recording, which is crucial for the Norwegian subcorpus, which contains both modern and old recordings, with 30–60 years between them. Finally, some recordings are distinguished according to genre: either interview or conversation.

The metadata can be used to create search filters for search in the corpus interface, as depicted in Figure 6.

The metadata is simply represented in a MySQL database, from which the corpus interface system Glossa picks the correct data according to the user’s needs.

Figure 2: Searching for two words in sequence. The first is transcribed phonetically: itte for the orthographic word ikke ‘not’

Figure 3: The Both button is ticked, in order to have both kinds of transcription presented in the search results

Figure 4: Part of the search result for the query in Figures 2 and 3

Figure 5: Querying for adjectives in the corpus

Figure 6: Metadata filter in corpus interface

Figure 7: Metadata on each informant is available via a clickable button
Informant metadata can alternatively be found by clicking on the i-button (i for information) on the left of each concordance line in the results view, as in Figure 4, yielding the information displayed in Figure 7.

5. Multilingual search

Users in the ScnDiaSyn network originally wanted the possibility for multilingual search. They imagined that if they wanted, say, all occurrences of the negation equivalent to ‘not’ in English, a full results list would appear for all languages. However, this would have required a full multilingual dictionary, which does not exist either in paper or digital format for the North Germanic languages.

Instead, we put a link on the search interface to a multilingual word-list (Tvärslå) compiled by several previous language technology projects, including ScanLex in which the first author of the present paper was also in charge. This way the user can look up the equivalents of particular words in the other languages. The multilingual list is far from comprehensive, and also contains wrong language equivalents, since it is partly developed using automatic methods.

The search system Glossa allows for disjunctive searches, making it possible for several strings to be looked up at the same time. This is illustrated in Figure 8, for the orthographic versions of ‘not’ for Faroese, ikki, Swedish, inte, Danish and Norwegian, ikke, and Icelandic, ekki.

6. Links to audio and video

The user can click on the film or sound symbol to get the desired multimedia display. Figure 9 depicts the display.

The transcriptions have time codes, implemented as XML tags, at regular intervals, inserted at the time of transcription. This way there is a direct link between text and audio and video files, to be used by the corpus search system. These files are made available in Flash or Quicktime (according to the user’s choice).

7. Results presented on maps

For a corpus aimed at dialect research, getting results on a map view is very useful. The place of origin for each informant is located by GIS coordinates and the Google Maps API is used. Since every item in the corpus is connected to an informant, it means that for each word, string, piece of word or syntactic construction, there is a geographical location.

We have incorporated two ways of displaying results via maps. One way is that all hits are simply marked on the map. Figure 10 shows a search that asks for all hits where in a subordinate clause the negation ikke or inte (Norwegian, Danish, Swedish) precedes the subject. The geographical distribution is shown in Figure 11 below.
It has been debated in the literature whether this word order is allowed (see Johannessen & Garbacz, 2011). The red dots on the map in Figure 11 show where the hits are. Even if there are more recording places in Norway than in Sweden and Denmark, cf. Table 1, we see immediately that there are many more places where this construction is found in Norway than in especially Sweden. Since stress patterns also interfere with the generalisations, it is necessary of the user to listen to selected results, but the first picture given by the map is a very useful start.

The other way to use maps is only possible for those search results that belong to a set of two transcriptions. All the phonetic varieties are presented on a chart with the option of colouring each according to any classification one might be interested in.

In Figure 12 a chart can be seen of all the phonetic versions of the word vi ‘we’ in Norwegian. We have chosen to colour those variants that are pronounced with an initial bilabial /m/ sound with a deep violet colour, while the initial /v/ sounds are coloured yellow. The result is shown in Figure 13.

It should be quite clear from the map example that the opportunity of using a corpus combined with maps is an excellent way of finding isoglosses. The geographical limits for a phenomenon are readily apparent on the map. It should be noted also that dialect maps are not a new thing. However, in the past, researchers rarely had the chance to cover many places, so the present corpus may contain data that has never been known before. Secondly, the old maps were rarely the result of spontaneous speech, but rather of words and lists given by the researcher to the informants. The present solution, with a corpus of spontaneous speech as a direct basis for maps, gives good opportunities for both a comprehensive and a correct view of the geographical language variation.
8. Conclusion
We have presented the Nordic Dialect Corpus. We have shown how challenges posed by researchers in this project initiated by linguists have been met. The corpus contains recorded speech from five different languages, provides access to audio and video, as well as transcriptions—many of which are both phonetic and orthographic. All transcriptions are tagged. There are different options for results handling that we have not focused on here. However, we have shown how the map options work, and how this way of combining a corpus with a map solution provides advanced possibilities for identifying and representing isoglosses in a simple way.

9. Acknowledgements
We are grateful to all the people who have taken part in the corpus data collection, and in particular to Øystein Alexander Vangsnes, University of Tromsø, who has been a central person and spokesman for the Nordic Dialect Syntax network. We are also grateful to our old and new, permanent and temporary, colleagues at the Text Laboratory, UiO, who have helped at various points in the process, from transcription via transliteration of transcriptions to tagging. This work has been funded by national research councils in the Nordic countries, and by universities and smaller research funds.

10. References
Glossa. Available at: <http://www.hf.uio.no/tekstlab/glossa.html>.
Nordic Dialect Corpus. Available at: <http://tekstlab.uio.no/glossa/html/?corpus=scandiasyn>
Formation and annotation of North AMPER project’s corpus

Regina CRUZ1,2, Ilma SANTO, Camila BRITO1,2,6, Rosinele LEMOS4, Isabel DOS REMEDIOS1,5, João FREITAS3, Elizeth GUIMARÃES3, Lurdes MOUTINHO3
1UFPA; 2CNPq; 3Universidade de Aveiro; 4SEDUC; 5Master student; 6Scholarship PIBIC

Abstract

The study of the linguistic diversity on Pará state has as its aim to understand the main factors which cause the linguistic diversity in this region and the importance of these factors on the verbal manifestations of the people who speak the language in use: the Amazon’s Portuguese variety. On this paper we present how the formed corpora for the study of prosodic features of Brazilian Portuguese (PB) linguistic varieties spoken in Amazon are being organized, processed and annotated. The Prosodic Multimedia Atlas of Northern Brazil aims to verify the prosodic variations of Amazon PB to provide a sociolinguistic configuration of prosodic level of Pará state. So far the formed corpora are from the following cities: Belém, Bragança, Baião and Cametá. There are also three corpora in progress: Abaetetuba, Belém islands and Marajó island, all of which were formed according to the guidelines of AMPER project, following strictly its methodology, from the selection of informants to the protocol of the data collection.

Keywords: AMPER project; prosodic variations; Amazon; Brazilian Portuguese.

1. Introdução

This paper aims mainly to present how the formed corpora are being organized, processed and annotated for the study of the prosodic characteristics of the linguistic varieties of Brazilian Portuguese (PB) spoken in Amazon. This study is closely linked to the AMPER project, whose aim is to supply the acoustic and prosodic characterization of the Romance languages, as well as an online multimedia atlas (Contini et al., 2002: 227-230; Moutinho et al., 2001: 245-252). In relation to the Portuguese system, eleven institutions participate for the description of its three main varieties: European Portuguese, insular European Portuguese and Brazilian Portuguese (PB).

UFPA has already been participating of this Project since 2007, responsible for the Multimedia Prosodic Atlas of Northern Brazil. Currently, four atlas are in progress: a) Belém (Brito, in progress; Guimarães, in progress); b) Abaetetuba (Remédios, in progress); c) Marajó (Freitas, in progress), Baião (Lemos, in progress). This project has one atlas already finished. It belongs to Cametá (Santo, 2011).

2. The AMPER-North project

Since its entry in the AMPER project, UFPA’s team has already formed corpora of spoken Portuguese in the following places: a) Belém (Santos Jr., 2008; Cruz et al., 2008; Cruz & Brito, 2011); b) Bragança (Castilho, 2009); c) Baião (Lemos, in progress); and d) Cametá (Santo, 2011; Santo & Cruz, 2011). The formation of these corpora was made according to the AMPER Project guidelines, following its methodology, since the selection of informants until the protocol of the data collection. A detailed description of these methodological procedures is shown on item 3.

As this project has as an aim to form a Prosodic Multimedia Atlas of Northern PB, other three corpora are still in prediction of formation: a) of Abaetetuba’s city (Remédios, in progress); of Belém’s isles (Guimarães, in progress; Brito, in progress) of Marajó’s isle (Freitas, in progress). We have also a prediction to the formation of corpora from the cities of Mocajuba, Óbidos, Santarém and Breves.

On the map below there is the localization of all the inquest points that are covered for this project in Para State nowadays.

Figure 1: Map 01 – The localities attained for The AMPER-North Project. It was adapted from Cruz (2012: 205)

Cassique (2006 apud Cruz, 2012) presents a new dialectal division of Pará State from Silva Neto (1957) that has been considered by UFPA’s researchers linked to AMPER-POR Project, so it has been used as base to the choice of the target localities of its project.

According to this dialectal division, the selected localities for this project’s investigation belong to the...
regional PB of Pará State (cf. Zone 1 of map 1). Bragança is the only one that belongs to another dialect called bragantino (cf. Zone 2 of map 1).

The PB spoken in Pará State is considered by Silva Neto (1957) how being of canua cheia de cucus de pupa a prua. It has the main dialectal mark that is a rising on the back vowels on the stressed syllable (Rodrigues, 2005).

For this reason, the Prosodic Multimedia Atlas of Northern PB will register exactly the prosodic variations of the PB spoken in Pará State and supply a sociolinguistic configuration on the prosodic level of this variety of PB.

On the first version of this project it was possible to move forward on the formation of corpora. Unfortunately, it was not possible to explore them yet, except for these two corpora that have been explored: Cametá city (Santo & Cruz, 2011; Santo, 2011) and Belém city (Cruz & Brito, 2011).

3. Methodological procedures that are adopted on the formation of the corpora

On this study they were adopted all the methodological procedures determined by the general coordination of the AMPER Project.

As one of the goals of the AMPER Project comprises a contrastive analysis of the studied dialects, the corpus was recorded for the varieties of Brazilian Portuguese. It is made up of six replicates of sixty-six sentences of the corpus-based AMPER project for the Portuguese language. Each constituent of the phrases have a corresponding image, since it is not allowed any contact of the speakers with the written sentences. Therefore, during the fieldwork, there is the visual representation of the sentences which means that slides are shown to informants as a way of graphic stimuli for the production of 369 sentences to be generated. The set of sentences that form the corpus of the project AMPER follows previously established phonetic and syntactic criteria.

Since the vowels have the most relevant information regarding the prosodic curve and taking into account the characteristics of the stress structure of the Portuguese, there have been chosen words that represent the different stress structures (oxytone2, paroxytone3 and proparoxytone4) in various positions on the sentence5.

The sentences were syntactically set up, so as for the present Subject – Verb – Complement (SVC). In relation to the intonation, they were designed to accommodate the neuter modes, affirmatives, declaratives and global interrogatives. Therefore, the sentences utilized on the recordings are of the type of SVC and its extensions to include prepositional phrases. As for the syntactic structure, all sentences have only: 1) three characters (Renato, pâssaro and bisavô); 2) three adjectival phrases (nadador, bêbado and pateta); 3) three prepositional phrases of place (‘de Mônaco’, ‘de Venezuela’ and ‘de Salvador’); 4) a single verb (gostar).

During the collection of the data, six repetitions of a set of phrases are asked to each speaker of a set of phrases in the corpus (in random order), being selected for acoustic analysis the top three repetitions, in order to establish the meaning of three acoustic parameters: duration, fundamental frequency (F0) and intensity.

As it was determined by the general project, for the selection of informants were taken into consideration the following criteria: 1) age (above 30 years old); 2) schoolar level (elementary school, high-school and college); and 3) residence time in the town (only local indigenous people). Based on these criteria, six informants were selected, three males and three females, who participated on the data collection. It is, therefore, a stratified sample. Each informant has received a code which contains information on his/her profile. On the Table 1 below, we can see the codification adopted by AMPER-North Project.

<table>
<thead>
<tr>
<th>Portuguese Dialect</th>
<th>Brazilian Region</th>
<th>City</th>
<th>School Level</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>B (Brazil)</td>
<td>E (North)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0 Belém</td>
<td>1</td>
<td>The Lowest school level</td>
<td>female</td>
</tr>
<tr>
<td></td>
<td>2 Abetetuba</td>
<td>3</td>
<td>high-school</td>
<td>female</td>
</tr>
<tr>
<td></td>
<td>3 Bragança</td>
<td>4</td>
<td></td>
<td>male</td>
</tr>
<tr>
<td></td>
<td>4 Curaninho</td>
<td>5</td>
<td>College</td>
<td>female</td>
</tr>
<tr>
<td></td>
<td>5 Cametá – urban region</td>
<td>6</td>
<td></td>
<td>male</td>
</tr>
<tr>
<td>F (North)</td>
<td>4 Mocajuba – rural region</td>
<td>2</td>
<td></td>
<td>female</td>
</tr>
<tr>
<td></td>
<td>5 Mocajuba – urban region</td>
<td>5</td>
<td></td>
<td>male</td>
</tr>
<tr>
<td></td>
<td>7 Outeiro – Belém's island</td>
<td>6</td>
<td></td>
<td>female</td>
</tr>
<tr>
<td></td>
<td>8 Mosqueiro – Belém island</td>
<td>7</td>
<td></td>
<td>female</td>
</tr>
<tr>
<td></td>
<td>9 Itaico</td>
<td>8</td>
<td></td>
<td>male</td>
</tr>
</tbody>
</table>

Table 1: Codification of the speakers adopted by AMPER-North Project

In total six sound files were obtained by investigating the localities. The rate of sample of the sign is 44.100 Hz, 16 bits, mono. All the data collection was made in the informant’s own house.

4. Characterization of the corpora

Therefore the AMPER-North Project’s corpus is composed of 198 sentences, in total 1.188 sentences by informant, which contains samples of the linguistic
varieties that are spoken in Belém, Cametá, Baião and Bragança. Below Table 2 contains the size in hours of recording of each formed corpus.

<table>
<thead>
<tr>
<th>City</th>
<th>Code</th>
<th>Duration corpus size</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Belém</td>
<td>BE0</td>
<td>3h 36 min. 25 sec.</td>
<td>Brito (in progress)</td>
</tr>
<tr>
<td>Baião</td>
<td>BF9</td>
<td>6h 33 min. 15 sec.</td>
<td>Lemos (in progress)</td>
</tr>
<tr>
<td>Bragança</td>
<td>BE3</td>
<td>2h 21 min. 21 sec.</td>
<td>Castilho (2009)</td>
</tr>
<tr>
<td>Cametá</td>
<td>BE5</td>
<td>3h 16 min. 39 sec.</td>
<td>Santo (2011)</td>
</tr>
</tbody>
</table>

Table 2: Total size of formed corpora of the AMPER-North Project in hours of recording

The Project in itself organizes the formed corpora, but the availability on line of the corpus is responsibility of the general coordination of AMPER-POR Project.

This project already supplied the list above of the varieties of Belém (BE0) and Cametá (BE5) to the general coordination and, therefore, the data on this list are already in the site of AMPER-POR Project⁶.

5. Tendencies of spoken Portuguese in the northern of Brazil: preliminary analysis

Until the present day, the obtained results refer to the physical parameters - intensity, duration and F0 – in relation to the kind of Portuguese stress and to the syntactic aspects controlled by AMPER Project on the construction of its corpus. At the moment, two Brazilian Portuguese varieties spoken in Amazon were analysed: Cametá and Belém.

The preliminary analysis made with the data (Santo & Cruz, 2011; Cruz & Brito, 2011) indicated that, in general, the F0, duration and intensity measures complement one another to establish a distinction between statement and interrogative in Brazilian Portuguese spoken in Cametá (PA).

We can equally to state that the important variations of the three controlled acoustic parameters, that establishes the difference between the two modalities, occur preferentially on the stressed syllable of the nuclear element of the phrases and/or on the last stressed syllable of the statement.

It is important to consider the meaning of F0 variations. Note that the more important variations occur just on the stressed syllable of the statement. We have showed above, on the two pairs of phrases - Figures 2 and 3 -, that the nucleus of the sintagma occupies firstly the position of subject of the phrase and after occupies the last position of the verbal complement to verify that the stressed syllable has the movement of variation of F0 which is more important on the sentence.

Figure 2: Comparison between the F0 variation meaning of the sentence twp – O Renato gosta do pássaro – on both the modalities – declarative (full line) and interrogative (dashed line), spoken by a female speaker with low educational level from Belém - BE0 (black) – and another speaker of the same social profile from Cametá dialect – BE5 (white)

Figure 3: Comparison between the F0 variation meaning of the sentence twp - O pássaro gosta do Renato - on both the modalities – declarative (full line) and interrogative (dashed line), spoken by a female speaker with low educational level from Belém - BE0 (black) – and another speaker of the same social profile from Cametá dialect – BE5 (white)

Figure 4: Comparison between the meaning of the duration (ms) on the sentence twp - O Renato gosta do pássaro – on both the modalities – declarative and interrogative - spoken by a female speaker with low educational level from Belém – BE0 and another speaker of the same social profile from Cametá dialect – BE5

---

⁶ http://pfonetica.web.ua.pt/AMPER-POR.htm
The parameter of duration (ms) seems to act like a complement of the variations of F0 on the distinction of the two modalities that were analyzed, just as it is possible to observe on figures 4 e 5.

While the parameters of F0 and duration seems to complement themselves on the characterization of both modalities declarative and interrogative on the varieties of the North of Brazil, the intensity seems not to be a significant physical parameter on the distinction of the two modalities in question, like we can note on the graphic of the Figures 6 and 7 below.

It is important to point out once again that the last stressed syllable of the phrase is the one that registers the more important movement of the distinction between both modalities. For this reason, it has been our base hypothesis to be verified on the corpora of the project here outline.

Figure 7: Comparison between the meaning of dB on the sentence pwt - O pássaro gosta do Renato - on both the modalities – declarative and interrogative - spoken from a female speaker with low education level from Belém – BE0 - and another speaker of the same profile of the Cametá dialect – BE5

6. Conclusion

The previous version of this project, whose period of execution includes since March 2009 until February 2012, has composed a corpora for the following Atlas: a) Belém – BE0 – (Cruz & Brito, 2011); b) Bragança – BE3 – (Castilho, 2009); c) Cametá – BE5 – (Santo & Cruz, 2011; Santo, 2011); and d) Baião – BF9 – (Lemos, in progress). Currently there is a planned fieldwork to the formation of other three corpora: f) Abaetetuba (Remédios, in progress); g) Belém islands (Guimarães, in progress); and h) Marajó island (Freitas, in progress). The Project has the exploration and the acoustic analysis of the corpora from Belém (Cruz & Brito, 2011) and from Cametá (Santo & Cruz, 2011; Santo, 2011).

7. References


Cruz, R. (2012). Alteamento vocálico das médias pretônicas no português falado na Amazônia


Desafios da formação de corpus nas zonas de migração do Norte do Brasil

Regina CRUZ\textsuperscript{1,2}, Edson GOMES\textsuperscript{1,3}, Jany Eric FERREIRA\textsuperscript{1,3}, Soelis MENDES\textsuperscript{1}, Emanuel FONTEL\textsuperscript{1}
\textsuperscript{1}UFPA; \textsuperscript{2}CNPq; \textsuperscript{3}Aluno de Mestrado
Av. Augusto Correa, s/n – Campus do Guamá – Belém (PA) – 66075-900
regina@ufpa.br, edsongomes@ufpa.br, eric_qf@yahoo.com.br, soelis@ufpa.br, emanelfontel@hotmail.com

Abstract

This work aims show how they formed sociolinguistic corpora for study of Brazilian Portuguese spoken in migration areas of Northern Brazil. We approach mainly difficulties happened during fieldwork of two UFPA’s teams: i) Vozes da Amazônia project's team linked to PROBRAVO and b) ALIPA project's team linked to ALIB. Both projects aim to identify and map Amazon dialects. We show here the whole methodology of these projects: aims, nature of study, research context, speaker selection, collection of data, composition of corpora and researchers report of their experience. Between some difficulties we found: i) to meet speakers whose profile is right for each project; ii) non-availability of speakers in order to collaborate of collection of data; iii) the rejection of the people to face the recorder; iv) the fact the interviewer is not of location (Mendes, in progress). In the other hand, we noted that when the researcher lives in target city and the project uses the same methodological criteria of Bortoni-Ricardo (1985) for formation of corpus, this frame changes and researcher obtains a strong collaboration of the speakers (Ferreira, in progress).

Keywords: sociolinguistic corpora; Amazon Brazilian Portuguese; interdialectal contact.

1. Introdução

O presente trabalho tem como objetivo principal demonstrar como se afigura árduo o processo de formação de corpus sociolinguísticos em zonas de grande fluxo de migração na Amazônia paraense. Darse-á principal enfoque às dificuldades enfrentadas no trabalho de campo realizado pela equipe do Projeto Vozes da Amazônia sediado na UFPA que, por sua vez, é vinculado diretamente ao Diretório de Pesquisa Nacional PROBRAVO\textsuperscript{1}. Três regiões foram selecionadas para uma nova fase de investigação desse projeto: Marabá (Mendes, em andamento), Aurora do Pará (Ferreira, em andamento) e Breves. Duas outras localidades estão previstas: Breu Branco e Paraapebas. No âmbito do ALIPA\textsuperscript{2}, outro projeto de formação de corpora também sediado na UFPA, identificam-se as mesmas dificuldades da equipe do Vozes da Amazônia. Por essa razão, um cotejo entre os dois projetos é aqui estabelecido. Para tanto, esposam-se considerações em torno dos objetivos e da metodologia empreendida em cada programa de investigação, além de breves relatos que ressaltam a atuação de cada pesquisador na respectiva comunidade linguística em que atua.

2. Projetos sociolinguísticos do Norte do Brasil

Em um Estado com dimensões continentais como é o caso do Pará, era de se esperar que houvesse forte variação no falar da população, principalmente pelo fato de essa constituição populacional ter-se dado por diferentes processos de ocupação territorial. Nesta seção apresentaremos os projetos Vozes da Amazônia e ALIPA, os quais têm o propósito de identificar e mapear os dialetos paraenses. Os projetos investigam os contatos interdialetais, resultado de processo de migração para a Amazônia paraense. No tocante a esse aspecto, a pesquisa em 6 (seis) pontos da mesorregião sudeste do Pará, empreendida por Gomes (em andamento), no projeto ALiPA, é o trabalho que trata mais especificamente da influência de outras regiões do Brasil nesta referida mesorregião.

2.1 Projeto Vozes da Amazônia

A versão atual do projeto Vozes da Amazônia prioriza uma investigação da identidade sociodiscursiva do amazônico nas regiões onde se atesta contato interdialetal decorrente de fluxo migratório intenso, motivado por projetos econômicos na região Amazônica, o que inclui o tratamento de aspectos culturais, sociais, históricos, político e ideológicos. Mapear a situação sociolinguística diagnosticada por Cruz (2012) relativamente à Amazônia paraense é o objetivo central do Vozes, em outras palavras o projeto busca identificar a influência de fatores extralinguísticos e identitários na configuração dos dialetos da Amazônia paraense, considerando o cenário sociohistórico da região e o fluxo migratório ali registrado. O projeto encontra-se vinculado a dois campi da UFPA - o de Belém e o de Marabá - e conta com a infraestrutura destes para a execução de suas atividades. A equipe atual, responsável pela condução das investigações, é composta por 2 (dois) alunos de Mestrado, 2 (dois) bolsistas de Iniciação Científica e 3 (três) pesquisadores titulados, todos com vínculo direto com a UFPA, além da coordenadora geral.

2.2 Projeto ALIPA

O ALIPA é um projeto de pesquisa ligado ao laboratório de linguagem da UFPA. Esse projeto tem por objetivo a construção do Atlas Geo-Sociolinguístico do Pará. Neste sentido desenvolve estudos cuja finalidade é identificar, analisar e mapear a variação linguística do português falado no Estado do Pará, integrando a dimensão social, que permitirá melhor compreensão dos mecanismos internos envolvidos na variação, especificamente, fonética, morfossintática e semântico-lexical. O projeto

\textsuperscript{1} relin.letras.ufmg.br/probravo.
\textsuperscript{2} http://www.ufpa.br/alipa.
utiliza metodologia do ALIB\textsuperscript{3}. Para sua a execução, foram selecionados no Estado 50 (cinquenta) pontos de inquérito. Desses cinquenta pontos, mais de quarenta já foram coletados, restando alguns pontos na mesorregião sudeste, dentre esses enquadram-se os seis pontos de inquérito de (Gomes, em andamento).

3. Regiões mapeadas
Como o objetivo de ambos os projetos – Vozes da Amazônia e ALiPA – é compor um panorama histórico, antropológico e social do Pará, assim como identificar fatores sociais favorecedores da variação dialetal do português da Amazônia paraense, falado nas regiões de forte migração interna, faz-se necessário relacionar aspectos de variação inter e intradialetal. Por essa razão, à medida que se caracteriza sociolinguisticamente o português falado em Marabá, Aurora do Pará, Tucurui, Curionópolis, por exemplo, obtém-se o panorama geral das zonas de migração do Estado. Na nova fase do projeto Vozes no Estado do Pará, os municípios de Breves, Aurora do Pará e Marabá, em destaque verde no mapa (1), foram selecionados para a realização da pesquisa, tanto nas suas zonas rurais quanto urbanas. No caso do Projeto ALiPA, há um número maior de regiões contempladas, entretanto, o presente trabalho trata particularmente das localidades de Tucurú, Itupiranga, São João do Araguaia, Curionópolis, Santana do Araguaia e São Felix do Xingu, indicadas de azul, no referido mapa.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{mapa.png}
\caption{Mapa 1: Indicações das localidades pesquisadas}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figura1.png}
\caption{Figura 1: Relação estabelecida entre as partes componentes do modelo utilizado por Bortoni-Ricardo (1985)}
\end{figure}

4. Procedimentos metodológicos adotados por projetos
Os projetos da UFPA aqui descritos, apesar de terem como ponto em comum o tipo de região investigada, no caso as zonas de forte fluxo migratório no Estado, a metodologia adotada por ambos na formação de seus corpora é bem diferente como veremos nesta seção.

