The annotation of the C-ORAL-BRASIL spoken corpus using an adaptation of the Palavras Parser

Eckhard Bick[1], Heliana Mello[2], Alessandro Panunzi[3], Tommaso Raso[2]
[1]University of Southern Denmark, [2]UFMG, [3]University of Florence
eckhard.bick@mail.dk, heliana.mello@gmail.com, alessandro.panunzi@unifi.it, tommaso.raso@gmail.com

Abstract

This article describes the morphosyntactic annotation of the C-ORAL-BRASIL speech corpus, using an adapted version of the Palavras parser. In order to achieve compatibility with annotation rules designed for standard written Portuguese, transcribed words were orthographically normalized, and the parsing lexicon augmented with speech-specific material, phonetically spelled abbreviations etc. Using a two-level annotation approach, speech flow markers like overlaps, retractions and non-verbal productions were separated from running, annotatable text. In the absence of punctuation, syntactic segmentation was achieved by exploiting prosodic break markers, enhanced by a rule-based distinctions between pause and break functions. Under optimal conditions, the modified parsing system achieved correctness rates (F-scores) of 98.6% for part of speech, 95% for syntactic function and 99% for lemmatization. Especially at the syntactic level, a clear connection between accessibility of prosodic break markers and annotation performance could be documented.

Keywords: morphosyntactic tagging, spoken corpora, constraint grammar.

1. Introduction

The C-ORAL-BRASIL corpus is a Brazilian Portuguese spontaneous speech corpus compiled with the same architecture and the same segmentation criteria as those found in the C-ORAL-ROM corpora for Italian, French, Spanish and European Portuguese (Cresti & Moneglia, 2005). The C-ORAL-BRASIL (Raso & Mello, 2012; Raso & Mello, 2010; Mello & Raso, 2009; Raso & Mittmann, 2009), at present, offers its informal component (208,130 words in 139 texts), divided in family/private context (159,364 words) and public context (48,766 words). In each context, the corpus is equally divided into monologues, dialogues and conversations.

In the C-ORAL corpus, prosodic segmentation was marked explicitly, at transcription time, using three different segmentation strengths:

1. major prosodic breaks (//), separating what functionally could be called utterances, which are considered the reference units for spoken language analysis;
2. discontinuation breaks (+) between utterances;
3. major prosodic breaks (//), separating what functionally could be called utterances, which are considered the reference units for spoken language analysis.

This paper will briefly present the procedure used for the automatic PoS, morphological and syntactic tagging of the corpus and the obtained results. As an earlier work on the Italian section of C-ORAL-ROM pointed out (Panunzi et al., 2004), one of the main spoken-specific tasks for automatic PoS tagging deal with the relation between the segmentation of speech flow and the disambiguation procedure. About these aspects, the work done on C-ORAL-BRASIL corpus introduces some crucial innovations that lead to a sensible improvement of performance in all annotation levels considered.

2. The tool

The PoS, morphological and syntactic tagging for C-ORAL-BRASIL corpus has been carried out using the Palavras parser (Bick, 2000) as a point of departure. Palavras is a Constraint Grammar (CG) parser that is mostly used for the annotation of written data. With lexical adaptation and various filter programs, the parser has been already used for non-standard language varieties, such as historical texts (Bick & Módolo, 2005) and the NURC speech corpus (Bick, 1998).

The Constraint Grammar paradigm (Karlsson et al., 1995), which the Palavras parser adheres to, can be described as a dualism of a robust, modular disambiguation methodology for Natural Language Processing (NLP) on the one hand, and a linguistic-descriptive convention on the other hand, encoding linguistic analyses as token-based tags and function-mediated dependency structures. Both the method and the descriptive tradition offer a number of formal advantages for the annotation of non-standard language data such as speech.

First, because CG systems have a modular architecture with a clear separation of lexica, analyzers and grammars (rule sets) for successive levels of analysis, it is relatively easy to add specialized lexica or morphological filters, as well as add specific grammar modules.