4.1 Como trabalhava o Vozes da Amazônia?
O Vozes da Amazônia parte do conceito de Redes sociais como um conjunto de ligações que se estabelecem entre indivíduos. Segundo Bortoni-Ricardo (1985), nesse tipo de estudo o foco da investigação está na caracterização das relações entre os indivíduos, através da qual se pode explicar seus comportamentos, inclusive comportamentos linguísticos. Outro conceito importante é o de grupo de referência, que serve de alavanca à construção da identidade do indivíduo, o qual tenta modelar seu discurso segundo o daqueles que atende às suas expectativas psicossociais e com os quais busca identificação. A figura 1, abaixo, ilustra as relações que podem explicar o comportamento linguístico, em conformidade com o que propõe a referida autora.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figura1.png}
\caption{Figura 1: Relação estabelecida entre as partes componentes do modelo utilizado por Bortoni-Ricardo (1985)}
\end{figure}

A composição do corpus ocorre a partir de dois grupos de informantes: ancoragem e controle. O grupo de ancoragem possui 24 informantes (12 de cada sexo) e o de controle, 12 informantes (6 de cada sexo) que devem necessariamente ter algum vínculo de parentesco com membros do grupo de ancoragem, como filhos, netos ou sobrinhos. Todos os informantes são distribuídos em três faixas etárias: a) de 15 a 26 anos; b) de 30 a 46 anos e; c) acima de 50 anos. A coleta de dados é realizada por meio de narrativas de experiência pessoal. O trabalho de Mendes (em andamento) atesta que esse tipo de procedimento metodológico tem sido eficaz nos dois grupos de informantes, para os quais se pergunta sobre a origem de cada um e sobre a percepção que cada informante detém da cidade antes de ele lá ter-se instalado etc. Além desses aspectos, firma-se atenção a todas as orientações da técnica apresentada por Tarallo (1988). Registre-se, ainda, que os dados estão sendo coletados por meio de gravadores digitais. Uma vez o trabalho de campo concluído, o tratamento dos dados seguirá todas as etapas previstas em um estudo sociolinguístico, a saber: (i) transcrição dos dados nos

\footnote{http://twiki.ufba.br/twiki/bin/view/Alib/MetodologiaGeral}
moldes da análise da conversação (Castilho, 2003); (ii) triagem dos grupos de força (Câmara Jr., 1969); (iii) transcrição fonética dos vocábulos que contenham marcas dialetais alvo, utilizando-se o alfabeto SAMPA; (iv) codificação dos dados e; (v) tratamento quantitativo VARBRUL.

4.2 Como trabalha o ALiPA?

O projeto ALiPA contempla o número de 50 (cinquenta) localidades, distribuídas por seis microrregiões do Pará, levando-se em consideração a extensão de cada região, os aspectos demográficos, culturais, históricos e a natureza do processo de povoamento da área. Para compor o corpus, foram selecionados 4 (quatro) informantes por localidades: dois do sexo feminino e dois do masculino, distribuídos nas faixas etárias de 18 a 30 anos e 40 a 70 anos. Eles devem ser filhos da localidade pesquisada, assim como os pais; devem ter, no máximo, a 4ª série do fundamental e exercer profissões que evitem mobilidade. A coleta de dados é realizada com uso dos questionários fonético-fonológico, morfossintático e semântico-lexical. No estudo de Gomes (em andamento), está sendo aplicado apenas o questionário semântico-lexical. Os dados estão sendo coletados em equipamentos sonoros, como gravador digital, gravados em CD e em outros equipamentos de informática para posterior tratamento. Vencida essa etapa, os dados serão transcritos grafematicamente e transferidos para as cartas lexicais. Também vem sendo usado o recurso fotográfico como meio de registrar através de imagens o homem e o espaço em que habita.

5. Caracterização dos corpora formados

Até o presente momento, o corpus formado conta com amostra de 14 (quatorze) informantes (sendo oito do grupo de ancoragem e seis do grupo de controle) da variedade linguística de Marabá (Mendes, em andamento), 18 (dezoito) informantes do grupo de ancoragem e 8 (oito) do grupo de controle da variedade linguística de Aurora do Pará, localidade que está sendo pesquisada por Ferreira (em andamento). No total, há 36 (trinta e seis) informantes por variedade. Ambos trabalhos fazem parte do Vozes da Amazônia. Para a pesquisa de Ferreira (em andamento), na fase de trabalho de campo, foram estabelecidas visitas prévias aos informantes, para um primeiro contato, o que permitiu prematuramente o surgimento de um certo vínculo de intimidade com os informantes. Esse grau de intimidade favoreceu para criar um clima, o mais descontraído possível, entre os participantes da pesquisa, o que é essencial para uma boa coleta de dados. O fato de contribuírem para o trabalho de alguém conhecido deixava os informantes bastante alegres e descontraídos, o que incentivou o estranhamento causado pelo aparato técnico, presente no momento da entrevista, que ocorreram na própria casa dos informantes. Há de se destacar, contudo, que, se por um lado o grau de intimidade que se estabeleceu entre entrevistador e informantes, representou, naquele momento, facilidade, a falta de informantes e de instrumentos técnicos suficientes para o grupo de pesquisa têm sido entraves para a composição dos corpora. Ferreira (em andamento) afirma que há enorme dificuldade para se identificar informante tanto do sexo masculino quanto do feminino na faixa etária de 26 a 46 anos. A maioria dos migrantes possui mais de 50 anos e instalou-se no município nas décadas de 60 e 70. Os de faixa etária mais jovem apresentam-se em menor quantidade, o que dificulta o trabalho para encontrá-los. Pesquisas dessa natureza, no qual se tem critérios para a seleção de informantes, nem sempre são fáceis de serem executados posto que o andamento da pesquisa depende do total de informantes necessários a sua realização. Isso tem ocasionado o atraso de todo o trabalho de campo de Ferreira (em andamento). O fato de não se ter uma quantidade suficiente de gravadores digitais para coletar os dados também constitui outro entrave. Os poucos investimentos injetados em trabalhos dessa natureza afetam diretamente em sua realização. Gomes (em andamento) até o presente momento conta com um corpus contendo amostra de 20 (vinte) informantes, sendo 10 (dez) homens e 10 (dez) mulheres, de um total de 24 (vinte e quatro) informantes. 10 (dez) informantes são da faixa etária de 18 a 30 anos e 10 (dez) são da faixa etária de 40 a 70 anos. De cada localidade - Santana do Araguaia, São Félix do Xingu, Tucuruí, Caruã, São João do Araguai e São Fructuoso - foram entrevistados 4 (quatro) informantes, faltando apenas 4 (quatro) informantes de Itupiranga. Para coletar os dados, Gomes (em andamento) teve que se deslocar em dois momentos: em julho de 2011, para Santana do Araguai, São Félix do Xingu e Tucuruí; em fevereiro de 2012, para Caruã, São João do Araguai e São Fructuoso. A coleta de dados ocorreu a partir de entrevistas realizadas, na maioria das vezes, nas casas dos informantes, o que não foi muito bom, devido às interferências de curiosos que, em alguns momentos, respondiam às perguntas. Outras, contudo, foram realizadas fora da casa, às margens do Rio Xingu, por exemplo, o que facilitou o trabalho, mas houve momentos difíceis, em que foi preciso fazer a entrevista sob o sol ou embaixo de chuvisco, por falta de local apropriado e para não perder o informante. Pelo fato de as entrevistas terem sido realizadas em localidades distantes, no sudeste do Pará, foi preciso montar um planejamento de deslocamento. Mesmo os informantes sendo pessoas desconhecidas do entrevistador, houve sucesso no trabalho, porque, em todas as cinco localidades, foram encontradas pessoas dispostas a ajudar na coleta de dados. O fato de alguém ter ido de longe e em condições desfavoráveis para campo, sensibilizava os informantes, os deixava satisfeitos e os tornava mais propensos a contribuir, dando informações dos seus respectivos lugares, muitas vezes esquecidos. A maior dificuldade por que se passou foi conseguir pessoas que se encaixassem nas exigências do projeto ALIIPA. As 20 (vinte) gravações foram realizadas em gravador digital Olympus Linear PCM Recorder LS-10. A próxima etapa constitui-se em
trabalhar os dados para verificação da variação que ocorre dentro da mesorregião objeto da pesquisa, e desta em relação às outras mesorregiões do Estado do Pará, para se obter um retrato, o mais fiel possível, do falar paraense. A seguir figura (5.1) apresenta a síntese do total de informantes por pesquisa, num total de 60.

![Informantes por pesquisa](image)

Figura 5.1: Síntese do Total de Informantes por pesquisa

6. **Dificuldades impostas pela realidade amazônica das zonas de migração**

Tanto Mendes (em andamento) quanto Fagundes (em andamento), que também faz parte da equipe do Vózes da Amazônia, estão tendo grande dificuldade na obtenção dos dados necessários. Uma das dificuldades encontradas para realização da pesquisa está em encontrar pessoas que se encaixem no perfil do projeto e a não disponibilidade dos falantes localizados para participar da pesquisa. Mesmo se tendo a preocupação de deixar o informante o mais à vontade possível com a presença da equipe e com a do gravador, a recusa da parte de algumas pessoas dá-se, na maioria das vezes, sem um motivo aparente. Invariavelmente, os que se recusam fazem-no simplesmente com a afirmação de que não aceitam participar e, diante disso, não são feitas mais investidas, pois é necessário que o informante sinta-se à vontade. Outras vezes, o falante, em função da presença do gravador, sente-se inibido e se recusa a participar. Além disso, há também a incidência daqueles que marcam a entrevista e não comparecem ao encontro. Diante dessas situações, muitas vezes, tenta-se marcar nova coleta, no entanto, isso não garante a presença do informante nessa nova oportunidade. A situação é mais grave, ainda, quando o entrevistador, além de não ser morador nativo da localidade, utiliza critérios sociolinguísticos para a formação de amostras mais adequados para estudos variacionistas clássicos, como, por exemplo, o critério de selecionar apenas informantes nativos da comunidade pesquisada ou que tenham ido morar para lá ainda criança, como é o caso de Gomes (em andamento), para quem os aspectos históricos e sociais da localidade investigada não são importantes, tendo em vista os objetivos da pesquisa que empreende atualmente. O fato de a pesquisa de Gomes (em andamento) estar localizada na mesorregião Sudeste do Pará, dificulta a coleta dos dados, porque alguns critérios adotados pelo projeto ALiPA, como a exigência de informantes nascidos na localidade, vão de encontro ao histórico da região, onde a população é constituída, em sua maioria, por migrantes de outras localidades do país. Enquanto nas outras mesorregiões a população rural é constituída, em sua maioria, por habitantes nascidos na localidade e portadores de característica “cabocla”, na mesorregião Sudeste (Sul do Pará) verifica-se exatamente o contrário: nas zonas rurais muitos habitantes moram nos Projetos de Assentamento (PAs), o que faz com que, muitas vezes, seja mais fácil encontrar um morador nascido na cidade que, embora desenvolva suas atividades na zona rural, nasceu em zona de cidade, ou seja, zona urbana, tanto que o deslocamento da população dessa região se dá mais para os Estados do Centro-Oeste, como Tocantins, Goiás, Brasília, do que para Belém, capital do Pará. Surpreendentemente, quanto à abordagem dos informantes, Gomes (em andamento) sentiu pouca dificuldade, pois quase sempre as pessoas estavam dispostas a colaborar na pesquisa. Sua maior dificuldade foi, sem dúvida, identificar informantes com o perfil exigido. Alguns funcionários da EMATER foram peças-chave para a localização de informantes, principalmente em Santana do Araguaia e São Félix do Xingu. Outro fato surpreendente foi o engajamento de outras pessoas que acabaram colaborando diretamente, ao terem se sensibilizado com o descolamento do pesquisador e com os objetivos de seu trabalho de pesquisa. Tais fatos acabaram aterciapor-se pontos positivos na implementação dos trabalhos e na colaboração dos moradores locais. Por outro lado, constatou-se que, quando o entrevistador é um morador da localidade alvo e se utiliza dos parâmetros propostos por Bortoni-Ricardo (1985) para a formação do corpus, este quadro de dificuldades não é verificado e o pesquisador consegue obter forte colaboração dos informantes, é o que relata Ferreira (em andamento). De qualquer forma, a experiência de estar realizando este tipo de pesquisa tem sido bastante rica, não só pelo fato de permitir perceber a grande recorrência do fenômeno analisado, no caso de Ferreira (em andamento), as vogais médias pretônicas, sem o que a pesquisa restaria inviável, mas também, e sobretudo, pelas possibilidades advindas do contato estabelecido com as pessoas nativas da região e pela interação com elas estabelecida, tanto na condição de amigo, de conhecido, quanto na de pesquisador que observa e analisa os ricos aspectos que envolve o fenômeno da variação linguística, evidenciada na fala natural. Nesse sentido, constata-se, ainda mais, a pertinência da ideia comum de que a interação independe de formalismos linguísticos. Fortalece o compreensão dos modos como o trabalho de campo imprime melhor reflexão ao trabalho de coleta e análise, além de permitir a verificação dos aspectos que, nessas condições, dão ensejo a ocorrência de fala natural, tão ansia pelos pesquisadores da sociolinguística. É preciso ressaltar que Mendes (em andamento), em sua pesquisa na comunidade linguística de Marabá, embora com dificuldades na seleção dos informantes e, em especial, na realização das entrevistas, constatou que
uma entrevistada ficou tão envolvida emocionalmente com a condução de sua narrativa, que, quando o tempo estipulado para todas as entrevistas chegou ao fim, ou seja, 15 minutos, a informante pediu para continuar contando casos de sua sofrida vida para Marabá. Outro ponto de destaque, que certamente poderá auxiliar na condução de pesquisas em desenvolvimento, refere-se a uma participação no Congresso Internacional de Linguística Histórica, em homenagem ao Prof. Dr. Ataliba Castilho, realizado na USP, no mês de fevereiro de 2012, quando Mendes, ao relatar à Prof. Dra. Odete Pereira da Silva Menon, (UFPR) a dificuldade vivenciada na realização das entrevistas em Marabá, foi orientada a contar com a autoridade de pastores evangélicos no contato com informantes. De fato, essa estratégia tem contribuído para selecionar novos informantes, permitindo, desse modo, vislumbrar, logo em breve, a consecução das entrevistas com os 18 (dezoito) falantes restantes.

7. Conclusão

Nos dias atuais, as mudanças ocorrem muito rapidamente e, em consequência, as transformações na língua também. Essa realidade tem impulsionado estudos linguísticos que visam a recuperar e/ou registrar os falares de diversas comunidades linguísticas, porquanto, assim procedendo, registram-se não só fenômenos linguísticos observados, mas também as memórias linguística e discursiva da comunidade da região estudada. Ao identificar e mapear dialetos nas regiões de migração do norte do Brasil, os projetos ALiPA e VOZES da AMAZÔNIA cumprem seu papel social, pelos motivos acima expostos. No entanto, essa tarefa nem sempre é tão simples e possível de ser concretizada, em virtude das dificuldades e desafios impostos pela pesquisa de campo. Desse modo, as dificuldades que envolvem a composição dos corporas, em regiões de migração, no norte do Brasil, apresentam-se tanto no projeto Vozes da Amazônia quanto no projeto ALiPA. Dentre as que merecem destaque, elencamos a dificuldade de encontrar pessoas que se encaixem no perfil dos projetos e a não disponibilidade dos falantes locais para participar da pesquisa; a recusa das pessoas perante o gravador e, por vezes, a escassez de equipamentos e ferramentas adequadas. O fato de o entrevistador não ser da localidade também dificulta a coleta de dados, havendo de se sobrepor a imperiosa e nobre tarefa do pesquisador de desmotivar aqueles que se interessam ou intencionam trilhar as veredas da pesquisa linguística. Do contrário, esperamos que as considerações aqui apresentadas sirvam para demonstrativo de que as dificuldades, sejam de cunho metodológico, sejam de outro caráter, não devem ser sobrepor a imperiosa e nobre tarefa do pesquisador de descrever o funcionamento da língua em todos os seus matizes e possibilidades.

8. Referências


Compiling a Multilingual Spoken Corpus
Ekaterina LAPSHINOVA-KOLTUNSKI, Kerstin KUNZ, Marilisa AMOIA
Saarland University
Universität Campus 66123 Saarbrücken Germany
e.lapshinova, k.kunz, m.amoia@mx.uni-saarland.de

Abstract
The present paper describes the compilation of the spoken part of an English-German corpus, which has been created for the investigation of cohesion. The corpus is one of the few existing resources supporting contrastive studies of cohesion and, to our knowledge, the only one permitting a contrastive analysis of spoken registers in the two languages. In addition, our corpus data offer further research potentials for contrastive linguistics and translation studies as well as for numerous NLP research areas.

Keywords: corpus compilation, spoken corpus, multilingual corpus, corpus annotation, cohesion.

1. Introduction
The present paper describes the compilation of the spoken part of an English-German corpus, which has been created for the investigation of cohesion. The corpus is one of the few existing resources supporting contrastive studies of cohesion and, to our knowledge, the only one permitting a contrastive analysis of spoken registers in the two languages. In addition, our corpus data offer further research potentials for contrastive linguistics and translation studies as well as for numerous NLP research areas.

1.1 Aims
The main objective of the present paper is to compile the spoken part of a multilingual corpus to investigate cohesion in German and English. Our long-term linguistic research interest is in the analysis of cohesive resources provided by both language systems and their instantiations in texts. More precisely, we are concerned with the exploration of contrasts in form, frequency and function of cohesive devices and meaning relations established to other textual elements. We aim to analyse these phenomena across and between languages, registers and modes.

1.2 Motivation
Comprehensive accounts of cohesion are only existent from a largely systemic and monolingual perspective, see e.g. (Halliday & Hasan, 1976; Brown & Yule, 1983; Schubert, 2008 and Esser, 2009) for English, and (De Beaugrande & Dressler, 1981; Vater, 2005; Brinker, 2005) for German. Empirical analyses (both monolingual and contrastive) in the area of cohesion mainly deal with individual cohesive devices, cf. (Bosch et al., 2007) or (Gundel et al., 2004). Empirical analyses of cohesion in spoken discourse exist for German, e.g. (Ahrenholz, 2007) and English, e.g. (Gundel et al., 2004 and 2005; Eckert & Strube 2001). These however, are limited to the investigation of individual phenomena, and mostly examine personal pronouns or demonstratives. To our knowledge, there is only one contrastive empirical analysis by (Schreiber, 1999) comparing English and German. It includes a relatively broad range of cohesive phenomena, however it uses excerpts of French and German corpora to illustrate particular phenomena rather than presenting a contrastive interpretation of findings from a statistical analysis.

These studies seem to suggest that particular cohesive devices exhibit a tendency to occur either in registers of spoken language only or with a much higher frequency than in written discourse, see e.g. (Schreiber, 1992; Ahrenholz, 2007). Our preliminary extractions from registers of written language 1 underpin these observations. For instance, they show that occurrences of the German demonstrative pronouns der, die, das and particular constructions of substitution are rarely traced in typical registers of written language and appear with a much higher frequency in those written registers that approximate spoken language, such as fiction or political speeches2. In addition, dialogic sequences of our fiction subcorpus point to instantiations of cohesive ellipsis which seem to be restricted to spoken discourse. These first findings call for a corpus which allows to integrate differences between written and spoken registers so as to establish a comprehensive model of cohesion in English and German. To our knowledge, there are no corpus resources to support our research goal. The existing ones are either monolingual, e.g. ICE, cf. (Greenbaum, 1996) for English or “Deutsch heute”, cf. (Brinckmann, 2008) German, or compiled for special purposes, e.g. SCOTS corpus, cf. (Anderson, 2007) or Verbmobil, (Hinrichs et al., 2000). Some of them also contain non-native data, e.g. ICLE described in (Granger, 2008) and LINDSEI, cf. (Gilquin et al., 2010).

2. Theoretical Background
There are substantial gaps in the area of text-based contrastive modeling for the two languages under analysis, especially text-based empirical accounts of mechanisms of textuality are absent. System-based text/discourse grammars commonly deal with specific questions of textuality only. While the literature in English mainly

1 cf. (Kunz et al., 2009; Klein, 2007 and Birster, 2007).
2 The extractions were done on the CroCo corpus, cf. (Neumann, 2005)
focuses on linguistic resources for establishing textuality, e.g. (Halliday & Hasan, 1976; Brown & Yule, 1983; de Beaugrande & Dressler, 1981), the German literature frequently takes as its starting point general pragmatic, cognitive and semantic principles of coherence, which are reflected in linguistic phenomena, cf. (Linke et al., 2001; Brinker, 2005; Vater, 2001). These methodological differences lead to noticeable differences in the range of phenomena considered. In general, monolingual text/discourse treatments inform us about the coherence-building systems of a language and are structured by type and/or function of the system (e.g. (co-) reference, conjunctive relation, lexical/semantic relations, etc). We define cohesive resources (devices) as a set of lexico-grammatical items that function as resources allowing to transcend the boundaries of the clause. For our classification of general categories, we follow the one by (Halliday & Hasan, 1976), according to which cohesion includes five categories: reference, substitution, ellipsis, conjunctive relations, lexical cohesion.

3. Corpus Compilation

3.1 Data Collection

Our multilingual spoken corpus contains two registers: interview and academic speech. These registers are added to the eight registers of written language of the already existing corpus, cf. (Kunz & Lapshinova, 2011): popular-scientific texts, tourism leaflets, political essays, corporal communication, instruction manuals and websites, prepared speeches, fictional texts. Especially the latter two registers are considered to lie at the borderline between written and spoken discourse. In order to create the German-English spoken corpus, we extract parts of already existing speech corpora and collect our own data, cf. Table 1.

<table>
<thead>
<tr>
<th>subcorpora</th>
<th>German (GO)</th>
<th>English (EO)</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTERVIEW</td>
<td>BACKBONE-DE</td>
<td>ELISA</td>
</tr>
<tr>
<td>ACADEMIC</td>
<td>GECCo spoken collection</td>
<td>BACKBONE-EN</td>
</tr>
<tr>
<td></td>
<td>MICASE</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Sources for the GECCo spoken part

For English, we use the data of the MICASE corpus, the English part of the BACKBONE corpus and the ELISA corpus. The Michigan Corpus of Academic Spoken English (MICASE) is a collection of nearly 1.8 million words of transcribed speech—almost 200 hours of recordings—from the University of Michigan and includes lectures, classroom discussions, lab sections, seminars, and advising sessions, cf. (Simpson et al., 2002). The BACKBONE pedagogic corpus contains corpora of video-recorded spoken interviews with native speakers of various European languages, cf. (Kohn, 2011). The ELISA corpus contains interviews with native speakers of English talking about their professional career (e.g. in tourism, politics, the media or environmental education), cf. (Braun, 2006). The data from the corpora were extracted according to criteria such as nationality of speaker, type of speech event, degree of speaker interaction. For German, we use the German part of the BACKBONE corpus, which contains interviews with German native speakers (including variants of German). This subset is comparable to the interviews in ELISA and the English part of the BACKBONE corpus. In addition, we compile our own corpus of spoken academic discourse consisting of lectures from all departments of the Saarland University. The lectures were recorder by VISU (Virtual University of Saarland) and have been transcribed by our team according to the transcription guidelines described below.

3.2 Problems in Spoken Data Compilation

In the process of data collection for the German part of spoken academic discourse, we have encountered a number of practical problems. For instance, we initially planned to include recordings of seminars for the analysis of dialogues. However, the seminars in Germany turn out to be less interactive and dialogic than assumed and hence, do not correspond to their English counterparts. Moreover, the collected student presentations constitute prepared speech and thus lack the authentic character of spontaneous speech. Therefore, our German academic corpus currently consists of lecture recordings which are comparable in their speech conditions to the English lectures.

Besides that, we had to apply manual transcription methods which is very labour- and time-consuming. Yet, the recorded data was found to contain too much noise to permit an automatic transcription (speech recognition). Moreover, manual transcription requires the formulation of transparent transcription guidelines. Since the English data was transcribed according to differing guidelines we elaborate a consistent scheme for tags in both languages to annotate extra-linguistic information (example (1)), linguistic variants (example (2)) and repairs and repeats (example (3)).

1) LAUGHTER:
<laugh>text</laugh>

CONTEXTUAL EVENTS:
<writing_on_board>text</writing_on_board>
<door>
<brake type="gasp">

2) EO-INTERVIEW
<turn speaker="Lauren"> <alternative text="yep"> Yes </alternative>, absolutely.
<alternative text="yeah">Yes</alternative>, I
<brake> <alternative text="yeah"> yes
<alternative>, absolutely </turn>

GO-INTERVIEW
<turn speaker="Stefan"> <alternative text="We'ner"> Wenn wir </alternative> die Netze erreicht </alternative> text="hand"> haben
<alternative>, <alternative text="weret"> werden </alternative> die Netze </alternative>
In order to guarantee comparability in frequency and function of cohesive devices between the written and spoken registers we had to restrict each register to 10-14 texts with around 34 thousand tokens each. The existing registers of written language contain both comparable and parallel texts of English and German. However, for the spoken registers, only comparable texts are available, cf. table 1. One possible solution for obtaining aligned texts would be to create interpretations for the existing originals. Interpreted texts, however, are produced under very specific conditions and are affected by various constraints such as time pressure, limited short-term memory capacity, linearity and others, see e.g. (Gumul, 2010) and (Pöchhacker, 2001). They are not considered as reflecting spontaneous speech on the one hand and differ considerably from translations, on the other hand. We thus consider to integrate transcriptions of films and their synchronizations in our corpus, although these are subject to other limitations described, for example, by (Herbst, 1994) and (Döhring, 2006).

4. Corpus Annotation

The spoken registers of the multilingual corpus are annotated on the same level as its written part:

1) **token level:** words, lemmas, parts-of-speech;
2) **chunk level:** sentences, syntactic and semantic chunks and their grammatical functions;
3) **cohesion level:** cohesive devices and cohesive chains;
4) **text level:** registers;
5) **extra-linguistic level:** meta information.

The automatic annotations of parts-of-speech, chunks and their grammatical functions are obtained with the help of the Stanford Parser, cf. (Márneff, et al., 2006). Cohesive devices, such as conjunctive relations, personal and demonstrative reference, substitution, ellipsis and lexical cohesion, are semi-automatically annotated with a tool based on the YAC recursive chunker, cf. (Kermes, 2003) which utilises the CWB Perl-Modules developed within the framework of YAC, cf. (Kermes & Evert, 2001) and (Kermes & Evert, 2002). We also apply the MMAX tool, cf. (Müller & Strube, 2006) for the manual correction of these annotations. Disambiguation of cohesive devices is based on the analyses described in (Kunz & Steiner, in progress) and (Kunz, 2010).

We also aim at annotating reference and lexical chains in our corpus. For this, we apply one of the existing systems for coreference resolution, the Stanford Coreference Resolution System described by (Lee et al., 2011). Our preliminary evaluation tests, see (Amoia et al., 2012), show that the system does not perform with the desired accuracy. Therefore, we also plan to manually improve annotations for this category of cohesion.

The corpus metadata include not only the information on speaker, such as age, sex (female, male, unisex, undefined), profession (translator, teacher, professor, student, etc.) and role (interviewer, interviewee, lecturer, etc.), but also the information on register analysis: field (experiential domain and goal orientation – argumentation, exposition, instruction, narration, description and persuasion), tenor (number of speakers, agentive role – monologic or dialogic, social role – equal, up or down, social hierarchy – expert to expert, expert to layperson, layperson to expert, layperson to layperson, social distance – formal or not) and mode (language role – ancillary or constitutive, channel – graphic, phonic or electronic, and medium – written, written to be spoken, spoken).

5. Corpus Querying

The corpus can be queried with the Corpus Query Processor (CQP, (Evert, 2005)), which allows us to detect candidates for cohesive devices by means of regular expressions, offering several functionalities for extraction (e.g., context expansion) and sorting purposes (e.g., counting, grouping of results). CQP allows two types of attributes: positional (e.g. for part-of-speech and morphological features) and structural (e.g. for chunks, registers or extra-linguistic information). These attributes are employed for CQP-based queries which include string, parts-of-speech, chunk, register and further constraints, cf. table 2.

<table>
<thead>
<tr>
<th>Query elements</th>
<th>meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>word=&quot;and&quot; &amp;</td>
<td>word <em>and</em> which is</td>
</tr>
<tr>
<td>cohesive_device=&quot;conj&quot; &amp;</td>
<td>cohesive</td>
</tr>
<tr>
<td>text_register=&quot;INTERVIEW&quot; &amp;</td>
<td>conjunction</td>
</tr>
<tr>
<td>tenor_numberOfSpeakers=&quot;2&quot; &amp;</td>
<td>in interviews only</td>
</tr>
<tr>
<td>speaker_ager=&quot;31-50&quot; &amp;</td>
<td>with 2 speakers only</td>
</tr>
<tr>
<td>tenor_socialRole=&quot;equal&quot;</td>
<td>aged between 31-50</td>
</tr>
<tr>
<td></td>
<td>in equal social role</td>
</tr>
</tbody>
</table>

Table 2: Example of a CQP query

The present CQP query delivers a list of concordances, as shown in example for the cohesive conjunction *and* (4).

4) 8: My name’s Norma Holt *and* I actually come from the Wirral Peninsula which is on the west coast of Liverpool, which is Lancashire...
Moreover, the sorting, counting and grouping functionality of CQP allows us to extract frequency information, as shown in table 3 (for English only as the German ACADEMIC part is still under construction). The obtained frequencies of cohesive phenomena can then be evaluated in terms of their distribution across registers, languages and modes.

For instance, table 3 displays the frequencies per million words of all cohesive occurrences of the form one in its function as nominal substitute. What the table nicely illustrates is that some registers show more commonalities in their distribution of cohesive one than others, and most notably that there is a considerable difference in frequency between the spoken and the written registers of our subcorpus. In addition, the two registers FICTION and speech are closer to the spoken registers then others. This may be due to the fact that FICTION contains text passages imitating spoken dialog and that SPEECH was written to be spoken. Thus, ACADEMIC seems to be at one end of the spoken written continuum of our corpus and SHARE at the other end (at least as far as cohesive one is concerned) with FICTION and SPEECH taking a somewhat middle position.

<table>
<thead>
<tr>
<th>register</th>
<th>Cohesive one per 1M</th>
</tr>
</thead>
<tbody>
<tr>
<td>spoken</td>
<td></td>
</tr>
<tr>
<td>INTERVIEW</td>
<td>949,84</td>
</tr>
<tr>
<td>ACADEMIC</td>
<td>2769,33</td>
</tr>
<tr>
<td>FICTION</td>
<td>378,42</td>
</tr>
<tr>
<td>SPEECH</td>
<td>199,65</td>
</tr>
<tr>
<td>ESSAY</td>
<td>85,72</td>
</tr>
<tr>
<td>SHARE</td>
<td>83,74</td>
</tr>
<tr>
<td>INSTR</td>
<td>110,60</td>
</tr>
<tr>
<td>TOU</td>
<td>124,02</td>
</tr>
<tr>
<td>POPSCI</td>
<td>142,26</td>
</tr>
<tr>
<td>WEB</td>
<td>166,18</td>
</tr>
</tbody>
</table>

Table 3: Frequencies delivered by CQP

6. Conclusion and Future Work

We have compiled a spoken corpus for English and German that is enhanced with annotations on several linguistic and extra-linguistic levels. Our corpus architecture not only allows a text-based contrastive analysis of cohesion in German and English but also permits a comparison of various spoken and written registers. Therefore, the findings based on our resources will not only complement the existing research gaps in cohesion but also enrich contrastive grammars with a systematic account of discourse phenomena in written vs. spoken mode. Moreover, both the developed resources as well as our findings on cohesion will provide valuable insights for language teaching and translator training and will open up new research options for various fields.

In the future, we aim at expanding corpus with further registers, e.g. internet forums, TV talk shows and reports. Besides that, we will develop further procedures to automatically annotate cohesive devices and relations. We also plan to enhance our spoken corpus with translations. The corpus will be available for querying online within the CLARIN-D initiative.

7. Acknowledgements

The project GECCo (German-English Contrasts in Cohesion) is supported by a grant from Deutsche Forschungsgemeinschaft (DFG, German Research Foundation). We thank our colleagues in the GECCo team – Katrin Menzel and Erich Steiner for their assistance. Besides that, we are especially grateful to Hannah Kermes for providing the necessary perl script for adaptation.

8. References


Brinckmann, C., Kleiner, S., Knöbl, R. and Berend, N.


Simpson, R.C., Briggs, S.L., Owens, J. and Swales, J.M.


LINDSEI-BR: an Oral English Interlanguage Corpus

Heliana MELLO, Luciana ÁVILA, Tufi Neder NETO, Barbara ORFANÒ
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

Corpus Linguistics has been more than instrumental in the study of interlanguage. It has made it possible for researchers not only to have access to large quantities of varied interlanguage samples but also process these data both for individual language features as well as for a host of other elements, such as interlanguage feature comparison. Presently there are many interlanguage corpora available to researchers and teachers, both written and oral, and this has afforded a spurt in interesting findings as far as the manyfold processes involved in language acquisition are concerned. In this paper, we will present a new English interlanguage corpus under compilation in Brazil, the LINDSEI-BR. It is associated with a larger project - the COBAI; the Brazilian Oral Corpus of Learner English is a repository of spoken interlanguage data that aims at gathering varied subcorpora of Brazilian learner English with the main purpose of providing data for the study of interlanguage features within the frame of second language acquisition research. The larger project was launched in 2011 and so far it is concerned with the compilation of the LINDSEI-BR, a component of the Louvain-based project Louvain International Database of Spoken English Interlanguage.

Keywords: interlanguage; learner oral corpus; Brazilian Portuguese; English.

1. Introduction

Corpus Linguistics has been more than instrumental in the study of interlanguage. It has made it possible for researchers not only to have access to large quantities of varied interlanguage samples but also process these data both for individual language features as well as for a host of other elements, such as interlanguage features at a given acquisition stage, comparative error analysis, among others. Presently there are many interlanguage corpora available to researchers and teachers, both written and oral, and this has afforded a spurt of interesting findings as far as the manyfold processes involved in language acquisition are concerned. In this paper, we will present a new English interlanguage corpus under compilation in Brazil. It is associated with a larger project - the COBAI. The Brazilian Oral Corpus of Learner English (COBAI) is a repository of spoken interlanguage data that aims at gathering varied subcorpora of Brazilian learner English with the main purpose of providing data for the study of interlanguage features within the frame of second language acquisition research. The project was launched in 2011 and so far it is concerned with the compilation of the LINDSEI-Brazil, a component of the the LINDSEI international project, which will be presented in this paper.

The Louvain International Database of Spoken English Interlanguage (LINDSEI) project is an international initiative coordinated at the Centre for English Corpus Linguistics, at the Université Catholique de Louvain (cf. Gilquin, De Cock & Granger, 2010). The LINDSEI project encompasses seventeen different interlanguage subcorpora, compiled with the same parameters and transcribed following the same guidelines. The LINDSEI project is the oral counterpart for the ICLE – International Corpus of Learner English, compiled by the same team of researchers under the direction of Sylviane Granger (cf. Granger, 2003; Granger et al., 2009).

The LINDSEI-BR is being compiled following the international project guidelines. At present we have achieved our recording goal of fifty recordings and their transcription is underway. The recording informants were university, high intermediate to advanced level students of English as a second language. The recordings covered three different tasks: a narrative about a chosen set topic by the informant, free discussion with the interviewer and the description of a pictured scene. Each recording is on average twenty minutes long and features quasi spontaneous speech patterns. For each recording there is an accompanying learner profile that covers the learner’s language history and other elements that might have contributed to her/his process of language acquisition, besides having information about the interviewer and the actual interview itself. The transcription guidelines include a code for each recording, speakers’ turns, and the marking of several speech features, such as: overlapping, pauses, backchannelling, contractions, truncation, among others.

2. LINDSEI-BR participant profiles

Following the LINDSEI guidelines, all participants recorded are third, fourth year students of English. The participants are recruited by the researcher and are aware that they are contributing their speech to the compilation of a corpus. All participants have to fill in willingly a learner profile in which information about their acquisition history is reported through the number of years of study, context of English learning, etc. In order for a recording session to be incorporated into the corpus, participant permission is necessary.

The participants in the LINDSEI-BR study in a major federal university in Brazil and have chosen English as their major. Many are already ESL teachers, although this per se does not mean a level stage of acquisition among informants, as there might be sever
different proficiency levels even among ESL teachers.

3. The recordings

Recording sessions took place in the first semester of 2011, and were carried at the Laboratory for Empirical and Experimental Language Studies (LEEL) at UFMG. The interviewer is a Brazilian with high proficiency level in English as a foreign language. It was not possible for the LINDSEI-BR team to arrange for a native speaker of English to carry the recordings. This is a shortcoming of the project since conversations might evolve differently between native and non-native speakers versus two non-native speakers of English with the same mother tongue background.

The recordings were carried with the following equipment: Recorder Marantz PMD660 Professional Solid State Recorder, unidirectional wireless microphones Sennheiser ME 4 clip-on (cardioid), cable Sennheiser CL100 (connectors XLR and mini jack 1/8”), receivers Sennheiser EM 100 G2 A, Sennheiser EK 100 G2 A, Sennheiser EK 100 G3 and transmitter Sennheiser SK 100 G2 A and Sennheiser SK 100 G3.

Recording files are wav format and in general have good acoustic quality. Some sessions have some background noise but this does not prevent understandability.

4. Some remarks about the transcriptions

Transcriptions are being carried at present by undergraduate research assistants. The transcribed files are revised by the project coordinator. No intertranscriber validation process has been carried so far but this is one of the goals the project to-do list.

Transcriptions follow the guidelines made available by the Leuven LINDSEI team and encompass the following aspects: a header, ch nt=“BR” nr=“BRXXX”, which indicates that participant number XXX is a native speaker of Brazilian Portuguese; turns are marked for interviewer <A> and interviewee <B>; each turn end carries the corresponding end tag, either </A> or </B>; overlapping is annotated at its onset with the tag <overlap /> in the underlying turn and also at the beginning of the overlapper’s turn, however its end is not annotated; the British orthographic convention is followed. There are several specific guidelines that cover empty pauses, filled pauses and backchannelling, unclear passages, anonymisation, truncated words, contracted forms, non-standard forms, dates and numbers, some phonetic features, among others. An example of a transcription is given below:

Example 1:

<B> <overlap /> and: I’m going talk about a movie. that I saw ... <overlap /> (erm), the: Inception.. with Leonardo DiCaprio. and I. I thought that. is a: very ..good movie .. very interesting ..and: <overlap /> . I don’t know it’s so .. (eh) . first of all the the: . photograph= of the movie . is amazing . the: . special effects . that they use . is very nice . (erm) and it was the first movie that I saw with my boyfriend= <laughs> <overlap /> and . we stayed for . three hours . in the cinema. and we:. (eh) tired and the . movie. (eh) (er) . how can I say </B>

As can be seen above, there are some specific markings such as: some end of words are followed by colons, which indicate last syllable lengthening (eg. and:); there are fillers such as (erm); non-verbal sounds are annotated (eg. <laughs>); truncated words are marked with = (eg. photograph=); silence is annotated through dots (eg. ..., meaning 1-3 seconds).

The transcriptions do not contemplate pronunciation interlanguage features. In order for phonetic-based studies to be carried using this material, further annotation must be added.

5. Future directions

LINDSEI-BR is still on the making; therefore, much remains to be done in order for it to be ready to be offered to researchers. However, plans are for the transcription process to be concluded within the year 2012. Additionally, some analysis has already been carried using data provided by this corpus, especially focusing on phonetic-phonological aspects of interlanguage speech (Medina, 2012). Future plans upon transcription completion include the addition of interlanguage feature annotation in order to facilitate researchers’ use of the corpus.

6. References


Bi-national bi-modal bi-lingual corpora of child language

Ronice Müller de QUADROS, Diane LILLO-MARTIN, Deborah CHEN PICHLER

Universidade Federal de Santa Catarina; University of Connecticut; Gallaudet University
ronice.quadros@ufsc.br, lillo.martin@uconn.edu, deborah.pichlerr@gallaudet.edu

Abstract

This paper discusses projects involving the building of corpora of sign language acquisition data. We developed a methodology to collect, to transcribe and to store data from different contexts of acquisition. The corpora include deaf children, from deaf parents; deaf children, from hearing parents; hearing children, from deaf parents (Codas) and deaf children with cochlear implants. There are two sign languages involved: Brazilian Sign Language and American Sign Language and two spoken languages, in the bilingual bimodal cases, that are, Brazilian Portuguese and American English. The complexity of building these corpora includes development of patterns of transcription and the organization of the same metadata system. In this process, we are developing manuals, database and software to make the data available and comparable across the languages. One example of software that we present in this paper concerns Sign ID, that is, it is software to indicate identities for each sign that is part of the database. The Sign ID software helps us make the annotations more consistent across transcribers. This kind of work is making it possible to compare data from these languages.

Keywords: sign language; corpora; and language acquisition.

1. Introduction

In order to address numerous linguistic research questions, we have been building several corpora of sign language acquisition data. Until recently, our focus had been on sign language only with deaf children, from deaf parents, acquiring sign language as native language. In this case, we built corpora of longitudinal data collected over a long period of time: these corpora included spontaneous data, with interaction of the child from 1-4 years old and an adult (usually the Deaf mother or a Deaf experimenter). On the Brazilian side, there is also data from deaf children with hearing parents. In this context, a Deaf experimenter interacts with the child in sessions alternating with the hearing mother. All the analyses done so far indicate that in the specific context of deaf children with deaf parents, the sign language acquisition is parallel to spoken language acquisition (see Lillo-Martin, 1999 and Newport & Meier, 1985 for reviews of some of this). However, there are also findings showing that certain aspects of language acquisition in this context show modality effects (e.g. Meier & Newport, 1990; Marentette & Mayberry, 2000; Meier, 2006). On the other hand, in the context in which the deaf child has limited contact with sign language, there is a lot of variability in the language development reported by different researchers, but it seems that even in these contexts in which input is not conventional, because the child has parents learning sign language and restricted or no access to sign language, the child develops his/her signing skills better than his/her parents, showing that the child is able to make better use of the mental language system (e.g. Singleton & Newport, 2004; Goldin-Meadow, 2003; Goldin-Meadow & Mylander, 1984, 1990, 1998; Quadros & Cruz, 2011).

Now we are expanding our work to include bimodal bilingual children acquiring both a sign language and a spoken language, building comparable corpora across two sign/spoken language pairs: Brazilian Sign Language and Brazilian Portuguese on the one hand, and American Sign Language and American English on the other. We are again collecting longitudinal data with babies from 1 to 4 years old, and adding experimental data with children from 4 to 7 years old.

We use different sets of researchers (deaf and hearing) to emphasize appropriate target language use, assuming the child’s interlocutor sensitivity (Petitto et al., 2001), but we also recognize that code-blending is simply a part of the language system being acquired.

We reorganized the form of the database used with the longitudinal data and we built a new database for the experimental studies. The experimental studies include a set of 24 tests, evaluating different language aspects, such as, morphology, phonology, syntax, discourse and pragmatics. The goal of the tests is to provide a comprehensive profile of each bilingual child’s developing competency in Libras (Brazilian Sign Language) and Brazilian Portuguese, or ASL (American Sign Language) and American English.

The data in sign and in speech adds considerable complexity to the already challenging prospect of corpus building. In this presentation, we explore some of the issues we have faced already and those we expect to face, in the context of our linguistic goals.

Recent research on childhood bilingualism has indicated that although children have two separate developing grammatical systems from very early on, there are instances of cross-linguistic influence, where grammatical structures from one language seem to exert a temporary influence on the child’s grammar of the other language (e.g. Hulk & Müller, 2000). An important question is to identify the loci of such influences based on linguistic criteria. In order for us to address such issues, we are developing corpora from individual children acquiring both a sign language and a spoken language. Many of the same data collection issues arise as those for projects investigating only sign language (see Baker & Woll, 2005 for some best practices in this domain). However, in our current project, it turns out that there are specific things for which additional practices are needed; for instance, we frequently observe code-blended...
language (the use of signs and speech produced simultaneously) as well as unimodal productions (Bogaerde & Baker, 2005, 2009; Emmorey et al., 2008). Language- or modality-specific properties as well as universals are found to be very interesting in these contexts. In this paper, we will present the organization of the sign language acquisition corpora developed on both sides of the project: Brazil and the United States of America.

2. Metadata
The metadata of the children is organized through documents that are shared with researchers involved in the different steps of the investigation: data collection involving filming, transcribers, people that organize the data for specific purposes and people that analyse the findings. The main topics of the documents are the following:

LONGITUDINAL
- Protocol of the child (nickname of the child, for example, EDU)
- Number of the section (from 000 up to the number of the sections collected, for example, EDU001, EDU002, EDU003)
- Date of the filming
- Age of the child (years;months.days)
- Target language
- Duration of the session
- Adults involved in the session
- Other participants involved in the session
- Comments
- Transcribers
- Checker/reviser of the transcription
- Organizer of the data for each purpose (for example, for WH analysis, for Modality analysis, etc.)

EXPERIMENTAL
- Name of the test
- Nickname of the child
- Condition (Coda, Deaf, CI, Coda adult)
- Date
- Age
- Language
- Duration
- Comments
- Transcriber
- Reviser

The whole database is organized in a computer server. See Figure 1 for an illustrative sample of this organization. There are two main folders: the original archive (“acervo”) and the production. The first one has the original videos. The second one has the compressed videos for manipulation by the people that access the videos, as well as transcription and analysis files.

The production folder includes the experimental data and longitudinal data in separate sections. First we discuss the longitudinal data. The basic organization is to list the children in separate folders. Each child’s folder will include the folders for each session containing the video and the transcript files (the basic one and the ones with the specific organization for specific purposes). The transcription is done using ELAN software producing eaf files with separate tiers of annotation capturing different types of information (see also below).

For the experimental studies, the basic organization is to have the folders with the places and years in which the fairs happened. Within each place, the folders are separated by test. These folders are further divided into two sets of data by child: one for those whose data is without restriction (“sem restrição”), and another for restricted data (“com restrição”). The restrictions are related to the kind of access people have to the videos. Some of the parents do not want students to have access to the videos of their child or for the researchers to use frames of the videos in conferences, for example. Within these two folders based on restriction, the children, then, are listed with the video and the eaf or the form of the test scanned with the results, depending on each test.

In the case of the experimental studies, the database is organized as well as using FileMakerPro (Figure 2 in Appendix). This database includes all four languages. Then, it facilitates the comparison among the experimental results over the four languages.

Figure 1: Example of the organization of the database
3. Designing annotation patterns

Following video collection, we invest considerable energy in the production of transcripts, to be used in conjunction with the videos for linguistic analyses. Following our earlier sign-only research, we use ELAN for time-locked videos with transcription (http://www.latmpi.eu/tools/elan/).

For bilingual research, we designed a different template so that both languages are parent tiers, to optimize the study of (sequential or simultaneous) bimodal productions. See Chen Pichler et al. (2010) for a detailed description of our ELAN tier structure and transcription conventions (cf. Figure 3 and Figure 4, in Appendix).

The general principles that guide the annotation of the data are to create a machine-readable record of language samples, not necessarily sufficient for the reader to reproduce in exactly the same way, but so that the records can be searched to find all occurrences of phenomena of interest (in the way described by Johnston, 2001; Johnston & Schembri, 2007; Miller, 2001; Pizzuto & Pietrandrea, 2001). In addition to having a basic annotation of the utterance in each language, we use multiple annotation parses focusing on different phenomena. This documentation of the data is the foundation for our analysis decisions.

Where it is possible, we follow the CHILDES conventions established for child language data (MacWhinney, 2000) in transcribing both speech and sign (though we do not use BTS) http://childes.psy.cmu.edu/manuals/chat.pdf. When the CHILDES conventions conflict with our sign-specific goals, we create new conventions to be followed for transcribing both sign and speech. It is important to keep the sign and speech transcriptions comparable.

4. Sign IDs

Finally, we see a number of important implications and extensions of the system we are developing. For example, we are creating a specific identification for each sign to be used in our transcripts (in the same spirit of Johnson, in preparation, for Australian Sign Language), what we call “Sign ID”. Because there is no commonly accepted writing system for sign languages, sign researchers generally rely on a system of glossing; however, traditional transcription does not assign a consistent gloss for each sign, but different glosses depending on context and other aspects of the signed utterance. This means that it is very difficult for researchers to identify the locations of interest in a transcript using a search function to discover all occurrences of a particular sign. Analysis must proceed at a much slower pace of hand searching transcripts one utterance at a time. In order to facilitate and expand the analysis of data collected in the parent project, we developed a sign ID lexicon containing the vocabulary items used most frequently by the children we are studying. Sign IDs are word labels chosen to represent each sign root systematically, so that every use of the sign has the same label, despite contextual or morphological differences which affect how the sign is interpreted. By using sign IDs in our transcripts, we are able to conduct our analyses more efficiently, using a wider range of data. The sign ID lexicon addresses the problem of transcript searchability and greatly facilitates the analysis of data collected for sign language corpora. This helps to standardize annotations and it can be more freely accessed by other researchers.

On the Brazilian side, we have been developing the sign IDs database by feeding it with the signs over which transcribers had doubts regarding transcription. We have periodic meetings to discuss these signs, then we christen each and add it to the ID list (www.idsinais.libras.ufsc.br) (see Figure 5 in Appendix for the Sign ID screen). The search system has filters based on sign language parameters (132 handshapes divided in 13 groups and 8 locations). An example with a group of handshapes chosen as a parameter to search for a specific sign is given in Figure 6 and the results of this search are shown in Figure 7, in Appendix.

The sign ID specifications include identification of the sign, Portuguese translation, English translation, written sign, handshape groups, handshapes, location and sign video. The searching may be done through handshapes, locations, handshape groups, location groups, the sign ID or the first letter of the sign ID.

On the American side, the development of an ID gloss database has taken into consideration the needs of different research groups across the country, each of which uses a different system for writing signs. The database was set up so that different local groups can enter their own information about each sign, and each group can also view the information entered by the others. This approach will facilitate the comparison of transcriptions used across different groups, and may eventually lead to greater convergence in the glossing systems used.

5. Conclusion

One of our major goals has been cross-site comparability, that is, establishing the same criteria, approach to data collection, ELAN template, and general transcription principles to be used across our three universities. The metadata and data are shared through the use of a common server, as well as online services including Google docs and Dropbox. The analyses of the results are being conducted through regular meetings and we are on the right track to answer our research questions (e.g., Lillo-Martin et al., 2010; Chen Pichler et al., 2010; Quadros et al., in press).

We have not yet resolved the following linguistic issues, but we hope that our project will contribute to their discussion in the field as a whole. Does bimodal bilingualism lead to cross-language influence different from that found in mono-modal bilingualism (e.g., due to code-blending, or use of non-manuals)? When bimodal bilinguals code-blend, are they choosing grammatical structures which are permitted in both languages for maximum accommodation? What kinds of syntactic
representations can account for code-blends? These are the types of research questions our project can address through the use of the corpora we are now building.

Our template and corpus-building decisions can be applicable to the development of adult only bimodal bilingual corpora. In addition, many similar issues arise in the study of co-speech gesture, and researchers in this area may take advantage of aspects of our procedures. And, we hope that our collaboration across continents may contribute to and promote cross-linguistic research on sign languages as well.

6. Acknowledgements

This research is supported by the U.S. National Institutes of Health – NIDCD grant #DC00183 and NIDCD grant #DC009263; by a Gallaudet University Priority Grant; and by the Brazilian National Council for Research, CNPq Grant #CNPQ # 304102/2010-5 and # 471478/2010-5.