Second, CG's token-based annotation, where even higher-level structural information is strictly token-based, allows a corpus project to maintain several layers of annotation in parallel (such as discourse markers as opposed to clause boundaries), even allowing rules handling one layer to make reference to tags from another

layer.
Technically, the Palavras parser is a chain of Constraint Grammar rule sets, successively handling ever higher (deeper) levels of analysis, progressing from morphological disambiguation and PoS tagging, over syntactic function mapping and dependency relations, to semantic role annotation, Named Entity Recognition and application-oriented modules. Input to this chain of grammars is provided by a preprocessor/tokenizer and a morphological analyzer program, supported by large lexica covering inflexional paradigms, valency potential, semantic class ontologies etc. All lexical information is encoded, CG-style, as token-linked tags on reading lines. Ambiguous reading lines for a given word are called a *cohort*.
PALAVRAS uses about 6.000 contextual CG rules that either remove, select, add, map or substitute tags/readings for ambiguous tokens. Given the general architecture and the rule methodology of the parser, three main tasks can be identified with regard to its application to oral data, affecting lexical recall on the one hand and contextual disambiguation on the other:

1. the text flow normalization, which includes the treatment of corpus meta information from the non-grammatical annotation layers (speaker names, overlapping and disfluency phenomena);
2. the treatment of non-standard word forms;
3. the definition, in the absence of ordinary punctuation, of the reference units that can provide delimited windows for contextual disambiguation of both PoS and syntactic dependencies.

3. Text flow normalization

C-ORAL-BRASIL uses a number of symbols and encoding conventions to handle data flow issues like turn taking, prosodic breaks, speaker overlap, retractions and interruptions. Such encoding is either in non-alphanumeric form (<, /, +), or not part of an utterance (speaker names), so they either cannot or must not be analyzed by the parser. To both maintain this meta-information and to provide text-only input to the parser, we opted for a two-level annotation, where meta-information is "stored" in angle brackets on separate lines as corpus meta-markup. PALAVRAS' annotation is transparent to such markup and will not change, remove or try to analyze it.
The main issue regarding the text-flow normalization is related to the treatment of the speech disfluency, i.e. the retracting phenomena and non-word occurrences. Retractions are manually pre-marked in the original transcripts at the start point of the retraction, providing the precise number of retracted words. Given this, our pre-processor module only needs to eliminate the words in question from the surface level to enable much smoother syntactic parses. Word repetitions or self-corrections, if allowed to persist at the surface level, would be problematic for CG rules at all levels, interfering not only with the implementation of linguistic universals like the uniqueness principle, but also with word class adjacency and agreement rules (see also Panunzi et al., 2004). The same procedure is used for so-called non-words (paralinguistic elements and incomplete words).
A special complication arose from the fact that overlap and retraction markings can be nested and/or overlapping as in the following example (with <...> for overlap, &.. for nonwords and [/..] for retraction):

**GIL: <eu &a [/2] eu acho que é> esse [/2] é esse aqui o'*
// <&he> + (I don't think that this here, see)

which required careful ordering of string matches, for instance to prevent retractions from getting "invisible" within (de-texted) speaker overlap markers. Also, since overlaps and non-words can appear within the scope of a retraction, they would change the latter's word count if removed too early, and possibly affect real words further to the left.

4. Non-standard words

In order to assign a morphological tag string and word class hypothesis, PALAVRAS tries to recognize unknown words as either (1) affix-derivations or (2) variations of standard forms, or a combination of both.
For the C-ORAL-BRASIL project, however, ordinary standardization was deemed not to be enough, first of all because certain oral word forms were transcribed in a phonetic fashion *as is*, creating in some cases unrecoverable differences from standard orthography, or the risk of ambiguity. As a side consideration, we also wanted to account for lexical gaps due to dialectal or otherwise rare forms. Therefore, two new modules were added to PALAVRAS' program chain, both with a manually maintained lexicon-file as input.
The first program (*coral.inter*) handles specific or systematic standardizations and is run after pre-processing, before morphological analysis. About 700 normalizations were listed in a special lexicon file, and though the standard analyzer could have handled a certain proportion on its own in terms of word class, the lexicon treatment also allowed us to add correct base forms or even semantic classification. A very phonetic example are abbreviations *(emedebê → MDB)* where even plural *(emeeles → ml)* forms and non-standard pronunciation *(emitivi → MTV)* were covered. Other groups concern non-standard inflexion *(fazido → feito, fize → fiz)* and derivation *(espim → espinhos, ladim → ladinho).* Finally, word-initial changes like a-drop (b2-4) had to be covered in order to prevent such forms from being guessed as (most likely) singular nouns.
Some normalizations are expressed as multi-word expressions *(a'=aqui -> olha=aqui, c'=ocês -> com=vocês)*, which both allows more complex rewritings and some implicit contextual disambiguation. Examples

of systematic, lexicon-independent normalisations are the treatment of s-drop in 1. person plural verb endings, or allowing plural forms of interjections. The normalized forms are fed to the Palavras analyzer module, while the original forms are stored as <OALT:....> metatags, invisible to the parser.