We sincerely thank the Deaf consultants, research assistants, children, and their families who work with us in our research.

7. References


Johnston, T. (in preparation). From archive to corpus: Transcription and annotation in the creation of signed language corpora, manuscript. Department of Linguistics, Macquarie University, Australia.


8. Appendix

Figure 2: FileMakerPro

Figure 3: ELAN in the context of Bibibi Project with the basic tiers for the child

Figure 4: ELAN in the context of Bibibi Project with the specific tiers for modality analysis
Figure 5: ID screen for Libras

Figure 6: ID searching system: Handshape selection

Figure 7: ID result of a search
C-Or-DiAL (Corpus Oral Didáctico Anotado Lingüísticamente) y la enseñanza del español

Carlota Nicolás MARTÍNEZ
Università degli Studi di Firenze
Via Santa Reparata, 93 – 50122 Firenze
carlota.nicolas@unifi.it

Abstract
En este artículo se describen las características, el proceso de elaboración y la utilidad didáctica de C-Or-DiAL. Este corpus está formado por 118.756 palabras que proceden de alrededor de diez horas de grabación de los siguientes géneros discursivos: conversación y diálogo (29%), entrevista informal (51%), conversaciones con tema preestablecido (13%) y clases o conferencias (7%). El texto etiquetado de la transcripción está precedido de una cabecera con informaciones generales sobre la elocución (participantes, situación, tema y palabras claves) y otras más específicas con indicaciones y propuestas para la enseñanza de la lengua (nivel del alumnado con el que usar la sesión, lista de palabras poco usadas, aspectos lingüísticos y funciones comunicativas que se pueden aprender en esa elocución). Este rico corpus está a disposición de quien quiera renovar el modo de enseñar y de aprender la lengua oral espontánea.

Keywords: Corpus orales del español; enseñanza de la lengua; géneros discursivos; base de datos.

1. Características y elaboración de C-Or-DiAL

1.1 Qué es
Es un corpus de la lengua oral espontánea recogida en grabaciones y transcrita ortográficamente, etiquetada prosódicamente y con las funciones comunicativas anotadas.

Es un corpus que además de ser un recurso para la investigación puede ser utilizado como material para la enseñanza de la lengua española. Para facilitar este uso se ofrecen en la cabecera de cada texto indicaciones y propuestas específicas para la enseñanza (nivel del alumnado con quien usar el texto, lista de posibles palabras desconocidas, observaciones lingüísticas).

1.2 Quién lo ha hecho

1.3 Cuándo se ha hecho
La primera grabación se hizo en el 2004.
El corpus se introdujo en la base de datos en noviembre del 2012. El libro sobre C-Or-DiAL se publica en julio de 2012. Las transcripciones son revisadas y corregidas periódicamente.

1.4 Cuánto material contiene
Es un corpus de dimensión media: 118.756 palabras transcritas que proceden de alrededor de diez horas de audio.
Son 240 sesiones compuestas por la transcripción de los correspondientes 240 audios; estos han sido extraídos de las 72 horas de grabaciones hechas en los últimos 9 años.
En la Tabla 2 (Appendix) se muestra el número de palabras de cada uno de los géneros que forman la estructura de C-Or-DiAL.

1.5 Dónde se ha hecho
Las grabaciones de C-Or-DiAL se han hecho en Madrid con el apoyo técnico del Laboratorio de Linguística Computacional de la Universidad Autónoma de Madrid.
Las transcripciones y toda la elaboración del corpus se ha hecho en la Università di Firenze.

1.6 Dónde consultar el corpus C-Or-DiAL

1.6.1. C-Or-DiAL base de datos
Las sesiones C-Or-DiAL se pueden extraer de la base de datos que lo aloja en LABLITA (Laboratorio de Linguística Italiana) Università di Firenze (lablita.dit.unifi.it/app/C-Or-DiAL/index.php).
En la Tabla 3 (Appendix) se muestra la página de acceso a las sesiones de C-Or-DiAL desde la que abrir cada texto y cada audio del corpus, y donde consultar las informaciones sobre cada sesión: Título y tema, Tipología de los textos, Número de hablantes, Situación, Número de palabras, Minutos, Uso didáctico (lablita.dit.unifi.it/app/C-Or-DiAL/corpus.php). Tambien se accede al corpus desde la Búsqueda avanzada (lablita.dit.unifi.it/app/C-Or-DiAL/search.php) que utiliza listas cerradas con informaciones sobre: Tipología de texto, Palabras clave, Nivel de uso didáctico y Funciones comunicativas.

1.6.2. Libro de C-Or-DiAL
Se ha editado en el 2012 el libro C-Or-DiAL (Corpus Oral Didáctico Anotado Lingüísticamente) publicación
de LICEUS EDICIONES en dos formatos: en papel, acompañado de un cd, y en formato electrónico. Esta publicación contiene el corpus con todas sus sesiones y, además, una detallada descripción sobre la elaboración, las características y los posibles usos didácticos de C-Or-DIAL.

2. Estructura y contenidos de C-Or-DIAL

2.1 Qué estructura tiene

2.1.1. Macroestructura
C-Or-DIAL contiene 240 sesiones compuestas por las transcripciones y los audios correspondientes.

Estas sesiones tienen diferentes tamaños y géneros discursivos.

La distribución de C-Or-DIAL en géneros discursivos se ve en la Tabla 1 (Appendix), en la que se evidencia, mediante cuatro parámetros de clasificación, el rasgo de espontaneidad en la lengua que predomina en este corpus.

Los tamaños de las sesiones son:

- 89 audios de hasta 2 minutos (01:58:37 horas);
- 67 audios de 2 a 3 minutos (02:46:48 horas);
- 57 audios de 3 a 4 minutos (03:09:47 horas);
- 27 audios de 4 a 8 minutos (02:18:15 horas).

2.1.2. Microestructura
Cada texto de las sesiones está compuesto por la cabeecera y el texto transcrto.

La cabecera tiene datos y metados:

- Informaciones sobre las características básicas de la sesión (número de minutos y de palabras, grabación de la que procede el fragmento, nombres de los archivos y de los transcriptores y revisores);
- Informaciones del contenido del texto (tema, informaciones sobre los participantes, situación en la que transcurre la elocución);
- Indicaciones y propuestas específicas para la enseñanza (nivel del alumnado con el que usar la sesión, lista de palabras poco usadas y de interés para ser estudiadas, aspectos lingüísticos y funciones comunicativas que se pueden aprender en esa elocución).

2.2 A quién se ha grabado

Los participantes de C-Or-DIAL, todos ellos anónimos, son más de 50. Denominados con tres letras mayúsculas que mantienen en todas sus intervenciones. La cultura de los participantes es media-alta (universitarios en general). Son personas de mediana edad (entre 30 y 60 años), solo ocho de estas personas tienen menos de 10 y más de 70 años.

El 99% de las personas es de Madrid.

2.3 Cómo se recoge el habla

El 30% de las grabaciones se han hecho sin que las personas supieran que se les estaba grabando; en estos casos se ha pedido el permiso para utilizarla al acabar la grabación. El 7% son grabaciones hechas en salas de conferencias y en aulas; el 63% restante se ha hecho pidiendo permiso a los participantes antes de iniciar la grabación, en estos casos la relación de amistad o la situación familiar hacía que la grabadora no fuera un impedimento para que se hablara con gran naturalidad.

2.4 Cuándo y dónde se recoge el habla

Las grabaciones se han hecho en distintos momentos del día.

Los lugares de las grabaciones han sido los normales de la vida cotidiana: casas particulares, cafés o bares. Han sido grabadas en lugares de trabajo las conferencias, las clases, las específicas sesiones de trabajo, y cinco entrevistas de las 20 realizadas.

2.5 Cómo se transforma el habla de la grabación en el texto de la transcripción

El primer paso para crear las sesiones ha sido fragmentar las grabaciones originales de larga duración en audios de pequeño tamaño (ver 2.1.1.). En cada audio se habla al menos de un tema claro, este ha sido el criterio de fragmentación.

Estos audios se fueron entregando a los alumnos de la Universidad di Firenze de los cursos de Lengua Española del 2005 al 2012 que tuvieron la obligación de hacer con cada uno su transcripción como parte del programa del curso.

Para la transcripción se usaron reglas que derivan de las usadas en C-ORAL-ROM.

Las correcciones y control final de todas las transcripciones es responsabilidad de Carlota Nicolás.

3. Utilización de C-Or-DIAL

3.1 Qué uso se puede hacer de C-Or-DIAL

C-Or-DIAL está diseñado como corpus para la investigación y para la enseñanza de la lengua oral.

Hasta ahora C-Or-DIAL se ha utilizado como valioso contenedor de muestras reales de la lengua oral espontánea, en el que analizar sus características. En estos años de trabajo con los alumnos de Lengua Española se ha constatado que hacer transcripciones es un modo muy válido para el análisis de la lengua oral.

La transcripción se ha revelado como un método de aprendizaje de gran impacto, pues despertó en el alumno actitudes hacia el aprendizaje poco desarrolladas al trabajar con otros métodos más usuales. La labor del transcriptor no solo es una práctica de minuciosidad y de concentración muy pedagógica, sino que aporta este patrimonio:

- la atención obligada para entender un audio habitúa a escuchar con especial atención;
- el traslado del audio a la escritura (aunque solo sea una transcripción ortográfica sin seguir las normas de puntuación) enseña a diferenciar.
estas dos modalidades;

- marcar los rasgos prosódicos que dependen de la percepción del transcriptor los hace reconocer conscientemente;
- la colocación en la transcripción de las etiquetas obliga a hacer un análisis solo posible si se han aprendido algunas características fundamentales de la lengua oral que son representadas por estas etiquetas.

3.2 Con quién y para quién utilizar C-Or-DiAL

C-Or-DiAL puede ser utilizado en la enseñanza de la lengua española con alumnos de todos los niveles, con ayuda del profesor o, sin ella, realizando su estudio en autonomía.

3.4 Cuándo y dónde utilizar C-Or-DiAL

En cualquier momento del proceso de aprendizaje de la lengua española se pueden incluir, para su estudio, las sesiones de C-Or-DiAL. Un profesor de lengua sabrá adaptar cada sesión al nivel del alumno. C-Or-DiAL es requerido para que el alumno tenga contacto con el español real espontáneo que es el español que necesita comprender y con el que se debe expresar.

Para trabajar con C-Or-DiAL es necesario el uso de un laboratorio informático para que el alumno pueda acceder a las transcripciones y a los audios individualmente y pueda con su propio ritmo trabajar con este material.

4. Relación entre el desarrollo de las habilidades personales del estudiante y la práctica de las destrezas lingüísticas

Al trabajar con C-Or-DiAL se activa la concentración, la percepción auditiva y la necesidad de segmentar lo escuchado para poder llegar a la comprensión oral.

La comprensión oral de los textos del C-Or-DiAL lleva a ejercitar el análisis, la deducción, la inducción y la síntesis.

A partir de los textos de C-Or-DiAL se pueden hacer ejercicios de imitación y recreación lo que conlleva la práctica de la expresión oral, la interacción oral y la expresión escrita.

5. Tres propuestas de actividades para distintos tipos de aprendientes

5.1 Contacto inicial con una sesión de C-Or-DiAL

Actividades preferidas por aprendientes de estilo pragmático y de estilo activo:

- Audición;
- Reconocer variantes prosódicas;
- Coger notas de lo que se oye;
- Buscar las palabras clave;
- Escribir el tema;
- Separar y reconocer palabras, locuciones y colocaciones;
- Reconocer funciones comunicativas;
- Observar aspectos gramaticales.

5.2 A partir de la sesión de C-Or-DiAL

Actividades preferidas por aprendientes de estilo pragmático:

- Dramatización a partir del texto
- Cambiar entonaciones del texto y observar los efectos;
- Escribir lo dicho en el texto con la estructura de una obra dramática;
- Hacer un guión cinematográfico añadiendo movimientos y situación;
- Escribir el resumen de lo sucedido;
- Inventar lo anterior o lo posterior dicho o sucedido entorno al texto.

5.3 Utilización de los recursos de C-Or-DiAL

Actividades preferidas por aprendientes de estilo teórico y de estilo reflexivo:

- Aprender particularidades prosódicas;
- Subdividir los enunciados;
- Analizar las peculiaridades discursivas;
- Reconocer diferencias entre géneros discursivos;
- Controlar las funciones comunicativas relevantes en el texto;
- Observar la estructura temática;
- Conocer la estructura dialógica;
- Aprender palabras, locuciones y colocaciones nuevas.

6. Conclusiones

El mejor modo de concluir esta descripción de C-Or-DiAL y de su uso es presentar una sesión en la que se observan algunas de sus cualidades. Es una conversación entre amigas que no sabían que eran grabadas. Se puede observar en ella su espontaneidad, un modo ejemplar de estructurar la narración, una cierta riqueza de vocabulario, además de otros muchos detalles que se pueden encontrar, y que serán especialmente apreciados por los profesores que buscan materiales reales y ricos para sus estudiantes.

7. Appendix

7.1 Transcripción

@ Archivos:
conv_03_UNA_CHIQUITITA_JAPONESA.txt,
conv_03_UNA_CHIQUITITA_JAPONESA.wav
@ Título: una chiquita japonesa
@ Participantes: CAR, Carlota (mujer, C, 3, profesora, Madrid, vive en Italia desde hace más de 20 años)
PIZ, Pizca (mujer, C, 3, archivadora, Madrid)
ANG, Ángeles (mujer, C, 3, traductora, Madrid, vive en Bélgica desde hace 25 años)
MAI: 1.7 y este verano tuvimos en casa a una japonesa / una criada de veintiún años / vino a casa [///] había estado tres meses en Sevilla / estudiando español // nos aparece / fuimos a buscarla a la estación de [///] de autobuses / te aparece una japonesita así jovencísima con un maletón / con su ordenador portátil / con el que &mm se comunicaba con su familia claro tal // dices pero esta chica / aquí / en España / primero a Sevilla / luego se viene a [///] a Madrid / a casa de un amigo / 6.16 que le decía a Ramón / pues porque somos gente decente / pero es que puedes aterrizar < en casa de un >

*PIZ: < yyy claro > //
*ANG: < en cualquier &sit > //
*MAI: se conocían de un foro en el que hay españoles que estudian japonés / y japoneses que estudian español / un foro de Internet /
*PIZ: yaa //
*MAI: ¿ sabes ? y dices / y de pronto cogen la maleta / y se < colocan en el otro lado del mundo > /
*PIZ: < del Japón > ...
*MAI: a casa de uno que se llama Ramón
*TTT: yyy
*MAI: y que lo has conocido ... y yo / luego le bromeeaba / porque después de casa / se marchó a Barcelona a casa de otro del foro / 6.16 yo le decía de broma / &eh ¿ ha llegado ya a casa del violador del Ensanche 6.16 1.7 ?
*CAR: yyy
*MAI: 2.2 porque es que dices / es que a mí me pone los pelos de punta / ¿no? 2.2 //
*ANG: sí //
*MAI: 1.7 los padres de esta chica se quedan tan contentos / en Japón //
*ANG: 2.2 bueno no / tan contentos no / es que veintiún años ya / si no están contentos ... < va a ser peor > 2.2 //
*TTT: yyy
*PIZ: < les va a dar igual ¿no? > //
*MAI: < y la niña / dando tumbos > ... había estado / en otros viajes en Australia / en el norte de Marruecos / en París / en no sé qué / dices / realmente es que para estos el mundo es todo // o sea su [///] su ámbito es todo // agarran la maleta se suben en un avión y < ¡hala! / por todas partes > 1.7 //
*ANG: < se largan > //
*XYZ: xxx

7.2 Tablas

<table>
<thead>
<tr>
<th>Géneros discursivos y porcentaje de tiempo en C-Or-DiAL</th>
<th>Parámetros de clasificación de la espontaneidad de los textos</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conversaciones 24% conv</td>
<td>Lazos familiares o de intimidad 100%</td>
</tr>
<tr>
<td>Diálogos 5% dial</td>
<td>Lugar familiar (casa, café, bar, jardín) 100%</td>
</tr>
<tr>
<td>Entrevistas 54% entr</td>
<td>Papel determinado de los hablantes -</td>
</tr>
<tr>
<td>Charlas 5% char</td>
<td>Tema u objetivo preestablecido 2%</td>
</tr>
<tr>
<td>Fin predeterminado 2% finp</td>
<td></td>
</tr>
<tr>
<td>Trabajo 6% trab</td>
<td></td>
</tr>
<tr>
<td>Clases 5% aula</td>
<td></td>
</tr>
<tr>
<td>Conferencias 2% sala</td>
<td></td>
</tr>
</tbody>
</table>

| Tabla 1: Clasificación de espontaneidad de los géneros discursivos de C-Or-DiAL |
Número de palabras de cada tipología de texto

- charlas
- clases
- conferencias
- fin predeterminado
- trabajo
- entrevistas informales
- conversaciones
- diálogos

Tabla 2: Proporciones de palabras por género

Tabla 3: Sitio de C-Or-DiAL en LABLITA. Página de acceso directo a los archivos
Extension of the LECTRA corpus: classroom LECture TRAnscriptions in European Portuguese

Thomas PELLEGRINI, Helena MONIZ, Fernando BATISTA, Isabel TRANCOSO, Ramon ASTUDILLO

1Spoken Language Systems Lab, INESC-ID, Lisbon, Portugal; 2Faculdade de Letras, Universidade de Lisboa, Lisbon, Portugal; 3ISCTE-IUL - Instituto Universitário de Lisboa; 4Instituto Superior Técnico, Lisbon, Portugal

thomas.pellegrini@inesc-id.pt, helena.moniz@inesc-id.pt

Abstract

This paper presents the recent extension of the LECTRA corpus, a speech corpus of university lectures in European Portuguese that will be partially available. Eleven additional hours of various lectures were transcribed, following the previous multilayer annotations, and now comprising about 32 hours. This material can be used not only for the production of multimedia lecture contents for e-learning applications, enabling hearing impaired students to have access to recorded lectures, but also for linguistic and speech processing studies. Lectures present challenges for automatic speech recognition (ASR) engines due to their idiosyncratic nature as spontaneous speech and their specific jargon. The paper presents recent ASR experiments that have clearly shown performance improvements on this domain. Together with the manual transcripts, a set of upgraded and enriched force-aligned transcripts was also produced. Such transcripts constitute an important advantage for corpora analysis, and for studying several speech tasks.

Keywords: lecture domain speech corpus, ASR, speech transcripts, speech alignment, structural metadata, European Portuguese.

1. Introduction

This paper aims at a description of the corpus collected within the national project LECTRA and its recent extension. The LECTRA project aimed at transcribing lectures, which can be used not only for the production of multimedia lecture contents for e-learning applications, but also for enabling hearing-impaired students to have access to recorded lectures. The corpus has been already described in (Trancoso et al., 2008). We describe the recent extension of the manual annotations and the subsequent automatic speech recognition and alignment experiments to illustrate the performance improvements compared to the results reported in 2008. The extension was done in the framework of the METANET4U European project that aims at supporting language technology for European languages and multilinguality. One of the main goals of the project is that languages resources are made available online. Thus, the LECTRA corpus will be available through the central META-SHARE platform and through our local node: http://metanet4u.12f.inesc-id.pt/.

Lecture transcription can be very challenging, mainly due to the fact that we are dealing with a very specific domain and with spontaneous speech. This topic has been the target of much bigger research projects such as the Japanese project described in Furui et al. (2001), the European project CHIL (Lamel et al., 2005), and the American iCampus Spoken Lecture Processing project (Glass, 2007). It is also the goal of the Liberated Learning Consortium1, which fosters the application of speech recognition technology for enhancing accessibility for students with disabilities in the university classroom. In some of these projects, the concept of lecture is different. Many of our classroom lectures are 60-minute long, and quite informal, contrasting with the 20-minute seminars used in (Lamel et al., 2005), where a more formal speech can often be found.

After a short description of the corpus itself and the annotation schema in Sections 2 and 3 respectively, ASR experiments are reported in Section 4. Section 5 describes the creation of a dataset that merges manual and automatic annotations and that provides prosodic information. Section 6 presents the conclusions and the future work.

2. Corpus description

The corpus includes seven 1-semester courses: Production of Multimedia Contents (PMC), Economic Theory I (ETI), Linear Algebra (LA), Introduction to Informatics and Communication Techniques (IICT), Object Oriented Programming (OOP), Accounting (CONT), Graphical Interfaces (GI). All lectures were taught at Technical University of Lisbon (IST), recorded in the presence of students, except IICT, recorded in another university and in a quiet office environment, targeting an Internet audience. A lapel microphone was used almost everywhere, since it has obvious advantages in terms of non-intrusiveness, but the high frequency of head turning causes audible intensity fluctuations. The use of the head-mounted microphone in the last 11 PMC lectures clearly improved this problem. However, this microphone was used with an automatic gain control, causing saturation in some of the recordings, due to the increase of the recording sound level during the students’ questions, in the segments after them. Most classes are 60-90 minutes long (with the exception of IICT courses which are given in 30 minutes). A total of 74h were recorded, of which 10h were multilayer annotated in 2008 (Trancoso et al., 2008). Recently additional 11 hours were orthographically transcribed. Table I below shows the number of lectures per course and the audio duration that was annotated, where V1 corresponds to the 2008 version of the corpus, Added is the quantity of added data, and V2 corresponds to the extended actual version.

---

1 www.liberatedlearning.com
The orthographic manual transcriptions were done using Transcriber\(^2\) and Wavesurfer\(^3\) tools. Automatic transcriptions are used as a basis that the transcribers corrected. At this stage, speech is segmented into chunks delimited by silent pauses, already containing audio segmentation related to speaker and gender identification and background conditions. Previously, the annotation schema comprised multilayers of orthographic, morpho-syntactic, structural metadata (Liu et al., 2006; Ostendorf et al., 2008), i.e., disfluencies and punctuation marks, and paralinguistic information as well (laughs, coughs, etc.). The multilayer annotation aimed at providing a suitable sample for further linguistic and speech processing analysis in the lectures domain. The extension reported in this work respects the previous schema, however does not comprise the morpho-syntactic information tier, since automatic classifications of part-of-speech (POS) tags and of syntactic parsing is automatically performed, initially by

\( \text{Marv (Ribeiro et al., 2003) and more recently by Falaposta (Batista et al., 2012). Thus, the extension of the annotation comprises the full orthographic transcription, enriched with punctuation and disfluency marks and a set of diacritics fully reported in Trancoso et al. (2008). Segmentation marks were also inserted for regions in the audio file that were not further analyzed (background noise, signal saturation).} \)

Three annotators (with the same linguistics background) transcribed the extended data. However, two courses could not benefit from the extension for different reasons: the IICT, since no more lectures were recorded, and the ETI due to the fact that the teacher did not accept to make his recordings publicly available.

Due to the idiosyncratic nature of lectures as spontaneous and prepared non-scripted speech, the annotators reported in the five sessions of the guidelines instructions two main difficulties: in punctuating the speech and in classifying the disfluencies. The punctuation complexities are mainly associated with the fact that speech units do not always correspond to sentences, as established in the written sense. They may be quite flexible, elliptic, restructured, and even incomplete (Blauw, 1995). Therefore, to punctuate speech units is not always an easy task. For a more complete view on this, we used the summary of grammatical and ungrammatical locations of punctuation marks for European Portuguese described in Duarte (2000). The latter is related to the different courses and the difficulty in discriminating the specific types of disfluencies (if it is a substitution, for instance), since the background of the annotators is on linguistics. To sum up, the guidelines given to our annotators were: the schema described in Trancoso et al. (2008) and the punctuation summary described in Duarte (2000).

The general difficulty of measuring the inter-transcriber agreement is due to the fact that two annotators can produce token sequences of different lengths. This is equivalent to measuring the speech recognition performance, where the length of the recognized word sequence is usually different from the reference. For that reason, the inter-transcriber agreement was calculated for pairs of annotators, considering the most experienced\(^4\) as reference. The standard F1-measure and Slot Error Rate (SER) (Makhoul et al., 1999) metrics were used, where each slot corresponds to a word, a punctuation mark or a diacritic:

\[
F1 - \text{measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \quad \text{SER} = \frac{\text{errors}}{\text{ref_tokens}}
\]

where \(\text{ref_tokens}\) is the number of words, punctuation marks and diacritics used in the reference orthographic tier, and \(\text{errors}\) comprise the number of inserted, deleted or substituted tokens.

The inter-transcriber agreement of the three annotators is based on a selected sample of 10 minutes of

\(^2\)http://trans.sourceforge.net/
\(^3\)http://www.speech.kth.se/wavesurfer/
\(^4\)The annotator in question had already transcribed other corpora with the same guidelines.
speech from one speaker involving more than 2000 tokens. The selection of the sample has to do with the reported difficulties of the annotators, in annotating disfluencies (e.g., complex sequences of disfluencies) and also punctuation marks. Table 2 reports the inter-transcriber agreement results for each pair of annotators. The table shows the number of (Cor)rect slots, (Ins)ertions, (Del)etions, (Sub)stitutions, (F1)-measure, and slot accuracy (SAcc), which corresponds to 1-SER. There is an almost perfect agreement between A1 and the remaining annotators, and a substantial agreement between the pair A2-A3. These results may well be the outcome of a thorough process of annotation in several different steps and with intermediate evaluations during the 5 guidelines instruction sections. Moreover, several other annotators for other corpora already tested the guidelines here in use.

<table>
<thead>
<tr>
<th>Annotator</th>
<th>Cor</th>
<th>Ins</th>
<th>Del</th>
<th>Sub</th>
<th>F1</th>
<th>SER</th>
<th>SAcc</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1-A2</td>
<td>1714</td>
<td>67</td>
<td>79</td>
<td>224</td>
<td>0.852</td>
<td>0.184</td>
<td>0.816</td>
</tr>
<tr>
<td>A1-A3</td>
<td>1632</td>
<td>38</td>
<td>34</td>
<td>351</td>
<td>0.808</td>
<td>0.210</td>
<td>0.790</td>
</tr>
<tr>
<td>A2-A3</td>
<td>1480</td>
<td>81</td>
<td>97</td>
<td>444</td>
<td>0.735</td>
<td>0.308</td>
<td>0.692</td>
</tr>
</tbody>
</table>

Table 2: Evaluation of the inter-transcriber agreement

4. ASR experiments

Transcribing lectures is particularly difficult since lectures are very domain-specific and speech is spontaneous. Except the ICT lectures where no students were present, students demonstrate a relatively high interactivity in the other lectures. Nevertheless, since only a lapel microphone was used to record the close-talk speech of the lecturers, the audio gain of the student interventions is very low. The presence of background noise, such as babble noise, footsteps, blackboard writing noise, etc. may difficult the speech processing, in particular the Speech / Non-speech detection that feeds the recognizer with audio segments labelled as speech. Typical WER reported in the recent literature is between 40-45% (Glass et al., 2007).

4.1 Overview of our ASR system

Our automatic speech recognition engine named Audimus (Neto et al., 2008; Meinedo et al., 2008) is a hybrid automatic speech recognizer that combines the temporal modeling capabilities of Hidden Markov Models (HMMs) with the pattern discriminative classification capabilities of Multi-Layer Perceptrons (MLPs). The MLPs perform a phoneme classification by estimating the posterior probabilities of the different phonemes for a given input speech frame (and its context). These posterior probabilities are associated to the single state of context independent phoneme HMMs.

The most recent ASR system used in this work is exactly the ASR system for EP described in (Meinedo et al., 2010). The acoustic models were initially trained with 46 hours of manually annotated broadcast news (BN) data collected from the public Portuguese TV, and in a second time with 1000 hours of data from news shows of several EP TV channels automatically transcribed and selected according to a confidence measure threshold (non-supervised training). The EP MLPs are formed by 2 hidden layers with 2000 units each and have 500 softmax output units that correspond to 38 three state monophones of the EP language plus a single-state non-speech model (silence) and 385 phone transition units which were chosen to cover a very significant part of all the transition units present in the training data. Details on phone transition modeling with hybrid ANN/HMM can be found in (Abad & Neto, 2008).

The Language Model (LM) is a statistical 4-gram model that was estimated from the interpolation of several specific LMs: in particular a backoff 4-gram LM, trained on a 700M word corpus of newspaper texts, collected from the Web from 1991 to 2005, and a backoff 3-gram LM estimated on a 531k word corpus of broadcast news transcripts. The final language model is a 4-gram LM, with Kneser-Ney smoothed modified, 100k words (or 1-gram), 7.5M 2-gram, 14M 3-gram and 7.9M 4-gram. The multiple-pronunciation EP lexicon includes about 114k entries.

These models, both AMs and the LM, were specifically trained to transcribe BN data. The Word Error Rate (WER) of our current ASR system is under 20% for BN speech in average: 18.4% for instance, obtained in one of our BN evaluation test sets (RTP07), composed by six one hour long news shows from 2007 (Meinedo et al., 2010).

4.2 ASR results

A test subset was selected from the corpus in 2008, by choosing one single lecture per course. In (Trancoso et al., 2008), preliminary ASR results were reported on this test set, showing the difficulty to transcribe lectures. Very high word error rates (WER), 61.0% in mean, were achieved for a subset of various lectures chosen as a test set. It has a vocabulary of around 57k words. Applying this recognize without any type of domain adaptation, obviously yielded very bad results.

Table 3 illustrates the performance of the old and the recent systems without and with adaptation of the LM for the recent system. Our recent system, which was described in the previous section, achieved a WER of 45.7% on the same test subset, hence showed a 25.0% relative reduction. The lexicon was almost twice the size of the one of the previous system. Further improvements were achieved with a 44.0% WER. This performance was obtained by interpolating our generic broadcast news 4-gram LM with a 3-gram LM trained on the training lecture subset. 100-best hypotheses were generated per each sentence and rescored with this LM and a RNN (implementation of the Brno University (Mikolov et al., 2011)). This RNN was trained only on the lecture train subset.

An analysis of the ASR errors showed that most of the misrecognitions concerned small function words, such as definite articles and prepositions, the backchannel word “OK” also appeared to be very often misrecognized. Then, words specific to each jargon of the courses also were error-prone. For instance variable names in the Linear
Algebra lecture, such as “álfa”, “beta”, “vector” were often substituted. In the PMC lecture, words such as “MPEG”, “codecs”, “metadados” (metadata), “URL” were subject to frequent errors.

<table>
<thead>
<tr>
<th>ASR system</th>
<th>LM adapt?</th>
<th>OOV (%)</th>
<th>WER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>no</td>
<td>_</td>
<td>61.0</td>
</tr>
<tr>
<td>2011</td>
<td>no</td>
<td>2.8</td>
<td>45.7</td>
</tr>
<tr>
<td>2011</td>
<td>yes</td>
<td>1.7</td>
<td>44.0</td>
</tr>
</tbody>
</table>

Table 3: Comparison of the ASR results reported in 2008 and obtained with our most recent system. OOV stands for out-of-vocabulary words.

5. Enriched annotations

The ASR system is able not only to produce automatic transcripts from the speech signal, but also to produce automatic force-aligned transcripts, adjusting the manual transcripts to the speech signal. Apart from the existing manual annotations of the corpus, automatic force-aligned transcripts have been produced for the extended version of the corpus, and will be available in our META-SHARE node. These force-aligned transcripts were updated with relevant information coming from the manual annotations, and finally enriched with additional prosodic information (Batista et al., 2012). The remainder of this Section provides more details about this process.

5.1 Automatic alignment

Force-aligned transcripts depend on a manual annotation and therefore do not contain recognition errors. A number of speech tasks, such as the punctuation recovery, may use information, such as pause durations, which most of the times is not available in the manual transcripts. On the other hand, manual transcripts provide reduced or error-free transcripts of the signal. For that reason, force-aligned transcripts, which combine the ASR information with manual transcripts, provide unique information, suitable for a vast number of tasks.

An important advantage of using force-aligned transcripts is that they can be treated in the exact same way as the automatic transcripts, but without recognition errors, requiring the same exact procedures and tools. However, the alignment process is not always performed correctly due to a number of reasons, in particular when the signal contains low energy levels. For that reason, the ASR parameters can be adjusted to accommodate the manual transcript into the signal. Our current force-alignment achieves 3.8% alignment word errors in the training, 3.1% in the development, and 4.5% in the evaluation sets.

5.2 Merging manual and automatic annotations

Starting with the previously described force-aligned transcripts, we have produced a self-contained dataset that provides not only the information given by the ASR system, but also important parts of the manual transcripts. For example, the manual orthographic transcripts include punctuation marks and capitalization information, but that is not the case of force-aligned transcripts, which only includes information, such as: word time intervals, and confidence scores. The required manual annotations are transferred by means of alignments between the manual and automatic transcripts.

Apart from transferring information from the manual transcripts, the data was also automatically annotated with part-of-speech information. The part-of-speech tagger input corresponds to the text extracted from the ASR transcript, after being improved with the reference capitalization. Currently, the Portuguese data is being annotated using Falaposta, a CRF-based tagger robust to certain recognition errors, given that a recognition error may not affect all its input features. It accounts for 29 part-of-speech (POS) tags and achieves 95.6% accuracy.

The resulting file, structured using the XML format, corresponds to the ASR output, extended with: time intervals to be ignored in scoring, focus conditions, speaker information for each region, punctuation marks, capitalisation, disfluency marks, and POS information.

5.3 Adding prosodic data

The previously described extended XML file is further improved with phone and syllable information, and other relevant information that can be computed from the speech signal (e.g., pitch and energy). The data provided by the ASR system allows us to calculate the phone information. Marking the syllable boundaries as well as the syllable stress are achieved by means of a lexicon containing all the pronunciations of each word together with syllable information, since these tasks are currently absent in the recognizer. A set of syllabification rules was designed and applied to the lexicon, which account fairly well for the canonical pronunciation of native words, but they still need improvement for words of foreign origin. Pitch ($f_0$) and energy ($E$) are two important sources of prosodic information, currently not available in the ASR output, and directly extracted from the speech signal. Algorithms for automatic extraction of the pitch track have, however, some problems, e.g., octave jumps; irregular values for regions with low pitch values; disturbances in areas with micro-prosodic effects; influences from background noisy conditions; inter alia. We have removed all the pitch values calculated for unvoiced regions in order to avoid constant micro-prosodic effects. This is performed in a phone-based analysis by detecting all the unvoiced phones. We also had a calculation cost to eliminate octave-jumps. As to the influences from noisy conditions, the recognizer has an Audio Pre-processing or Audio Segmentation module, which classifies the input speech according to different focus conditions (e.g., noisy, clean), making it possible to isolate those speech segments with unreliable pitch values.

After extracting and calculating the above information, all data was merged into a single data source. The existing XML data has been upgraded in order to accommodate the additional prosodic information.
6. Conclusions

This paper described our lecture corpus in European Portuguese, and its recent extension. The problems it raises for automatic speech recognition systems were illustrated. The fact that a significant percentage of the recognition errors occurs for function words led us to believe that the current performance, although far from ideal, may be good enough for information retrieval purposes, enabling keyword search and question answering in the lecture browser application. ASR performance is still poor but as stated in Glass et al. (2007), “accurate precision and recall of audio segments containing important keywords or phrases can be achieved even for highly-errorful audio transcriptions (i.e., word error rates of 30% to 50%)”. Together with the manual transcripts, a set of upgraded and enriched force-aligned transcripts were produced and made available. Such transcripts constitute an important advantage for corpora analysis, and for studying a number of speech tasks. Currently, the LECTRA corpus is being used to study and perform punctuation and capitalization tasks, and spontaneous speech phenomena. We believe that producing a surface rich transcription is essential to make the recognition output intelligible for hearing impaired students. Six courses of the corpus will be soon available to the research community via the META-SHARE platform.

7. Acknowledgements

This work was partially funded by the European project METANET4U number 270893, by national funds through FCT – Fundação para a Ciência e a Tecnologia, under project PEst-OE/EEI/LA0021/2011, and by DCTI - ISCTE-IUL.

8. References


A constituição de um corpus de italiano falado para o estudo de pedidos e pedidos de desculpas: considerações sobre a validade interna e externa dos dados

Elisabetta SANTORO
Universidade de São Paulo
Av. Prof. Luciano Gualberto, 403 Cidade Universitária
CEP 01060-970 – São Paulo – SP – Brasil
esantoro@usp.br

Resumo

O texto aqui apresentado pretende discutir questões ligadas à constituição de um corpus de italiano falado, coletado a partir de gravações em áudio e vídeo, para um estudo que se insere no âmbito da pragmática linguística e procura investigar dois atos de fala específicos, a saber, pedidos e pedidos de desculpas. Propõe-se, em especial, uma reflexão sobre a validade externa e a validade interna dos dados, sendo que, a partir desses conceitos, será possível pensar nas características das pesquisas realizadas com dados coletados a partir de diferentes metodologias, além de se poder imaginar uma “hierarquia” de metodologias, da mais livre à mais controlada. Se, por um lado, metodologias muito abertas permitem uma elevada validade externa dos dados, mas não são muito adequadas para o estudo de fenômenos específicos, além de ser também dificilmente replicáveis; por outro, metodologias nas quais a produção dos informantes é mais controlada podem produzir dados mais facilmente comparáveis e ajudar a circunscrever aspectos específicos da língua.

Palavras-chave: pragmática linguística; corpus; metodologias de coleta de dados; role play.

1. Corpus e língua falada

Realizar uma pesquisa a partir de um corpus de língua falada pressupõe decisões importantes sobre a metodologia de coleta dos dados, pois, antes mesmo de iniciar o planejamento do trabalho, é preciso avaliar com extrema atenção vantagens e desvantagens de cada uma das possibilidades. Se o objetivo da pesquisa for, por exemplo, estudar a língua falada sob diferentes pontos de vista (da fonética, da fonologia, da prosódia, do léxico, da morfologia, da sintaxe etc), será essencial dispor de material linguístico que seja diastreticamente e diafasicamente o mais variado possível, para que se possam fazer afirmações que, mesmo dizendo respeito às amostras de língua coletadas, possam “representar” o todo. Quando, ao contrário, o pesquisador estabelece metas mais detalhadas e pretende se dedicar a fenômenos específicos da língua falada, pode ser necessário utilizar metodologias que deem subsídios de outra natureza para a análise a ser desenvolvida.

Pretendemos aqui discutir brevemente algumas das alternativas que se colocam para o pesquisador, pensando, em especial, nas escolhas feitas para um estudo realizado com o italiano contemporâneo, que se insere no âmbito das pesquisas em pragmática linguística e procura investigar dois atos de fala específicos, a saber, pedidos e pedidos de desculpas, a partir de gravações em áudio e vídeo.

Poderia ser útil – e trazer ainda outras questões, entre as quais nem sempre há consenso entre os pesquisadores – analisar também as definições de corpus, inclusiva colocando-as em relação com os objetivos de pesquisa e com o tipo de análise a ser realizado. No entanto, não faremos isso aqui e iremos nos concentrar em considerações relativas à validade interna e externa dos dados coletados, para podermos refletir sobre as diferentes abordagens e metodologias que impedem ou, ao contrário, permitem a execução de um determinado tipo de pesquisa.

2. Validade externa e validade interna

Em primeiro lugar, cabe explicar o que entendemos quando falamos de validade externa e de validade interna dos dados. Começando pela validade externa, podemos dizer que esta se julga dada, quando é possível generalizar os resultados de uma pesquisa, que, a partir das amostras escolhidas, podem ser considerados válidos para a língua em análise como um todo. Para tanto, reputa-se imprescindível gravar os informantes em situações que eles não sintam como “estranhas”, isto é, que não sejam distantes de sua habitual prática linguística.

A validade interna, ao contrário, refere-se à interpretabilidade da pesquisa e deve permitir dizer se as variações presentes nos dados podem ser tratadas como uma consequência das variáveis analisadas. A validade interna está relacionada aos fatores que podem influenciar diretamente os resultados e é avaliada levando em conta se as diferenças encontradas na variável dependente (que medimos para ver quais são os efeitos da variável independente sobre ela), se relacionam diretamente com a variável independente (aquela que pode “causar” o resultado). A validade interna implica, portanto, que os dados sejam mais controlados, e precisa de instrumentos de coleta que permitam isolar variáveis de modo a garantir sua adequada avaliação separadamente e em sua interação com outras.

Há muitos fatores que podem comprometer a validade interna dos dados de uma pesquisa, entre os quais, por exemplo, as características e o comportamento dos participantes, o equipamento utilizado, a atitude do pesquisador que coleta os dados e a situação em que isso é feito.

Além disso, é importante não esquecer que, em geral, estudos com elevada validade externa sofrem em relação à validade interna, porque o respeito à integridade do contexto impede que sejam controladas as variáveis –
como é possível fazer, por exemplo, com um protocolo experimental – assim que afirmações de natureza causal ou o estabelecimento de relações entre os dados serão sempre problemáticos ou até impossíveis.

Tendo isso em vista, no caso de uma pesquisa que visa a estudar a realização de atos de fala específicos na interação entre dois falantes e que se propõe a descobrir eventuais relações entre variáveis, para poder entender o funcionamento de uma determinada língua natural em uso, uma vez controladas as características ligadas aos informantes, será necessário abdicar, ao menos em parte, da validade externa da fala espontânea não controlada e procurar metodologias de coleta dos dados que permitam também análises causais, que relacionem os dados.

3. Corpora no estudo da pragmática linguística

Em muitas pesquisas que se propõem a constituir corpora para a investigação da pragmática linguística, especialmente quando ligadas à mais ortodoxa análise da conversação1, atenta-se principalmente para a validade externa dos dados a serem estudados, isto é, procura-se coletá-los de modo que haja a maior correspondência possível entre os fenômenos observados ao longo da investigação e os que acontecem, ou se presume aconteçam, na vida real. Em outras palavras, os dados considerados de maior relevância para o estudo da pragmática das línguas, principalmente em contextos cotidianos, são os ditos “dados naturalísticos”, coletados de preferência sem que o informante tenha ciência, no momento em que os fornece, de participar de uma pesquisa e, possivelmente, sobretudo no caso de gravações só em áudio, com os aparelhos escondidos, de modo que o informante nem mesmo saiba que sua fala está sendo gravada. É o caso das ditas “gravações secretas”, nas quais se procuram voluntários dispostos a colaborar nas pesquisas, que, em geral, gravam conversações das pessoas de seu convívio, revelando só depois de concluída a gravação sua participação no projeto.

Não citaremos aqui as questões éticas e legais que procedimentos como esses envolvem (para isso, sugerimos, por exemplo, a leitura de Bazzanella, 1994). Embora isso não seja considerado admissível por alguns, pois alteraria por si só a validade e a confiabilidade dos dados, basta que os informantes sejam avisados e aceitem ser gravados – como acontece nas gravações que chamaremos “consentidas” – para que essa dificuldade seja superada.

Mesmo assim, os dados produzidos a partir desse tipo de metodologia podem ser, segundo alguns, menos “naturais”, pois os informantes, ao sabermos que estão sendo gravados, alterariam sua fala. É preciso, contudo, lembrar que a própria definição de dado naturalístico não é isenta de problemas.

De fato, é suficiente pensar nas observações sobre o “paradoxo do observador” de Labov (1970) ou nos questionamentos de Ochs (1979) sobre a impossível neutralidade do processo de transcrição, para concluir que a realidade linguística não poderá nunca ser colhida em toda a sua complexidade e que o pesquisador sempre irá intervir para recortar do material coletado as partes mais significativas para o seu projeto de pesquisa, eliminando em alguns casos o contexto e produzindo, assim, alterações que também deveriam ser levadas em conta.

Desta forma, do nosso ponto de vista, pode ser por vezes desmedida a atenção dada à definição do que se pode considerar fala espontânea ou semi-espontânea: se acreditarmos que a gravação e a transcrição em si já alteram o contexto da fala e precisariam, portanto, ser levadas em conta na hora de analisar os dados, deveríamos também relativizar a rigidez que muitas vezes acompanha o julgamento das maneiras como foram coletados.

Não obstante, é claro que há distinções a serem feitas entre as possíveis maneiras de eliciar dados e que é necessário ter consciência de quais são, sempre atentando também para os objetivos de cada pesquisa. Como dizíamos no início, diferente será, por exemplo, coletar um corpus com o objetivo de estudar fenômenos gerais da língua, não ligados a situações peculiares e importantes pela sua recorrência em diferentes contextos comunicativos; ou tentar delimitar e fixar na gravação o mesmo fenômeno que se repete diversas vezes, de modo que suas manifestações, em contextos de partida idênticos, possam ser comparadas e estudadas.

Na hipótese em que se queira, como no exemplo da pesquisa de que falamos aqui, verificar se o mesmo pedido é realizado com o uso de formas linguísticas distintas, caso intervenha uma determinada variável, será necessário controlar a variável escolhida e comparar o maior número possível de ocorrências realizadas a partir do mesmo input. É evidente que isso só poderá ser feito se os dados forem coletados com metodologias que prevejam o controle das variáveis e ser praticamente impossível com gravações “livres”.

Visando a contribuir para uma maior clareza sobre as diferenças na coleta de dados para o estudo da língua falada, há estudiosos que prepararam listas e propuseram hierarquizações das metodologias, colocando-as em uma ordem que vai do menor ao maior grau de controle sobre a produção dos dados, isto é, da maior validade externa à maior validade interna (se veja, Pallotti, 2001).

Tentaremos fazer aqui algo parecido, refletindo, em especial, sobre as pesquisas relativas ao estudo da pragmática linguística, intercultural e interlinguística.

Citemos acima dois procedimentos que se propõem a “capturar” a realidade linguística assim como ela é e que se impõem para esse fim várias e, muitas vezes, rígidas limitações metodológicas.

Pensando ainda em termos de validade externa e interna, podemos observar que no outro extremo em relação às metodologias mencionadas acima, em especial quando a perspectiva é a da pragmática intercultural ou interlinguística, é prática comum coletar os dados utilizando instrumentos que possuem um elevado grau de

controle sobre as variáveis. São, de fato, frequentes os casos nos quais, para a coleta dos dados, se escolhem DCT (Discourse Completion Tests) escritos, nos quais os informantes, utilizando-se da escrita para fornecer dados que deveriam pertencer à oralidade, escrevem o que diriam em determinadas situações (ver, por exemplo, Hudson, Detmer & Brown, 1995) ou até realizam atividades de escolha múltipla, em que o informante deve apenas assinalar qual das alternativas apresentadas considera mais adequada para responder à situação comunicativa descrita.

<table>
<thead>
<tr>
<th>Controle mínimo sobre as produções dos informantes (elevada validade externa)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gravação secreta</td>
</tr>
<tr>
<td>Gravação consentida</td>
</tr>
<tr>
<td>Gravação participante</td>
</tr>
<tr>
<td>Role play aberto</td>
</tr>
<tr>
<td>Role play semi-aberto</td>
</tr>
<tr>
<td>Role play fechado</td>
</tr>
<tr>
<td>Discourse Completion Test oral</td>
</tr>
<tr>
<td>Discourse Completion Test escrito</td>
</tr>
<tr>
<td>Escolha múltipla</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Controle máximo sobre as produções dos informantes (elevada validade interna)</th>
</tr>
</thead>
</table>

Tabela 1: Algumas metodologias para coleta de dados

Com essas metodologias as produções dos informantes são muito controladas e, além disso, os dados assim coletados requerem um baixo dispêndio de tempo e energias, pois não precisam de equipamentos de gravação em áudio e vídeo e podem ser gerados em grande número até em uma única sessão.

É evidente que metodologias dessa natureza afastam demasiadamente os dados coletados daqueles que se considerariam “naturais”. De fato, tentando reproduzir por escrito aquilo que diriam nas situações dadas os informantes eliminam completamente os traços característicos da língua falada (falsas partidas, reformulações, hesitações etc) e “limpam” suas manifestações pragmáticas de todos os elementos que caracterizam a fala. Além disso, escrever no lugar de dizer significa eliminar completamente a interação oral entre dois indivíduos; e dispor de um tempo maior antes de produzir o dado o priva da imediatidade característica da língua falada, na qual se reage a um estímulo oral, sem ter a chance de refletir ou de se preparar.

Citamos acima apenas as metodologias mais livres, de um lado, e mais controladas, do outro. A seguir, apresentamos nossa proposta de uma escala de metodologias para a coleta dos dados, pensada a partir das escolhas mais frequentes feitas para estudos de pragmática linguística. As metodologias foram hierarquizadas de modo que a menos controlada e com maior validade externa foi colocada na parte superior da tabela, enquanto a mais controlada e com menor validade externa na parte inferior.

4. Um corpus de italiano falado para o estudo de pedidos e pedidos de desculpas

Para a constituição de um corpus de italiano falado que se propõe a analisar pedidos e pedidos de desculpas, optamos por uma metodologia de coleta dos dados que se coloca na posição intermediária da tabela apresentada acima. Trata-se do role play semi-aberto que, em relação a outras opções controladas de coleta de dados, possui, para começar, a vantagem de criar uma verdadeira interação oral entre dois interlocutores, mantendo, portanto, as características da língua falada, embora a interação não seja a consequência de uma necessidade real dos interlocutores e seja induzida pelo pesquisador.

Em geral, a distinção se faz apenas entre role play aberto, que envolve a interação entre dois ou mais indivíduos, reagindo a uma determinada situação; e role play fechado, nos quais é apresentada aos participantes uma situação específica, à qual devem responder, em geral, com um único turno de fala. Considera-se que o role play fechado pode não refletir dados que poderiam ocorrer naturalmente, enquanto o aberto os refletiria mais exatamente, por prever a interação e uma reação “livre” à situação dada. Para o tipo de role play utilizado na nossa pesquisa preferimos utilizar a categoria “role play semi-aberto”, pois aos interlocutores foi pedido que reagissem a uma situação comunicativa específica e em um contexto dado.

2 A esse respeito Mackey & Gass (2005: 91) afirmam: “Open role plays, on the other hand, involve interaction played out by two or more individuals in response to a particular situation. [...] Closed role plays suffer from the possibility of not being a reflection of naturally occurring data. Open role plays reflect natural data more exactly [...].”
As instruções eram dadas por escrito a apenas um dos dois participantes ao qual cabia também iniciar a interação. Isso foi feito com o objetivo de o input ser idêntico para todos os participantes da pesquisa que receberam as instruções com o mesmo papel e, portanto, com exatamente as mesmas palavras e do mesmo modo. Além disso, essa escolha permitiu que um dos dois informantes reagisse diretamente à fala do outro, sem saber antes a que situação teria que reagir.

A decisão de como transformar em palavras e em interação com o outro a situação descrita no papel recebido era “livre” quanto às formas linguísticas e sem limites de tempo ou turnos de fala limitados. Procuramos ainda, sempre que possível, recrear o contexto (setting)³, para que os informantes pudessem mais facilmente evocar as rotinas linguísticas utilizadas em situações do mesmo tipo. Assim, as situações que aconteciam “na rua” foram efetivamente gravadas na rua e assim também foram gravadas em casas as situações do contexto “casa de outrem”. Com um fim parecido, procuramos também definir contextos e situações nos quais todos os informantes poderiam se encontrar na vida real, de modo a não levá-los a representar um “papel” no qual dificilmente se encontrariam na vida real. Mais um corretivo aplicado ao role play é que os interlocutores não mudaram sua relação da vida real e se trataram no role play como se tratariam fora dele.

Com essa metodologia poderiam ser controladas as variáveis independentes. Brown e Levinson (1987: 76) identificam três variáveis para os atos de fala: a distância social entre os interlocutores que cria um eixo horizontal; o poder relativo entre eles, a partir do qual se estabelece um eixo vertical; e o grau de imposição de um ato de fala, ou seja, a relação custo/benefício que a realização do ato representa para os interlocutores.

No nosso caso, se é verdade que a escolha quanto a respeitar a identidade e a relação real entre os informantes limitou ou até impossibilitou a seleção de situações com claras diferenças de poder relativo e distância social (para imaginá-las teria sido necessário pensar em contextos como o ambiente de trabalho, nos quais isso é mais evidente), é também verdade que o grau de imposição, variável que pode produzir notáveis diferenças nos atos de fala, pôde ser incluído. De fato, a um maior grau de imposição corresponde em geral um aumento da atenuação, dos modificadores, da necessidade de justificar um pedido ou de procurar reparar o prejuízo provocado no caso de pedidos de desculpas.