The second program (*postlex_pt*) is regular morphological analyzer in its own right, with its own lexicon and inflexion rules, overriding PALAVRAS' own analysis, removing the error risk created by heuristic readings. It allows both fullform and base form entries in its lexicon (*newlex_pt*). In the actual lexicon (2000 entries), due to the good coverage of PALAVRAS, there are very few regular Portuguese nouns, and those there are could mostly have been recognized by PALAVRAS' derivational analysis *(e.g. fazeção N F S)*. However, the lexicon resolves inflexional ambiguity *(e.g. caça-talentos* as plural) and prevents wrong endings-based analysis of foreign words. 2/3 of the total were proper nouns *(Tinina, Timoftol)*. Though relatively easy to identify (uppercasing), only lexical treatment of proper nouns will allow parsing rules to make reference to features like gender or semantic type (e.g. <org>, <hum> etc.).

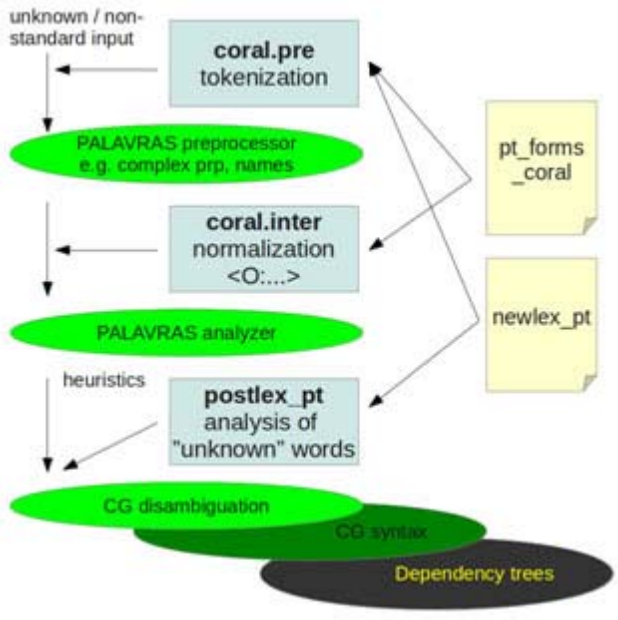

Fig.1: Speech genre adaptation modules in Palavras.

Because of potential overlap with inflected forms of other words, C-ORAL-specific lexicon additions were not allowed to override ordinary morphological analysis, but rather added to readings cohorts *before* contextual disambiguation.

5. Syntactic speech flow segmentation

While written language data provide paragraph markers, line breaks, full stops and other punctuation to deduce syntactic and informational structure, such segmentation is not explicit in spoken language transcriptions. At the surface level, speech data lacks punctuation and has unclear sentence and clause boundaries. The necessary information to segment speech resides in prosody (i.e. rhythm, stress and intonation) as well as nonverbal signals. Depending on whether and how this information is encoded in the transcription, a parser may simply lack the segmentational information to work properly (Bick, 1998). Syntax should be even more affected than PoS/morphological tagging by the absence of boundaries, since long-distance contexts are more important for capturing syntactic relations (Bick, 2000).

One can conclude that for the annotation of speech data it is paramount to provide the parser with some kind of delimiter clues concerning clause and phrase structure, if its global rules are to work optimally.

In the C-ORAL-BRASIL tagging, the pre-existing markup of prosodic breaks has been exploited in order to provide this kind delimiter. The // “major break” was substituted with a semicolon, while the / “secondary break” was re-tagged as a comma with two potential readings, <break> and <pause>. Only the former represents a syntactic break, while the latter is allowed inside phrases and between verb and complement. CG-rules (such as a-c below) were written to distinguish between these two readings, and run as a separate prosodic segmentation module before