Procuramos, portanto, organizar as situações previstas para os role plays em pares, nos quais sempre havia uma situação com um baixo grau de imposição (1), isto é, com um pedido ou um pedido de desculpas que previa um ônus baixo para o interlocutor, e outra, no mesmo contexto, com alto grau de imposição (4), ou seja, com um ônus elevado. Cabe dizer que, para garantir que fosse claro o diferente grau de imposição entre os pares de situações no mesmo contexto, procuramos escolher pedidos e pedidos de desculpas em que as diferenças fossem muito marcantes. Assim, por exemplo, no contexto “casa de outrem”, o pedido –I do informante que chega à casa de outra pessoa é um copo d’água, enquanto o pedido +I é poder tirar a roupa molhada e tomar um banho, porque a pessoa que chega foi surpreendida por um forte temporal e estava sem guarda-chuva.

Os contextos em que as situações foram colocadas eram ao todo três – a rua, o trem e a casa de outrem – e havia, para cada um deles, dois pedidos e dois pedidos de desculpas, chegando-se assim a 12 situações gravadas pelas 30 duplas de informantes que participaram da pesquisa e realizaram interações orais a partir do mesmo input.

Vale acrescentar que foi considerado na elaboração dos role plays que haveria diferentes graus de familiaridade entre os participantes e foi assim decidido dividí-los em duas grandes categorias, tratando como um grupo os que declararam ter um grau de conhecimento de 1 a 5 (desconhecidos, conhecidos, pessoas que acabaram de se conhecer), e cada um segundo grupo os com um grau de conhecimento de 6 a 10 (amigos ou parentes).

Apenas para os pedidos realizamos também gravações em estabelecimentos públicos e comerciais, de três diferentes cidades italianas, nos quais pudemos contar com a participação das pessoas que habitualmente atendem o público. Para essas gravações foi dada aos informantes uma instrução oral reduzida ao essencial para que pudessem realizar a ação prevista (do tipo: “entre na loja e compre um presente”).

Além de permitir o controle das variáveis e, portanto, uma validade interna elevada que possibilita um estudo sistemático das ocorrências, o corpus coletado é caracterizado pela replicabilidade e pela possibilidade de ser ampliado. Pretendemos, de fato, constituir um corpus com as mesmas características para o português brasileiro que possibilite a realização de estudos de pragmática intercultural, comparando a realização dos mesmos atos de fala por falantes nativos de italiano e de português brasileiro. Foram iniciadas também coletas de dados e pesquisas com aprendizes brasileiros de italiano que poderão representar a base para analisar a pragmática interlinguística, isto é, como um aprendiz brasileiro desenvolve sua competência pragmática em italiano, que tipo de relação essa competência possui com os conhecimentos gramaticais, e se as instruções explícitas podem ter efeitos reconhecíveis.

5. Conclusões

Para um projeto dessa natureza, o role play, realizado com os corretivos antes mencionados, representou uma forma de coletar dados que, por um lado, permitiu conservar as peculiaridades da língua falada e, por outro, ofereceu a possibilidade de isolar variáveis e analisar as alterações que poderiam ser provocadas por cada uma delas. Isso significou criar um corpus com características homogêneas, capaz de fornecer primeiros dados comparáveis para o estudo de pedidos e pedidos de desculpa. Os role plays gravados em áudio e vídeo pelo

mesmo par de informantes, mas com diferentes graus de imposição nos permitem observar a maior e menor presença, por exemplo, de modificadores e atenuadores, ou a presença/ausência de uma justificativa para um pedido ou para um pedido de desculpas e isso pode nos ajudar a reonduzir as escolhas a variáveis pré-determinadas, dando-nos assim a possibilidade de identificar as prováveis “causas” de específicas manifestações linguísticas e fornecendo-nos dados para criar relações entre o mundo e a língua.

6. Agradecimentos
A pesquisa à qual nos referimos no texto foi realizada durante um estágio pós-doutoral na Itália financiado pela CAPES. Foi de grande valia a colaboração com o grupo de pesquisa que trabalha no projeto Lira (Lingua e cultura Italiana in Rete per l’Apprendimento) e, em especial, com o Prof. Gabriele Pallotti da Università di Modena e Reggio Emilia.

7. Bibliografia
SPEECH TECHNOLOGY AND DATA BASES
SweDia 2000 – A Swedish dialect research database

Anders ERIKSSON
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg, Box 200, SE-405 30 Gothenburg, Sweden
anders.eriksson@ling.gu.se

Abstract

The SweDia 2000 dialect database (SweDat as we refer to it in our daily work) is a speech database containing recordings of Swedish dialects from all over Sweden and Swedish speaking communities in Finland. The database contains recordings of at least 12 speakers per dialect from 107 locations. A little over 1300 speakers have been recorded and the total recording time is about 800 hours. Each dialect is represented by two generations of speakers, an older generation 55–75 years of age and a younger generation 20–35 years of age. Each age group is represented by an equal number of male and female speakers. The data is organised in two separate databases – one publicly available database containing four short samples from each dialect and primarily intended for educational purposes, and a research database containing the entire material but with access rights limited to researchers. In this paper we will describe the criteria behind the selection of locations, speech types etc., the collection of data, the linguistic structure and properties of the database, examples of how the material is used and finally what we are presently doing to preserve the data for future generations of researchers.

Keywords: dialect; database; e-science.

1. Introduction

The SweDia 2000 database (we often refer to it as SweDat) is the result of two project efforts. The first project, SweDia 2000 – Phonetics and phonology of the Swedish dialects around the year 2000, was funded by the Bank of Sweden Tercentenary Foundation (grant 1997-5066/01/02 and ran between 1998 and 2003. During this period all the data was collected and a first version of the database set up. The goal of the present work on the database is to update data formats and to make the database available to the research community over the Internet. This is done within a follow-up project, SweDia 2000 – A Swedish dialect database, funded by The Swedish Research Council (grant 825-2007-7432) for the period 2007–2011.

In the following we will describe the considerations behind the selection of recording sites and the data collection procedure itself. The general properties and linguistic structure of the database will then be described and finally a description of the state of development we are in now and examples of the many different uses of the database for education and research.

2. General considerations

The goal was not, as is often the case in traditional dialectology, to find the most archaic samples of the selected dialects, but to collect samples representative of the linguistic varieties used in the daily lives of socially active people in the selected speech communities.

The chosen recording sites are evenly distributed over Sweden and the Swedish speaking communities in Finland taking into account both geographical dispersion and population density. The selection was done in close co-operation with dialect experts at the Swedish and Finnish dialect archives. Where there was more than one site that fulfilled the above mentioned two criteria, the site was chosen based on the amount of earlier material available in the dialect archives in order to maximize the possibility of historical comparisons. Only rural dialects were considered, no major towns are included in the data. The reason behind this decision was that the driving forces behind language change are quite different in the rural communities and major cities. In the cities change is driven by the influx of new inhabitants from other linguistic areas whereas the situation in the rural communities is almost the opposite, here non-mobility has been the major factor. Another consideration was the fact that rapid linguistic levelling is going on in many smaller communities and we wanted to capture the situation before that levelling had gone too far.

3. Data collection

The bulk of recordings were made during the summer of 1999. But a number of preliminary recordings were made already during 1998. These recordings were made to test the procedures with respect to recording techniques, interview types and logistics (travel arrangements, lodging facilities, time consumption etc.).

We were also not quite sure how many recordings per site would be necessary in order to control for inter-speaker variation. So the choices were more recordings per site and fewer sites or fewer recordings and more sites. Our goal was to collect data from two age groups, young adults aged 20–35 years of age and an older generation 55–75 years of age, and an equal number of male and female speakers in each age group. In the trial round we tested two alternatives, 5 or 3 subjects per group, that is a total of 20 or 12 speakers per location. Subsequent analyses indicated that 12 speakers per location would be sufficient. When all relevant factors had been considered the decision was to collect data from 107 different recording locations including the ones recorded in the trial round (see Appendix!).

As was mentioned above the goal of the project was to collect data representative of the speech used by socially active people in their daily lives. We therefore required that the participants should either still be working or should take active part in the social activities in their communities in some other way. For the younger
informants we also required that they should be second generation native speakers of the dialect. This was not a formal requirement for the older generation, but it turned out that most of them met the requirement anyway.

The plan was to record most of the dialects during the summer holidays of 1999. To be able to accomplish this, very careful preparations were of the essence. According to the time plan, a site should be completed in one work week and there was really no margin of error. For this to be possible everything had to be prepared well in advance. Data collection was made by linguistics students at the universities of Umeå, Stockholm and Lund. They were recruited in the beginning of 1999 and spent the spring term of 1999 planning the work. Informants were recruited via municipal organisations, social clubs etc. When the recordings began, 12 speakers per location had already been contacted and agreed to participate. We also had a few extra contacts in case anyone should be unable to participate, for example due to illness.

The field teams were recruited mainly among the students who had been responsible for the preparations. A team consisted of two students working together, taking turns in interviewing and handling the recording equipment. They had a rented car at their disposal, a credit card for expenses, Digital Audio Tape (DAT) recorders with lapel microphones, a mobile phone to manage contacts and a lap top computer for making notes.

They had all been thoroughly trained for the task both in terms of interviewing techniques and handling of the equipment. The students performing the field work were not generally native speakers of the dialects of the informants but often spoke some similar dialect. For some of the more deviant dialects we chose, however, to recruit students who were themselves native speakers of the dialect.

4. General properties of the database

The SweDia 2000 database has some properties, which as far as we are aware, are not common in otherwise comparable databases.

**Synchronicity:** All recordings were made within a narrow and precisely defined time slice. They therefore represent the dialectal variation at a precisely defined moment in time.

**Consistency:** The material has three well controlled parts that represent three fundamental, phonological properties – the quantity system, the accent system, and the phoneme inventory. It is thus possible to analyze and compare speech material of identical types for all dialects.

**Completeness:** The recorded material also contains about 30 minutes of spontaneous speech per speaker. This gives us additional information about how observed phonological rules are realized in everyday speech. It may also be used for other types of studies; for example studies of syntax and morphology (see below!).

5. Linguistic structure of the database

Linguistically, the database may be divided into two major parts – structurally controlled material and semi spontaneous speech.

The data in the controlled material consist of words or phrases repeated 3–5 times exemplifying the phoneme inventory of the dialect, the phonetic realization of quantity, and certain prosodic properties (word stress, tonal accent and phrasal focus).

The part intended for phoneme inventory analyses contained everyday words which could be assumed to have existed in the dialect (albeit not necessarily with the same pronunciation) for a very long time (several hundred years). The word lists were constructed in close co-operation with experts on historical dialectology from the departments of Swedish and researchers at the national dialect archives in Sweden and Finland.

The quantity word lists consist of minimal word pairs differing in quantity only. Old Swedish had a four-way quantity system (V:C, V:C, VC:, VC). In modern Swedish only two of the contrasts are still used (V:C and VC). There are, however, several dialects that still have a three-way system where VC is also contrastive. Many such examples exist in the recordings of semi spontaneous speech but unfortunately the quantity word list did not include such examples.

Swedish is not a tone language in the strictest sense of the term, but has nevertheless a contrastive tonal accent. Examples of tonal accent as well as word stress and phrasal stress may be found in the prosody part of the elicited material.

In order to influence the pronunciation of the target words and phrases in the controlled parts as little as possible, crossword-like word games were used to elicit the intended targets.

Most of the spontaneous material consists of informal interviews where the interviewer had been instructed to interfere as little as possible. In some cases, dialogues between two speakers of the dialect were used as an alternative.

6. Further development of the database

Maintaining the data and making it accessible for research is of course an important factor. This may seem as a fairly trivial task, but it is not. Sound format standards, for example, are changing over time. At the time of creating the database, we used an analysis package called ESPS/Waves. Neither the sound file format nor the format of the time aligned transcriptions are commonly used anymore and before not too long they will be completely outdated. It is therefore necessary to regularly update the file formats used in the database. There is simply no other way of long term preservation than regularly migrating the whole database to the currently favoured formats.

As mentioned above, the data consists of audio recordings and time aligned transcriptions. The original data in the ESPS/Waves format have now been converted to the currently most widely used formats – wav for the sound files and Praat TextGrid for the time aligned transcriptions.

Basic data about the speakers recorded for the database may be of great value for certain types of
linguistic studies. Minimally these data should contain information about speaker sex and age, educational level, vocational training and work experience. Some of the information presented in this paper about recording techniques, project descriptions (background and financing) and addresses to the people responsible for maintaining the database and monitoring access right should also accompany the database, ideally in the form of a meta database directly connected to the recorded data. There exists a now partly outdated version of such a meta database in the IMDI format developed by the Max Planck Institute. We are currently working on updating this database. Whether we will stay with the IMDI format has not been decided at the time of writing. We are also considering a move to a more modern and somewhat more flexible type of meta database, CMDI, also developed by Max Planck Institute.

7. Preserving the data for future generations

In the previous paragraph we have described one problem connected with maintaining digital databases – continuous format changes. Other factors influencing accessibility is constant technological change and mobility among the people involved. If we want to preserve the data for future research and guarantee its availability, the data must be secured in a way that does not depend on specific individuals, formats, server locations etc. Trying to solve this problem is one of our main concerns at the present stage. Fortunately we are not the only ones who are actively looking for solutions to these problems. There is considerable activity going on in this field.

For the purpose of long term preservation only, a copy of the database will be hosted by the Swedish National Data Service (NDS). But we are also working on a more advanced solution providing services specifically designed to service the speech research community. This service, Speech & Language Repository (SLDR) is hosted at the Aix-Marseille University in France.

8. Examples of research based on the material in the database

Intonation as a function of dialect has been studied for Swedish for a long time. The first study appeared already in the thirties (Meyer, 1937). This has been followed by many more studies over the years. Based on data in the SweDia database, a group of researchers at Lund University have developed models to simulate the prosodic variation among Swedish dialects. This work has been done within a project called SIMULEKT and the results have been described in a number of publications (e.g. Bruce et al., 2007, 2008).

Helgason has studied preaspiration in Nordic languages based, among other data, on material from the SweDia database (e.g. Helgason 2002, 2003).

Many more examples of studies using data from the SweDia database may be found in the publications list from the SweDia project (see. link at the end of this paper!).

9. Language variation from a somewhat different angle

Traditionally, the driving forces behind language variation and change are considered to be geographical dispersion and isolation of groups of speakers as well as renewed contact as a result of migration. These factors are no doubt important, but if that were all there is, the observed variation is likely to be more chaotic than what we seem to observe. A basic tenet in the SweDia project is the belief that although there is certainly a random element involved in language change it is primarily rule governed. One way of approaching this question is to look for coherence, or clustering of phonological properties within the entire speech community rather than assuming any specific areal distributions.

Promising results along these lines have been obtained by approaching the description of regional distribution from an angle that does not assume any geographically based constraints at all. In three studies (Leinonen, 2010; Lundberg, 2005; Shaeffler, 2005) based on the SweDia 2000 data, cluster analysis has been used as a means of creating dialect “areas” based only on acoustically grounded phonological properties. In those studies, geographical areas are defined by dialects whose properties cluster together. This approach could, in principle result in a very scattered picture with no obvious geographical coherence. This did not, however, turn out to be the case. On the contrary, dialects grouped into geographical areas that in many cases closely resemble those suggested in traditional dialectology. If the clustering had been based on the same considerations as in the traditional analyses this would have to be seen as a rather trivial finding, but this is not the case at all. In all the above studies, cluster analyses were based solely on acoustic properties like formant frequencies (Leinonen; Lundberg) or segment durations (Shaeffler) never considered in traditional dialectology. The results in a study by Livijn (2010) on the articulation of coronals, using a similar approach but without using cluster analysis, point in the same direction. Moreover, there is considerable overlap between the areas resulting from these studies. This lends support for the assumption that dialectal change is rather strongly constrained by the compatibility of internal factors.

10. Additional uses of the database

In addition to the research database, there is also a limited version of the database developed for educational purposes in university courses on dialectology, secondary schools and study groups of interested individuals. This database contains speech samples from all dialects represented by short sound files (30–50 seconds) from one speaker per category (age/sex) together with simplified phonetic-like transcriptions and translations to standard Swedish. This database may be accessed over the Internet. At present the interface exists only in Swedish. There are no immediate plans to translate the
interface.

A group of researchers at Lund University are using material from the database for studies of dialect syntax. They are part of a Nordic network of dialect syntax researchers (ScanDiaSyn). Studies of this kind were not envisaged when our data were collected, but we are pleased to see that the data can be fruitfully used also for such studies. To support their efforts we supply the ScanDiaSyn database hosted at the University of Oslo with data for their studies.

Although the data were collected for the primary purpose of studying language variation and change in the phonological domain, the usefulness is not necessarily limited to that area. As mentioned above, the data is now used also for the study of dialect syntax.

The database contains data from speakers of ages ranging from 20 years of age up to 75 years of age for both male and female speakers. That means that in addition to language variation data the database can be used to study speaker variation as a function of age. This has been done in a series of studies by Schötz. In her doctoral dissertation (2006) she studied the variation of parameters such as fundamental frequency, formant frequencies, jitter, shimmer and speech rate as a function of age. These results were then used as a basis for a model that could be implemented in speech synthesis to simulate speaker age. This has been further developed in later studies (e.g. 2007).

Another successful use of the data is as a reference database for automatic speaker recognition for forensic purposes. This has been described in Lindh and Eriksson (2009).

11. Summary

In this paper we have presented the SweDia project and the database created and developed within the project and in the last paragraphs we have given many examples of various uses of the data, not only uses which are primarily in the area of dialectology or even linguistics in a restricted sense. This may be seen as an example of what is often referred to as e-science, that is re-using existing data for new research, not envisioned when the data was collected but made possible because the data now exist.

12. Acknowledgements

The present work on the research database is supported by a grant from the Swedish Research Council (grant # 825-2007-7432).

13. References


CMDI home page. Available at: <www.clarin.eu/cmdi>.


IMDI home page. Available at: <www.mpi.nl/imdi>.


Reference list of Swedia 2000 publications. Available at: <www.ling.gu.se/~anders/SWEDIA/publ_eng.html>.


SweDat project presentation. Available at: <http://swedat.ling.gu.se>.

Swedish National Data Service. Available at: <http://snd.gu.se/en>.

The educational database. Available at: <http://swedia.ling.gu.se>.
14. Appendix

Figure 1: The geographical distribution of recording sites
Easyalign for Brazilian Portuguese: a (semi) automatic segmentation tool under Praat

Jean-Philippe GOLDMAN1, Maíra Avelar MIRANDA2, Cirineu Cecote STEIN3, Antoine AUCHLIN1

1University of Geneva, Switzerland; 2Federal University of Minas Gerais, Brazil; 3Federal University of Paraíba, Brazil jeanphilippegoldman@gmail.com, mairavelar@gmail.com, cirineu.stein@cchla.ufpb.br, Antoine.Auchlin@unige.ch

Abstract

This communication presents an automatic phone-text alignment system, EasyAlign, in its latest adaptation to Brazilian Portuguese. Automated steps are crucial in large corpora prosodic investigation as opposite to time-consuming human alignment, they are both more consistent, and reproducible. They also are open to adaptation and improvements. One issue is the tool’s precision in alignment at phone level.

Keywords: automatic segmentation; automatic alignment; Brazilian Portuguese; EasyAlign.

1. Introduction

Phonetic alignments (or phonetic segmentation) purpose is to determine the time position of phone, syllable, and/or word boundaries in a speech corpus of any duration, on the basis of the audio recording and its orthographic transcription. Resulting aligned corpora are widely used in various speech applications, such as automatic speech recognition, speech synthesis, as well as prosody and phonetic research.

Conducting fully manually an accurate segmentation would require as many as 800 times real-time; i.e. 13 hours for a one-minute recording (Schiell & Draxler, 2004). Processing time is a major drawback for manual labeling, especially facing with very large spontaneous speech corpora. This is why an automatic phonetic alignment tool is highly desirable. Such an automatic approach, besides, is not only consistent (i.e. has the same precision throughout the corpus), it also is reproducible (i.e. can be repeated, within a short time interval, and many times). An alignment tool can save time, but speech, especially spontaneous speech, presents unpredictable phonetic variations that can decrease process’ accuracy. Even with precise computational tools and data preparation, automatic systems can make errors that a human would not. Thus, manual or automatic post-processing detection of major segmentation errors is needed to improve accuracy. Automatic approaches are never fully automatic nor straightforward and instantaneous as claimed by existing systems. It is a matter of balance between time, expected precision and computational skills. The determination comes from the needed degree of accuracy; i.e. a corpus-based text-to-speech (TTS) system needs a high precision, but other studies (at syllable level) require a lesser precision.

For automatic phonetic alignment, several methods have been designed: some borrow techniques from the automatic speech recognition (ASR) domain. But, the alignment task is much easier than speech recognition as the alignment tool does not need to determine what the segments are but only their position in time. For that, HMM (Hidden Markov Models)-based ASR systems are used as a forced-alignment process for segmentation like HTK in (Young & Woodland, 2000). Other approaches combine a TTS system and a Dynamic Time-Wrapping (DTW) algorithm. In this case, the orthographic transcription is used to synthesize speech, which is compared to the authentic speech to segment as in (Malfrère, 2003). The DTW algorithm finds the best temporal mapping between the acoustic features of the two enunciations. A dual system based on these two approaches (first HMM then TTS+DTW) is presented in (Sérgio & Oliveira, 2004), with better results. Finally in Van Santen and Sproat (1999), contour detection techniques are borrowed from image processing, providing relevant results. Although these systems are usually freely available and give good results, it should be noted that they are not ready to use, as a training of the acoustic models is required.

The presented system, named EasyAlign, relies on HTK (for HMM ToolKit), a well-known HMM package. It can be seen as a friendly layer within Praat software (Boersma & Weenink, 2009), which does the whole alignment process as it is provided with a grapheme-phoneme conversion system and embeds already trained acoustic models.

2. EasyAlign

EasyAlign (Goldman, 2011) is a plugin developed for Praat. It produces semi-automatically a multi-tier annotation with a phonemic, syllabic, word and utterance segmentation from a sound recording and the corresponding orthographic and phonetic transcriptions. The plugin is made of Praat scripts, but it also includes two external components: a grapheme-to-phoneme conversion system and a segmentation tool for the alignment at the phone level. Consequently, the whole procedure is a succession of 3 automatic steps in between which some manual adjustments may be necessary. The 5 resulting tiers are grouped in TextGrids and named as phones, syll, words, phono and ortho as illustrated in Figure 1.
The tool is already designed for French, Spanish and Taiwan Min and has been recently adapted to Brazilian Portuguese (BP). It is freely available and works on Windows only. It is distributed as a self-installable plug-in, and comes with the already trained acoustic models of the phones.

The segmentation of a speech file occurs as follows: from a speech audio file and its corresponding orthographic transcription in a text file, the user has to go through three automatic steps; manual verifications and adjustments can be done in-between to ensure even better quality. More precisely, these three steps are as mentioned in (Goldman, 2011).

2.1 Macro-segmentation at utterance level
The first automatic step of EasyAlign creates a macro-segmentation as a TextGrid with one tier named ortho, on the basis of the previously loaded Sound object (the sound file to segment) and Strings object (the transcription). After this step, the Sound and the new TextGrid are opened. The internal algorithm computes a heuristics based on signal duration and utterance transcription length to estimate utterance duration and also relies on pauses.

2.2 Grapheme-to-phoneme conversion
The second step duplicates the ortho tier into a phono tier (i.e. with the same boundaries) but replaces the orthographic transcription by a phonetic transcription according to the SAMPA phonetic alphabet. The grapheme-to-phoneme module of a full text-to-speech system named eLite and developed at Faculté Polytechnique de Mons (Belgium), does a linguistic analysis of the orthographic transcription to produce a phonetic transcription based on a phonetic dictionary and pronunciation rules.

2.3 Phone segmentation
The third step aims at creating the phones, syll and words tiers. For each utterance, the orthographic and phonetic transcriptions are used by a well-known speech recognition engine named HTK (HMM ToolKit) set to a “forced alignment” mode to catch the temporal boundaries of phones and words. The syllabification is rule-based. Two main principles are: 1. there is one and only one vowel per syllable; and 2. the sonority principle is used to split the consonant clusters. The pauses are also used as syllabic boundaries.

2.4 Result
The result is a multi-tier TextGrid, the annotation format within Praat, with phones, syllables, words and utterance segmentation, as showed in Figure 1. It is important to highlight that, before performing each of these steps, it is necessary to make some manual adjustments, as showed in Figure 2.

As it can be observed, at first, we have a preliminary manual step (if the transcription is in a paragraph format and/or without punctuation), in which the user has to reformat the transcription file with one utterance per line. After that, an utterance segmentation script is run, which creates a TextGrid with an interval tier ortho containing the transcription. The user, then, manually verifies the utterance boundaries. Next, it is done the automatic grapheme-to-phoneme conversion: the script duplicates the ortho tier to the phono tier, generating a phonetic transcription with major variations. At this point, the user may validate the phonetic transcription to ensure sporadic phonological variant of pronunciation. This optional time-consuming task might be skipped. Finally, the phone segmentation is automatically performed: the script is run and generates the phones and words tiers, then the syllables tier.
3. EasyAlign for Brazilian Portuguese

3.1 Development

EasyAlign adaptation to a new language needs: some speech data and a grapheme-to-phoneme conversion system. First of all, we selected two audio samples, that have a total duration of 20 minutes (10 minutes produced by a male reader and 10 minutes by a female one), from a corpus (Barbosa et al., 2004) composed by 4 subjects, 2 males and 2 females, reading sentences. The corpus was manually aligned. Then, we integrated the “Conversor Grafema-fone v1.6” (Grapheme to Phone conversor 1.6) phonetizer within EasyAlign to convert the orthographic transcription to the phonetic one.

This phonetizer was developed by Fala Brasil team (http://www.laps.ufpa.br/falabrasil/) at the Federal University of Pará (UFPA) Brazil (Siravenha et al., 2008). Despite its good performance, some problems must be solved, such as the specification of a dictionary of exceptions for open vowels (the ones which are not predicted by phonological rules, or are no longer predicted through orthography, due to the latest official orthographic changes in Portuguese). That improvement is on course.

As shown in the *phon* tier of Figure 1, the grapheme-phoneme conversion tool provides a phonetic transcription, in SAMPA alphabet, on the basis of the orthographic transcription. The phonetic transcription of each utterance was manually checked so as to exactly match the produced utterance.

In the end, a HTK-based stochastic training was performed with this speech material and their phonetic transcription. The result is a collection of acoustic models. The Figure 2 shows the necessary steps to train then to use EasyAlign.

3.2 Evaluation

According to (Goldman & Schwab, 2011), “the evaluation of a semi-automatic system can be seen in two ways: i) its automatic performance, i.e. how robust and accurate the automatic tool is, and ii) its ergonomics, i.e. how the whole process is made easier and how many times real-time it takes”.

The automatic performance has been evaluated on the basis of a corpus of twenty-four minutes, that was manually annotated by one phonetic expert (reference alignment) and was compared to the automatic alignment. Among the speakers, two were “internal” speakers, used in the training corpus and two were new “external” speakers, taken from political debates broadcasted by Record TV channel. The internal corpus was composed by twenty minutes (10 minutes produced by a male speaker and 10 composed by a female one) and the
EASYALIGN FOR BRAZILIAN PORTUGUESE: A (Semi) Automatic Segmentation Tool Under Praat

An external corpus was composed by four minutes (2 minutes produced by a male speaker and 2 by a female one).

Evaluation was performed according to three approaches: a boundary-based, a duration-based and a segment-based approach. In each of the evaluations, the pauses were discarded, and only phones were taken into account. As some segments might be very short, especially in spontaneous speech, the evaluation was done with two thresholds: the 20ms (as mentioned above) and a narrower one set at 10ms.

3.3 Boundary-based evaluation
In this first evaluation, we computed the absolute difference (in ms), for each phone (n = 11650), between the automatic and the manual initial boundaries. Results showed that 43.2% of the differences between automatic and manual boundaries lie within 10ms, and 73.1% within 20ms for the internal corpus.

<table>
<thead>
<tr>
<th>Differences</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Within 10 ms</td>
<td>43, 2%</td>
</tr>
<tr>
<td>Within 20 ms</td>
<td>73,1%</td>
</tr>
</tbody>
</table>

Table 1: Boundary-based evaluation for the internal corpus

3.4 Duration-based evaluation
For each phone, we looked at the difference between the automatic and manual segment durations.

<table>
<thead>
<tr>
<th>Differences</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Within 10 ms</td>
<td>33, 2%</td>
</tr>
<tr>
<td>Within 20 ms</td>
<td>42,5%</td>
</tr>
</tbody>
</table>

Table 2: Boundary-based evaluation for the external corpus

3.5 Segment-based evaluation
According to Goldman and Schwab (2011), in the segment-based evaluation, we computed, for each phone, the Overlapping-rate, a speech-rate independent measure (Sérgio & Oliveira, 2004), which represents the ratio between the common part of the automatic and manual segment and the maximal duration of the segment considering initial and final boundaries of both automatic and manual segmentations. A rate of 0 means that there is no overlap between the automatic and manual segments, while a rate of 1 means that the overlap is total. According to Sérgio and Oliveira (2004), a segment with an overlapping rate of 0.75 is considered well segmented.

<table>
<thead>
<tr>
<th>Differences</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>OVR Mean</td>
<td>0.671</td>
</tr>
<tr>
<td>Sdev</td>
<td>0.239</td>
</tr>
</tbody>
</table>

Table 5: OVR evaluation for the internal corpus

The mean value is much higher for the internal corpus than for the external corpus, which indicates a better overlapping rate for the internal corpus, and thus the need of a better training.

4. Conclusion
The 3 kinds of evaluation were done – boundary-based, duration-based and segment-based. All of them showed promising results, with a good training of the training corpus. On the other hand, the external evaluation corpus was under-represented and, consequently, generated poor results. We need, then, to increase the size of the training corpus to obtain a more accurate training and make good generalizations from the external evaluation corpus.

To sum up, EasyAlign appears as an friendly and
efficient tool which helps aligning speech from an orthographic transcription within Praat. The tool is freely available online and is complemented by a demo mode and a tutorial. It can be downloaded from this link:

http://latlnic.unige.ch/phonetique/easyalign

To our knowledge, such a tool was not, until now, available for Brazilian Portuguese.

5. Acknowledgements
This research is partly funded by The Swiss National Science Foundation - FNS Grant nr 100012_134818.

6. References


DB-IPIC: an XML database for informational patterning analysis

Lorenzo GREGORI, Alessandro PANUNZI
Università di Firenze
E-mail: lorenzo.gregori@unifi.it, alessandro.panunzi@unifi.it

Abstract

DB-IPIC is a linguistic web resource for the analysis of spoken language based on the Informational Patterning Theory of E. Cresti and M. Moneglia. The corpora stored inside the database take parts of the C-ORAL-ROM and C-ORAL-BRASIL projects and enrich them with informational and PoS tagging. This paper focuses on DB-IPIC’s construction, from the annotation processes of acoustic sessions to the retrieval capabilities of the web interface. In the first part we give a short overview of the theoretical framework on which the database has been structured and we describe the annotation procedure of speech sessions. In the second part we explain the XML data model and the conversion process from annotated data to XML. Finally, we describe the steps that have been followed to build DB-IPIC itself, along with its querying capabilities; in particular we’ll describe the web interface and its features for extracting information patterns and analysing results.

Keywords: DB-IPIC, XML database; information patterning; C-ORAL-ROM.

1. Introduction

DB-IPIC is a database of transcribed and annotated spoken language: in this paper we are going to describe this resource, focusing on the data types comprising the database and on the tools provided by the web interface to query it. At the moment, the database stores a corpus of 74 spoken Italian language texts chosen from the Informal section of Italian C-ORAL-ROM (Cresti, Panunzi & Scaranò, 2005). The whole corpus has been tagged with respect to the informational structure and exploited to build a queryable XML database for the study of linear relations among Informational Units in spoken language (Panunzi & Gregori, 2012). In addition to this we inserted a subset of the C-ORAL-BRASIL (20 texts; Raso & Mello, 2012) corpus and provided an Italian collection with the same size for comparison with the Brazilian one (Mittmann & Raso, 2012).

Besides the database, DB-IPIC includes a web interface that provides an easy means of extracting complex data from the corpora. With this tool it’s possible to query the database, crossing different kinds of information stored in different logical levels (the logical model is explained in paragraph 2). The DB-IPIC web interface is specifically designed for the search and analysis of information patterns and the comparison of the informational values and prosodic profiles of linguistic structures (Mittmann et al., in this volume). Beyond this, DB-IPIC provides more search features, such as part-of-speech filtering and communicative context restriction.

1.1 Theoretical framework

DB-IPIC is built in accordance with Language into Act Theory and Informational Patterning Theory (Cresti, 2000; Cresti & Moneglia, 2010). These two theoretical models form a framework that can be productively applied to the annotation of spoken language.

The framework identifies two different pragmatic levels in oral production: the first one is macro-pragmatic and deals with the informational structure. Both of these levels are governed by prosody which splits the speech flow into terminated sequences and tone units using terminal and non-terminal breaks, respectively. These breaks are pragmatically defined as perceptually relevant prosodic variations in the speech flow (Cresti & Moneglia, 2005: 17) and acoustic source analysis has revealed a link between prosodic breaks and F0 behaviour.

At the macro-pragmatic level, the oral performance is structured into Utterances, which correspond to the pragmatic reference unit for spoken language. An Utterance is a sequence of words that can be pragmatically interpreted and corresponds to a Speech Act. On the prosodic side, an Utterance is included in a Terminated Sequence (TS), which ends with a perceptually identifiable terminal break. So, at this level, we have the TS that is prosodically recognizable by the terminal break and even pragmatically interpretable, since it achieves an Utterance.

At the micro-pragmatic level, Utterances can also be divided into sub-elements that are coherent with respect to the information value they carry. These elements are called Information Units (IU): on the prosodic side, IUs are segmented by non-terminal breaks, which split the TS into a sequence of Tone Units (TU). The Comment is the core IU of an Utterance and corresponds to the expression of an illocutionary force, being necessary as it ensures pragmatic interpretability. The Comment can be surrounded by other IUs, each with a specific informational value. The IUs can be divided into two main classes: the textual units that participate in the construction of the semantic content of the Utterance (Topic, Appendix, Parenthesis, Introducer), and the dialogical units that are devoted to the successful pragmatic performance of the Utterance in the communicative context (Incipit, Phatic, Allocutive, Conative, Connector, etc.). The full tagset with descriptions is available in Table 1 in appendix.

Within the proposed theoretical framework, each Utterance consists of a pattern of IUs that is roughly isomorphic to a pattern of TUs (informational patterning principle). Therefore, there is a strict connection in the...
definition of the two pragmatic levels of spoken language:
(a) An Utterance is defined as the linguistic expression of a Speech Act, but it can be also viewed as a pattern of IUs;
(b) the informational pattern necessarily contains a Comment IU, which properly accomplishes the illocution and corresponds to a single TU.

There are “exceptions” in which the theoretical principles explained previously cannot be applied. Two cases in particular should be mentioned:

- Illocutionary patterns: these structures occur within an Utterance when the Comment carrying the illocution doesn’t “fit” inside a unique TU; in these cases a Multiple Comment is produced which consists of a pattern of two or more TUs (linked together through a compositional informational/prosodic model) with an overall illocutionary value;
- Stanzas: these structures are oral performances in which there is more than one Comment unit in sequence with weak illocutionary force. These sequences do not form compositional units, but they are produced by progressive adjunctions, outside of any informational/prosodic model. Stanzas correspond to a linguistic “activity” whose primary intention is the production of an oral text.

Thus, three types of Comment unit are defined:

- Comment (COM): the standard Comment IU, which accomplishes the illocutionary force of the Utterance and corresponds to a single TU;
- Multiple Comment (CMM): a complex IU composed of two or more TUs and forming an illocutionary pattern;
- Bound Comment (COB): occurring within Stanzas and corresponding to a non-patterned sequence of Comments with weak illocutionary force.

In short, there are two referring units to which a TS can correspond: the Utterance, which mostly aims at an interactive exchange with the interlocutor (Speech Act performance), and the Stanza, which intends the construction of a “text” by the speaker. Within the database, these units are distinguished with different tags (different attributes of the TS element). Furthermore, single and multiple Comments are recognizable inside the data model because of the different labels applied to the IUs in these two structures (COM vs. CMM).

1.2 Annotation procedure

Following the sketched theoretical framework, the prosody drives the annotation procedure. Since the prosodic segmentation of speech flow is strictly connected to the pragmatic features, the first annotation step consists of marking terminal and non-terminal breaks. This is done manually and occurs in parallel with the transcription procedure. For this step the annotators used WinPitch (Martin, 2005), a tool which allows one to listen an audio recording and carry out the text-sound alignment of a transcription, as well as to analyse the acoustic features of the source (in particular, an F0 real-time examination is required to safely determine the breaks).

The second annotation step, also performed manually, consists of tagging the IUs and exploits the informational patterning principle: once the prosodic boundaries of a TS and the internal pattern of the tonal units have been detected, it is possible to mark each TU with its informational value to get the informational pattern (Scaranò, 2009; Moneglia, 2011). This is divided into two stages: first the Comment unit is identified and the TS type determined (Utterance or Stanza), and then the other TUs are tagged.

Finally, general session metadata, comprehending details about the audio source, the participants features, and communicative context are added to the annotation. Data and metadata are written in a CHAT-like format (MacWhinney, 2000; Moneglia & Cresti, 1997).

The result of this work is a set of sessions with audio, transcription, text-sound alignment, prosodic annotation, and information structure annotation. A corpus created following this multi-level annotation procedure is a resource rich in data that can be used as reference for studies on spoken language. But, as we will see in the chapter below, the original structure of the corpus does not make it effectively accessible to the scientific community: for these reasons, the production of an integrated database has been necessary.

2. Annotation tree and XML model

As a result of the manual annotation procedure, we have three files for each spoken session: an RTF file containing metadata, a transcription, and annotation tags, the WAV audio file of the recording, and the WinPitch XML file containing the text-sound alignment information. For the purpose of building a queryable resource, this representation format has several problems: data are sparse, RTF is not a real standard, annotations and transcriptions are written in a non machine-readable format, and all information is inserted inline into the text file without considering the dependence among different annotation levels. For these reasons a new representation model was developed.

As mentioned previously, we have two main structures involved in the segmentation of speech flow, Terminated Sequences and Tone Units, and one is superordinate to the other (the informational features can simply be added as labels belonging to the elements).

We then have high level metadata specifying session features and low level data that includes transcription and prosodic annotation. A peculiarity of this multi-level annotation is that it is structured as a tree, in which logical levels are linked in a hierarchical data model. This is one of the reasons that led us to use XML as the standard format for the corpus representation (Gregori, 2011).

XML has many good features causing it to be widely
used, especially for encoding and sharing information throughout the web, but, generally, corpora with multi-level annotation cannot be easily stored in an XML model. Commonly, each annotation level is independent from the other ones and a tree is too rigid for representing the data structure: we can say that it’s typically difficult to encode a multi-level annotated corpus into XML without losing human-readability feature, as more than one file is needed per session. Otherwise, our collection fits well into an XML tree and the features of this language make it a good choice for storing DB-IPIC: firstly, the XML format allows an efficient standardization of the annotated data and formal validation. Moreover, XML is able to encode information that requires different kinds of representation (category, structural and relational information) and its elements are organized into a hierarchical model. Finally, the XML “family” comprehends query languages directly applicable to the annotated texts. The necessity to find a representation format for the IPIC collection led to the development of an XML data model and of software for automatic data migration in the new format.

An additional feature we decided to insert into the corpus is PoS-Tagging, so another annotation level has been inserted into the XML model. This additional information is derived automatically using TreeTagger. TreeTagger execution runs automatically inside our internal software that converts the corpus into XML format.

Each session of the corpus is composed of the following data types:

- an audio stream, containing the audio recording;
- general metadata, containing details about the session (audio quality, communicative context, etc.);
- transcription, consisting of a text that reports the word sequence;
- prosodic annotation, containing speech flow segmentation;
- information structure annotation, specifying the informational rule of any TU;
- morpho-syntactic annotation, automatically induced by TreeTagger;
- text-sound alignment, generated at the transcription procedure time by the WinPitch software.

All these data has been structured into an XML model according to the theoretical framework and considering the following relational rules among the levels:

- transcription data are at the lower level of the annotation tree; each transcription element is qualified depending on its nature (word, break, fragment, paralinguistic);
- part-of-speech and lemma are properties of words;
- TUs are superordinate to transcription elements and IUs are isomorphic to them, so informational values are properties of the TU elements;
- TSs are superordinate to TUs and they have a number and a type that depends on the relative reference unit: Utterance or Stanza. Alignment data are also a property of TSs, since it specifies their start and end times;
- general metadata is independent from the annotation tree: it depends only on the session;
- the session is the root level and includes all the other data.

This model has been translated into XML: objects and properties have been transformed into elements and attributes, preserving their logical difference. Figure 1 in appendix shows the structure.

### 3. DB-IPIC resource

As the IPIC collection is stored in XML files, we decided to use an XML database to index it and make it queryable. Even if this kind of storage technology is not as efficient as common relational databases, the choice is justified by the fact that we have a unique data format for representation and querying. In addition to this, the corpus size is adequate enough to yield a good response time for any query.

We chose eXist-db, which is an open source software that runs as a server and can be queried via web protocols using the standard query language “XQuery”. A user-friendly web interface has been developed in PHP to allow the extraction of informational patterns from the database (Figure 2). With this tool it’s possible to query the corpus at different levels in relation to the logical structure of the data set. In particular DB-IPIC can operate on five levels:

1. data source: it is possible to query the whole corpus or to specify a subset of sessions; different corpora can be managed in DB-IPIC;
2. metadata: sessions can be filtered by their properties, specifying the communicative context (familiar or public) and the interaction type (monologue, dialogue, or conversation);
3. informational patterns: the user can select the TSs by specifying their IU pattern;
4. information units: it’s possible to search TSs containing or not containing specific IUs independently of their informational pattern;
5. words: finally, the user can refine their search by including or excluding words with a specific form, PoS, or lemma.

As mentioned, the main purpose of the DB-IPIC resource is the search of information patterns and for this reason it provides advanced features for searching the objects inside the corpus. The following actions are allowed:
definition of multiple sequences of IUs at once by using regular expressions: each element of the IU pattern can be extended to a variety of IUs using the W3C regular expressions syntax (Peterson et al., 2012);
• selection of the linear relation among the IUs of the pattern by specifying what IUs can optionally interrupt the sequence. There are five possible choices, from the more rigid, in which the IUs must be adjacent, to the freest, in which there are no restrictions about the IUs that can interrupt the sequence;
• specifying the content of each IU of the pattern in terms of word form, part-of-speech, and lemma. For this feature we developed a lightweight CQL\(^1\) parser and a graphical tool that helps the user to write the restriction in the correct syntax.

In addition to the information pattern definition, DB-IPIC allows one to make complex queries through the intersection of the five logical levels described above.

![DB-IPIC web interface](image)

Figure 2: DB-IPIC web interface

We can take Figure 2 as an example of the search capabilities of the resource: we decided to retrieve the dialogues in public context and excluding Stanzas (General filter section) from the Italian corpus (Source selection section); each Utterance must contain an Appendix of Comment and the lemma “essere” and cannot contain a Multiple Comment (Utterance restrictions section); Utterances must include the Topic-Comment pattern, in which the Topic contains a

noun and is the first IU of the Utterance (Search for Information Pattern section).

You can see in this query that all five logical levels described previously are involved. The results are shown in Figure 3 in appendix. Query results are displayed in the CHAT format: the interface shows the list of Utterances matching the query parameters. Audio is directly accessible, through the exploitation of the alignment data and the three buttons located on the right of each entry correspond to the functions available for each TS: online audio playing, audio file download (in WAV format), and opening of the acoustic stream with WinPitch for deep analysis. Finally, it’s possible to download all the search results in a format compatible with spreadsheet applications (CSV file), by clicking the icon in the upper right of the page.

The DB-IPIC web resource is available at the project’s homepage \(^2\) and freely usable. Though it’s possible to query the corpora using the XQuery language by following the public XML Schema, this approach is not recommended due to the complexity of the XML model. DB-IPIC is already designed to support data retrieval at different levels, from general metadata to words in the transcription.

4. Conclusions

In closing, we want to remark that the annotation is based on prosodic features that are perceptually relevant. The inter-annotator agreement of such an annotation has been proved by a statistic analysis done for C-ORAL-ROM, which points out an agreement of more than 95% in the distinction of breaks (Moneglia et al., 2005). The high reliability of these data is an important quality of the corpus and, in general, of the whole annotation procedure that is founded on a universally agreed feature of speech. Moreover, this validates the choice to consider TSs and TUs as the structural elements of our data model.

On the other hand, we don’t have statistics about the accuracy of the informational tagging, because the full revision of the corpus has not yet been done. A corpus validation session is necessary for informational data, requiring an inter-rater agreement approach, and can lead to data alterations in the database. On this point we want to underline that data inside DB-IPIC are easy to modify: this is an important feature that we obtained from the creation of a structured data model and the usage of an XML database.

We also note that information about parts-of-speech and lemmas is induced automatically with software that uses a probabilistic model: with this approach errors are frequent, especially in a spoken language context. A manual revision of PoS-tagging would be desirable and would allow us to produce a gold standard for the informational annotation of spoken language resources.

---

\(^1\) CQL (Corpus Query Language) is a language developed from the University of Stuttgart to make lexical queries on corpora.

\(^2\) http://lablita.dit.unifi.it/ipic
5. References


Treetagger. Available at: <http://www.ims.uni-stuttgart.de/projekte/corplex/Tre eTagger/>.


eXist-db. Available at: <http://exist.sourceforge.net/>.
6. Appendix

Figure 1: The XML data model

Figure 3: DB-IPIC results page
<table>
<thead>
<tr>
<th>Class</th>
<th>Name</th>
<th>Tag</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comment Units</td>
<td>Comment</td>
<td>COM</td>
<td>Comment IU accomplishes the illocutionary force of the Utterance, and it is therefore necessary and sufficient to perform an utterance</td>
</tr>
<tr>
<td></td>
<td>Multiple-Comment</td>
<td>CMM</td>
<td>A complex IU comprised of two or more Comments, forming an <em>illocutionary pattern</em></td>
</tr>
<tr>
<td></td>
<td>Bound Comment</td>
<td>COB</td>
<td>A sequence of Bound Comments with weak illocutionary force, produced by progressive adjunctions following the flow of thought, out of any model of informational patterning</td>
</tr>
<tr>
<td>Textual Units</td>
<td>Topic</td>
<td>TOP</td>
<td>It identifies the domain of application for the illocutionary act expressed by the comment, providing the Speech Act with a cognitive reference and allowing the Utterance displacement from the actual context</td>
</tr>
<tr>
<td></td>
<td>Topic List</td>
<td>TPL</td>
<td>A chain of Topics forming a pattern of Topics</td>
</tr>
<tr>
<td></td>
<td>Appendix of Comment</td>
<td>APC</td>
<td>It integrates the text of the Comment and concludes the Utterance</td>
</tr>
<tr>
<td></td>
<td>Appendix of Topic</td>
<td>APT</td>
<td>It gives a delayed integration of the information given in the Topic adding specification for the addressee</td>
</tr>
<tr>
<td></td>
<td>Parenthesis</td>
<td>PAR</td>
<td>It adds information to the utterance with a meta-linguistic value having &quot;backward&quot; or &quot;forward&quot; scope; always bears a modal value.</td>
</tr>
<tr>
<td></td>
<td>Locutive Introducer</td>
<td>INT</td>
<td>It is used for introducing a sequence of IUs that have a strong and unitary &quot;point of view&quot;, as in reported speech and reported thought</td>
</tr>
<tr>
<td>Dialogic Units</td>
<td>Incipit</td>
<td>INP</td>
<td>It opens the communicative channel bearing a contrastive value starting a dialogic turn or an utterance</td>
</tr>
<tr>
<td></td>
<td>Conative</td>
<td>CNT</td>
<td>It pushes the listener to take part in the dialogue in an adequate way, or stops his non collaborative behavior</td>
</tr>
<tr>
<td></td>
<td>Phatic</td>
<td>PHA</td>
<td>It is dedicated to controlling the communicative channel, ensuring its maintenance; it stimulates the listener to the social cohesion needed by the dialogical exchange and/or ensures the reception of the utterance</td>
</tr>
<tr>
<td></td>
<td>Allocutive</td>
<td>ALL</td>
<td>It specifies to whom the message is directed keeping his attention. Simultaneously it plays a cohesive and empathetic function, bringing the interlocutor to share the point of view of the utterance</td>
</tr>
<tr>
<td></td>
<td>Expressive</td>
<td>EXP</td>
<td>It works as an emotional support. It stresses the sharing of a common social affiliation with the interlocutor, searching for social cohesion.</td>
</tr>
<tr>
<td></td>
<td>Discourse Connector</td>
<td>DCT</td>
<td>It zips different parts of the discourse (e.g. utterances within a turn), signaling to the addressee that the discourse is going on and that the entity which follows holds a relation with the previous ones.</td>
</tr>
<tr>
<td>Non-informative Units</td>
<td>Scanning</td>
<td>SCA</td>
<td>It occurs when the corresponding prosodic unit has no informational function and its locutive content is part of a larger IU (by default occurring on its right)</td>
</tr>
<tr>
<td></td>
<td>Interrupted</td>
<td>EMP</td>
<td>Interrupted units which cannot be evaluated</td>
</tr>
<tr>
<td></td>
<td>Time Taking</td>
<td>TMT</td>
<td>Time taking units for programming needs</td>
</tr>
<tr>
<td></td>
<td>Unclassified</td>
<td>UNC</td>
<td>Unclassified Units</td>
</tr>
</tbody>
</table>

Table 1: Information units
NURC digital: uma proposta de preservação dos dados do projeto NURC

Miguel OLIVEIRA Jr.
Universidade Federal de Alagoas
Maceió, Alagoas, Brasil
miguel@fale.ufal.br

Resumo
O presente artigo descreve um projeto de pesquisa tem por objetivo central propor um modelo de informatização de um dos corpora mais influentes na pesquisa linguística do Brasil: o corpus do Projeto NURC. Partindo de recomendações de órgãos internacionais especializados em práticas de codificação e transmissão de dados digitais, um corpus de dados representativos do Projeto NURC será organizado e apresentado aos coordenadores de todas as capitais brasileiras que sediam o Projeto NURC como possível modelo a ser adotado para a informatização, preservação e disponibilização de seu acervo, que atualmente se encontra em sério risco de deterioração devido à ação do tempo.

Palavras-chave: NURC; preservação; dados orais.

1. Introdução

Os dados que fazem parte do acervo do Projeto NURC têm sido utilizados para a elaboração de um grande número de trabalhos acadêmicos, incluindo dissertações de mestrado, teses de doutorado, artigos publicados em periódicos nacionais e internacionais, e trabalhos apresentados em encontros científicos ao redor do mundo. A Gramática do Português Falado (Castilho, 1990; Castilho, 1993; Castilho & Basílio, 1996; Ilari, 1992; Kato, 1996; Koch, 1996; Neves, 1999; Abaurre & Rodrigues, 2002), grande e ambicioso projeto nacional que envolveu entre 1988 e 2002 cerca de cinquenta pesquisadores na área da linguística, resultou em uma série de volumes, todos contendo análises de materiais extraídos dos dados do Projeto NURC. AGramática do Português Falado é um livro que contém muitos dos textos reunidos pelo Projeto NURC, e os pesquisadores que usaram os dados do Projeto NURC, assim como os pesquisadores que usaram os textos da Gramática, têm a oportunidade de consultar os textos originais e utilizá-los para o estudo de diversas características da oralidade, que vão desde aspectos discursivos, tais como a análise de narrativas inseridas na conversação (Oliveira Jr., 1999) e de questões discursivas e ideológicas presentes nas diversas modalidades de gravações feitas pelo Projeto (Cunha, 2003), até aspectos mais formais, tais como a análise de elementos argumentativos e pragmáticos, da intertexualidade e da organização interacional e sintática presentes no texto oral (Sá, 2004).

2. Justificativa
Entende-se por corpus, nos estudos linguísticos, uma “coletânea de porções de linguagem que são selecionadas e organizadas de acordo com critérios linguísticos explícitos, a fim de serem usadas como uma amostra da linguagem” (Percy et al., 1996: 4). O corpus do Projeto NURC é uma coletânea de dados de fala de informantes com formação universitária completa (chamados cultos), organizada para servir de estudo da modalidade oral da língua portuguesa culta falada no Brasil. O material do Projeto NURC foi – e tem sido – largamente utilizado para o estudo de diversas características da oralidade, que vão desde aspectos discursivos, tais como a análise de narrativas inseridas na conversação (Oliveira Jr., 1999) e de questões discursivas e ideológicas presentes nas diversas modalidades de gravações feitas pelo Projeto (Cunha, 2003), até aspectos mais formais, tais como a análise de elementos argumentativos e pragmáticos, da intertexualidade e da organização interacional e sintática presentes no texto oral (Sá, 2004).

A maior parte dos estudos desenvolvidos a partir dos dados do Projeto NURC deriva de uma série de publicações feitas com transcrições de material selecionado pelos grupos de pesquisadores atuantes em cada uma das capitais em que o Projeto era desenvolvido. Essas coletâneas de transcrições publicadas a partir da década de 80 ficaram conhecidas por Materiais Para o Seu
O principal objetivo do presente projeto de pesquisa é propor uma metodologia de organização de um corpus representativo do acervo do Projeto NURC, em formato digital, que servirá como possível modelo a ser adotado para a informatização de todo o material pertencente ao arquivo do Projeto NURC. Para isso, serão levados em conta procedimentos internacionais estabelecidos para a construção de corpus linguístico digital. Este projeto representa, assim, um importante passo no processo de preservação do valioso material do Projeto NURC, de forma que o mesmo possa ser utilizado mais eficazmente não apenas no presente, mas por futuras gerações de pesquisadores.

### 3. Objetivos

O principal objetivo do presente projeto de pesquisa é propor uma metodologia de organização de um corpus representativo do acervo do Projeto NURC, em formato digital, que servirá como possível modelo a ser adotado para a informatização de todo o material pertencente ao arquivo do Projeto NURC. Para isso, serão levados em conta procedimentos internacionais estabelecidos para a construção de corpus linguístico digital. Este projeto representa, assim, um importante passo no processo de preservação do valioso acervo do Projeto NURC, que atualmente se encontra em sério risco de deterioração ocasionada pela ação do tempo. Além disso, os resultados provenientes da execução do projeto aqui proposto beneficiarão diretamente a comunidade científica, que passará a ter disponíveis para consulta otimizada dados – anteriormente de difícil acesso – em formato digital de alta qualidade, devidamente catalogados, etiquetados e transcritos.

Como objetivos específicos, o projeto aqui proposto pretende:

1. contribuir para a formação de pesquisadores nas áreas da documentação linguística, da linguística
de corpus e da análise da oralidade;
ii. digitalizar todo o acervo do Projeto NURC/Recife, originalmente gravado em formato analógico, respeitando os padrões recomendados pelos órgãos internacionais de codificação e transmissão de dados digitais;
iii. catalogar e armazenar em formato digital todas as informações referentes ao material de áudio digitalizado;
iv. informatizar os dados de transcrição referentes a parte do material de áudio digitalizado (o corpus compartilhado do Projeto NURC/Recife, tornando-os alinhados, o que propiciará uma utilização mais proveitosa dos mesmos;
v. propor um sistema de anotação/etiquetagem multi-nível para os dados do Projeto NURC;
vi. anotar / etiquetar um corpus representativo dos dados do Projeto NURC, com informações multi-níveis;
vii. arquivar os dados informatizados em bancos de dados internacionais, assegurando assim a sua preservação;
viii. elaborar um documento com proposta de digitalização, preservação e anotação dos dados do Projeto NURC, elaborada a partir de discussão com todos os coordenadores do Projeto NURC, levando-se em conta as recomendações de órgão internacionais especializados em arquivamento de dados digitais;
ix. republicar os Materiais para o Seu Estudo em formato digital, contendo todos os dados do corpus compartilhado (transcrição, anotação e áudio);
x. editar um volume Estudos, composto de artigos feitos a partir do corpus compartilhado do NURC/Recife;
xi. disponibilizar o corpus compartilhado digitalizado e anotado para a elaboração de trabalhos os mais variados (artigos, capítulos de livro, dissertações e teses), dentro do âmbito do projeto.

4. Metodologia

O presente projeto de pesquisa tem por objetivo informatizar um corpus representativo do material do Projeto NURC, com o propósito de sugerir uma metodologia padrão, baseada em recomendações feitas por órgãos internacionais de codificação e transmissão de dados digitais, para ser adotada no Projeto NURC como um todo, preservando, assim, o seu precioso acervo, e permitindo que ele seja utilizado de maneira mais eficiente no futuro. Todo o acervo do Projeto NURC/Recife será digitalizado. Parte deste acervo será também anotado. O material a ser anotado corresponderá ao corpus compartilhado do Projeto NURC Recife. Justifica-se a escolha desse material pelo fato de ser o proponente deste projeto pesquisador do Projeto NURC Recife desde 1990, tendo, portanto, acesso ao acervo daquela capital. Além disso, cumpre notar que a sala do Projeto NURC Recife foi recentemente inundada, devido às fortes chuvas de abril/maio de 2011 naquela região. Ainda não se tem ideia da proporção dos estragos causados por esse incidente no que diz respeito ao material ali arquivado. Entretanto, o incidente por si só já justifica a necessidade – e mesmo a urgência – de se estudar uma estratégia de arquivamento mais eficiente para o acervo do Projeto NURC em geral, e do acervo do Projeto NURC/Recife em particular.

Os dados de áudio do corpus compartilhado do Projeto NURC/Recife – material selecionado para compor o corpus representativo deste projeto – serão digitalizados observando-se as recomendações propostas pelo Open Archival Information System (OAIS), que é um modelo de referência, com padrão ISO (14721:2003), adotado pelos bancos digitais de dados linguísticos mais recentes, e pelo Comitê Técnico da IASA para objetos digitais (Bradley, 2009; Von Arb & Lars, 2005). As informações referentes aos arquivos de áudio e às transcrições (metadados) serão registradas seguindo o padrão Dublin Core e o protocolo da Open Archives Initiative Protocol for Metadata Harvesting, também adotados por bancos de dados internacionais. As transcrições dos dados serão registradas no aplicativo ELAN, que possibilita o seu alinhamento com os arquivos de áudio a que se referem, além de permitir que áudio, transcrições e metadados sejam pesquisáveis local e virtualmente. Durante toda a fase de digitalização e tratamento do material do Projeto NURC, backups regulares serão realizados em lugares diferentes do local onde os dados primários estarão custodiados, garantindo assim a preservação dos mesmos.

Os inquéritos do Projeto NURC foram gravados em condições variadas. Em geral, as gravações eram realizadas com microfones dinâmicos omnidirecionais, apoiados em uma mesa. Todos os inquéritos foram registrados em fita de rolo. A dependência do tipo de inquérito, as gravações eram realizadas em salas específicas, em salas de aula, em auditórios e, em algumas casas, nas casas dos informantes. Portanto, a qualidade acústica das gravações do Projeto NURC é bastante heterogênea, não sendo possível descrever um perfil das gravações como um todo em termos de relação sinal-ruído. Diante deste cenário, não é viável que se aponte como objetivo do presente projeto disponibilizar arquivos sonoros com qualidade suficiente para análises acústicas sofisticadas, embora, em alguns casos, a depender das condições da gravação, isso será perfeitamente possível.

Como indicado acima, todos os cuidados metodológicos, recomendados por órgãos internacionais especializados em arquivamento de dados digitais serão considerados no processo de digitalização dos arquivos de áudio, procurando-se, na medida do possível, preservar as características originais do sinal analógico. Quando necessário, técnicas automatizadas de redução de ruído (como, por exemplo, ruídos de pitch fixo – hum e whistles –, associados geralmente a gravações analógicas em fitas magnéticas) serão empregadas. Entre as técnicas mais
comuns de redução de ruídos associados a fitas magnéticas estão a utilização de filtros de frequências.

Experiência prévia de digitalização de arquivos do Projeto NURC, como, por exemplo, a realizada pelo Projeto NURC do Rio de Janeiro, com apoio financeiro do CNPq, demonstra que a proposta aqui apresentada é exequível.

A anotação / etiquetagem do corpus compartilhado do Projeto NURC/Recife será feita a partir da utilização esquemas previamente utilizados com sucesso para o português brasileiro, como, por exemplo, o tagset proposto pelo Núcleo Interinstitucional de Linguística Computacional (NILC), o NILC Tagset (Aires et al., 2000), e o etiquetador morfofissantático MXPOST (Ratnaparkhi, 1996).

5. Contribuições da Proposta

O corpus informatizado será arquivado localmente, nos servidores da Universidade Federal de Pernambuco e da Universidade Federal de Alagoas, em um site dedicado ao Projeto NURC/Recife, para livre consulta pela comunidade científica, e depositado em bancos internacionais, tais como o do IMDI (http://www.lat-mpi.eu/archive/), com o intuito de garantir a sua preservação.

Uma vez constituído e devidamente arquivado, o corpus digitalizado será apresentado aos atuais coordenadores do Projeto NURC, em todas as capitais, como modelo a ser discutido e, eventualmente, adotado, para a informatização e preservação de todo o material coletado por este importante projeto na área da linguística.

6. Referências


Analyzing (-r) with R

Livia OUSHIRO
University of Sao Paulo
Av. Prof. Luciano Gualberto, 403 - Sala 16 - Cidade Universitária - 05508-010 - São Paulo - SP
livia.oushiro@usp.br

Abstract
This paper presents a script written for the free software R (Gries, 2009; Hornik, 2011), which has been employed in the analysis of variable (-r) in Paulistano Portuguese (Oushiro, 2012a) to automatically (i) identify and mark tokens of the variable; (ii) extract tokens into a spreadsheet file precoded with social factors; and (iii) extract a balanced sub-sample of a specific number of tokens per speaker (Wolfram, 1993). It describes the tasks to be performed by R, and discusses the script's advantages and existing shortcomings. The script seems to work better with phonetic and morphological variables, and naturally does not exempt the researcher from a thorough qualitative analysis of their corpus (for example, for identifying possible exclusions). On the other hand, the script can be adapted to studying a number of variables, its different tasks can be performed separately, it allows the researcher to handle data in a more consistent manner and, by reducing the time spent in preparing the token file, it allows more time to perform statistical analyses and interpret results.

Keywords: variable (-r); software R; data handling; Paulistano Portuguese; variationist sociolinguistics.

1. Introduction
Quantitative analyses of sociolinguistic variation (Guy, 1993; Bayley, 2002) often involve handling hundreds or thousands of tokens of a variable, especially in studies of phonetic variation. Analyses in softwares such as GoldVarb X and RBrul should be preceded by the identification, isolation, coding, and extraction of variants within a variable context. These tasks are mechanical, time-consuming, tiresome, and subject to errors due to lapses of attention on the part of the researcher. In fact, there have recently been a number of initiatives for automatizing certain tasks of sociolinguistic quantitative analyses (see e.g. Cieri & Strassel, 2010; Rosenfelder & Labov, 2010).

This paper presents a script written for the free software R1 (Gries, 2009; Hornik, 2011), which was employed in the analysis of variable coda (-r) in Paulistano Portuguese (Oushiro, 2012a) in a corpus of 102 hour-long sociolinguistic interviews (about 1.5 million words), which yielded 63,994 tokens of the variable.

The software R allows researchers to perform a number of tasks, including corpus linguistics data handling, statistical analyses, plotting graphs (Gries, 2009). The aforementioned tasks for handling such a number of tokens were greatly minimized by the use of the software R, which was employed to automatically: (i) identify tokens of variants in the speech of informants; (ii) extract tokens with preceding and subsequent context into a precoded spreadsheet file; and (iii) extract a balanced subsample of a specific number of tokens per speaker (Wolfram, 1993).

The scripts are largely based on Gries (2009) and the internet discussion list “Corpling with R” (https://groups.google.com/group/corpling-with-r). In the scripts below, the relevant functions are in bold. Although it is possible to simply copy the scripts and substitute the relevant variables marked below as "X," the reader is also advised to consult R manuals, such as Gries (2009), since most functions are not described in details here. Section 2 presents the full scripts and discusses some of their main functions; Section 3 discusses their applicability to other sociolinguistic variables and some of its present shortcomings.

2. Full Scripts

2.1 Identifying tokens
This script identifies tokens of a variable in transcript files and marks them with ",". Take the excerpt below as an example:

(1)
S1: não foi assim que eu escolhi a Mooca acho que a Mooca me escolheu [risos] eu [hes.] não foi assim pensado ah eu quero morar naquele bairro porque eu nem’ a minha irmã mora aqui há muitos anos mas eu vinha aqui só a passeio né? mas [hes.] é depois que você muda para cá ai você não quer mais sair não quer mais mudar [sabe]?
D1: [ah que legal]
S1: é bem [hes.] é bem gostoso aqui
D1: então assim [hes.] a senhora diz que morou a maioria/ a maior parte do tempo em São Mateus
S1: é próximo a São Mateus

The task R was to perform was finding the tokens of coda (-r) (e.g. in the words morar, porque, irmã etc.) in the speech of informants (S1), marked here in bold, but not the tokens in the speech of the interviewer (D1) (e.g. maior, parte), marked in italics. The desired output is shown in (2):

(2)
S1: não foi assim que eu escolhi a Mooca acho que a Mooca me escolheu [risos] eu [hes.] não foi assim pensado ah eu quero morar <> naquele bairro

1 Available at: <www.r-project.org>.
porque < eu nem/ a minha irmã <> mora aqui há muitos anos mas eu vinha aqui só a passeio né? mas [hes.] é depois que você muda para cá aí você não quer <> mais sair <> não quer <> mais mudar <> [sabe]? D1: [ah que legal] S1: é bem [hes.] é bem gostoso aqui D1: então assim [hes.] a senhora diz que morou a maioria/a maior parte do tempo em São Mateus S1: é próximo a São Mateus

To do this, the script requires the user to follow the steps 1 to 5 below and, after that, copying and pasting the whole script into R ("#" below is a comment character, which means that R will not read what follows it as functions). The trickiest part is certainly defining the variable (Step 4) in a way that the software R will correctly identify all instances of relevant variants. Identification of variable (r), for example, employed the following specification:

```
(3) \[b.*?[aâæéêioóöúû]\r[bcçdfgjklmnpqstvwxz]\]|\b.*\]?.*\b
```

The line in (3) instructs R to look for instances of a vowel \[aâæéêioóöúû\], followed by the grapheme "r" and a consonant or end of a word \[bcçdfgjklmnpqstvwxz\]\|\b.*\]?.*\b. It further instructs R to respect word boundaries \(\backslash b\), because the intended output should be, e.g., porta <> and not por<>ta. Finally, the symbols .*[b] indicate that there can or cannot be other characters before or after the (vowel-r-consonant) sequence within the same word.

In Script 1, R identifies all .txt files (dir) in the directory specified in Step 2 as the working directory (setwd); creates an empty list (list) into which all the transcript files will be loaded (scan); marks the relevant tokens of the variable specified in Step 4 (gsub) in the speech of the informant S1 with <>; and saves (cat) the transcript files with variant markings in a new directory (setwd) specified in Step 3.

```
#SCRIPT 1

#Step 1. Create 2 folders:
#(i) for the original transcript files (copy files there);
#(ii) for the files which will contain marked tokens.

#Step 2. Specify the directory where the original transcript files are by substituting "X" below for their complete path.
#e.g. "C:/Users/Documents/Transcriptions/"
#Transcript files should be in plain text format (.txt).
originalfiles<-paste("X")

#Step 3. Specify the directory where transcripts with token identification are to be stored by substituting "X" below for their complete path.
```

```R
setwd(originalfiles)
files<-dir(path=originalfiles, pattern=".txt", all.files= T)
all.corpus<-list()
for (i in files) {
  all.corpus[[i]]<-scan(i, what="char", sep="\n", skip=0)
}
for (i in 1:length(files)) {
tokens<-gsub(thevariable, "\\1 <> ", all.corpus[[i]], ignore.case=T)
naos<-paste("[a-rt-z],[S][2-9],[S1: 2.*])<>",
so.S1<-gsub(naos, "\\1", tokens, ignore.case=T)
so.S1<-gsub(naos, "\\1", so.S1, ignore.case=T)
so.S1<-gsub(naos, "\\1", so.S1, ignore.case=T)
so.S1<-gsub(naos, "\\1", so.S1, ignore.case=T)
so.S1<-gsub(naos, "\\1", so.S1, ignore.case=T)
so.S1<-gsub(naos, "\\1", so.S1, ignore.case=T)
markedfiles<-gsub(" +", " ", so.S1)
setwd(markings)
cat(markedfiles, file=files[i], sep="\n")
}
```

```
#END OF SCRIPT 1

2.2 Extracting tokens
```

Script 2 finds tokens of variables (previously identified by "<>") and extracts them into a spreadsheet file with 4 words of the preceding context and 4 words of the subsequent context in the interview. Information of the speaker's social profile is extracted from the transcript file header, according to our research group's transcription norms:

```
D1: Lívia Oushiro
Duração total: 01h14min49seg
Início: 00h00min00seg
```

# Step 4. Specify the variable by substituting "X" below for a 'syntactic definition' of the variable.
thevariable<-paste("X")

# Step 5. Copy and paste this script in R (from "SCRIPT 1" to "END OF SCRIPT 1").
For instance, the second line of this header informs: the year the interview was recorded (2009), the speaker's sex (F), age (63), level of education (M, for "Ensino Médio" or high school), type of school (U for "public"), area of residence (E for "expanded central area"), zone of residence (O for "zona oeste" or West Zone), generation of the family in the city (0, i.e. none of the parents are from São Paulo), family's place of origin (x for "mixed"), and mobility (M for "has lived in different zones of the city"). The other lines respectively inform the speaker's parents' place of origin, the speaker's pseudonym, neighborhood, interviewer, total time of recording, at which point the transcription starts and ends, and additional comments.

Other researchers should naturally be aware of their own transcription norms in their corpora and make the necessary adjustments.

Similarly to Script 1, in Script 2 R first sets the working directory specified in Step 1 (setwd) and identifies all .txt files there (dir) to be loaded into a list (list) by the (scan) function. R then reads the second line of each transcript file as the speaker's social profile (all.tokens[[i]][2]->spkprofile) and breaks this line into characters (substr) to identify the codes for year of recording (characters 5-8 in the second line), sex (character 9), age (characters 10-11) etc. These pieces of information are stored in separate vectors, which will be called later on. After that, R breaks the transcript files into words (gsub("\W","\1", all.tokens[[i]])) and looks for instances of the variants specified in Step 2 (grep). Finally, R creates a new file specified in Step 3, which outputs each token in a different line. Each line also contains the respective speaker's social characteristics separated by a tab stop (\t), four words from the preceding context (tokens2[max(0,i-5):max(0,i-2)]) and four words from the subsequent context (tokens2[(i+1):min(i+4,length(tokens2) +1)]). When opening the newly created file in a spreadsheet program such as Excel or Calc, the tab stops can be read as different columns, thus separating each social variable.

#SCRIPT 2

# Step 1. Specify the directory where the marked transcript files are by substituting "X" below for their complete path.
# e.g. "C:/Users/Documents/Markedfiles/
markings<-paste("X")

# Step 2. Specify what should be extracted to the spreadsheet file.
# if you wish R to extract all tokens, don't change the command line below
# if you prefer to extract tokens of specific variants, substitute "<" for the code of the variants between <> and separate variants by "]":
# e.g. "<R><T>

# Step 3. Define a name for the coding spreadsheet file by substituting "X" below for the name of the file; don't delete ".txt".
# e.g. "DataR.txt"
codingfile<-paste("X.txt")

# Step 4. Copy and paste this whole script in R (from "SCRIPT 2" to "END OF SCRIPT 2").
# The coding file will be in the directory specified in (1) above, with the name specified in (3).
# To open the file in a spreadsheet program (such as Excel or Calc), right-click on the file and choose "Open with... .

#########################################################################

setwd(markings)
files2<-dir(markings, pattern=".txt", all.files=F)
all.tokens<-list()
for (i in files2) {
  all.tokens[i]<-scan(i, what="char", sep="\n", skip=0, encoding="UTF-8")
}
for (i in 1:length(all.tokens)) {
  all.tokens[i][1]<-spkprofile
  substr(spkprofile, 5,8)->year
  substr(spkprofile, 9,9)->sex
  substr(spkprofile, 10,11)->age
  substr(spkprofile, 12,12)->leveled
  substr(spkprofile, 13,13)->typeschool
  substr(spkprofile, 13,13)->region
  substr(spkprofile, 14,14)->zone
  substr(spkprofile, 15,15)->generation
  substr(spkprofile, 16,16)->porigin
  substr(spkprofile, 17,17)->mobility
  all.tokens[i][3]->origin
  all.tokens[i][4]->inf
  all.tokens[i][5]->neighborhood
  all.tokens[i][6]->intr
  tokens1<-gsub("\W","\1", all.tokens[i])
tokens2<-unlist(strsplit(tokens1," "))
  (matches.in.corpus<-grep(token.tokens2, ignore.case=T))
}
for(i in matches.in.corpus) {
  cat("\1", substr(token2[i], 2,2), "\n", year,
  "\n", inf,
  "\n", intr,
  "\n", sex,
Script 3 employs the R package NCStats to randomly select a number of tokens per speaker from the complete coding file, in order to reduce the time spent on coding linguistic factor groups and to avoid data biasing by possible idiosyncratic speakers (Wolfram, 1993). In the analysis of variable coda (r), we selected 50 tokens per speaker, which yielded a final coding file containing 5,100 tokens (50 tokens x 102 speakers), an amount which can be much more easily handled than the original 63,994 tokens.

In this script, after setting the working directory specified in Step 2 and reading the coding file (read.table), R is asked to show the data structure (str). It invokes the package NCStats (library), which will be required because of its srsdf function. It then identifies all speakers in the sample, which had been specified in column 4 in the coding file, by making a list of all the different values (unique); then, for each speaker, R randomly (srsdf) selects a subset of tokens (subset), given the number specified in Step 4, and creates a new coding file with the name specified in Step 5. The .txt file can be opened in a spreadsheet program by right-clicking on it and selecting "Open with...".

#END OF SCRIPT 2

### Sampling

2.3 Sampling

For instance, analyses of diminutives (e.g. menininha 'little girl') (Mendes, 2011) and variable nasal /el (Oushiro, 2012b) in Brazilian Portuguese have employed the following definitions respectively:

(5) \b.*?inh[oa]s?\b
(6)
The line in (5) specifies that R should look for instances of words ending in "-inho," "-inha," "-inhos" or "-inhas". The line in (6) instructs R to look for sequences of the vowel "e", followed by "n" or "m", and by a consonant or end of a word.

The automatization of the task of identifying tokens does not exempt the researcher from a thorough qualitative analysis of his/her corpus. As an example, the definition specified in (5) above correctly identifies tokens such as copinho 'little cup' and bonitinhas 'pretty', but will also mark instances of the words minha 'my, mine', tinha 'had', vinha 'came', which are not diminutives. In this case, the researcher can run Script 1 first as a general survey, make a list of all unwanted tokens, and further exclude words that should not be included by employing the gsub function:

(7)  
clean.files<-gsub(minhas?\|tinha\|vinha) \(\rightarrow\) all.corpus[i], ignore.case=T)

The command line in (7) instructs R to look for instances of minhas <> minhas <> tinhas <> and vinha <> and to substitute them only by the first element(\(\rightarrow\))(minha, minhas, tinha, vinha), i.e. to delete <> in these sequences.

In general, it seems that it is easier to create "syntactic definitions" for phonetic and morphological variables than for syntactic and discourse variables, if one is working with an unannotated corpus. This is because such analyses will probably require further information about the syntactic or discourse function of words in a given corpus. Take, for instance, an analysis of the variable use of Wh-interrogatives in Brazilian Portuguese (Oshiro, 2011), which has four variants: (i) Onde você mora?; (ii) Onde você mora?; (iii) Onde é que você mora?; and Você mora onde? 'Where do you live?' Finding wh-words (o que 'what', que 'what, which', quem 'who', qual 'which', quando 'when', onde 'where' etc.) is not a difficult task, in principle; however, these words do not only function as wh-words, but can also be relative pronouns (Cf. a casa onde morei 'the house where I lived,' o homem que gosta de TV "the man who likes TV") or complementizers (Cf. ele disse que viria 'he said (that) he would come'). An unannotated corpus would miss these distinctions.

On the other hand, these different scripts can be employed independently from one another. If a researcher already has transcript files marked with the relevant tokens, Script 2 can be applied to extract them into a spreadsheet file regardless of Script 1 having been used. If a researcher simply wants to produce a subsample of a larger token file, Script 3 can be applied independently from scripts 1 and 2.

4. Conclusion

The software R enables the sociolinguist to automatize certain mechanical tasks in preparing a coding file for quantitative analyses. The scripts presented here allow the researcher to handle data in a more consistent manner and, by reducing the time spent in preparing the token file, it allows more time to perform statistical analyses and interpret results. Adapting the scripts should allow the analyses of a number of sociolinguistic variables, especially phonetic and morphological ones.

5. Acknowledgements

This research is funded by FAPESP (Grant 2011/09122-6).

6. References


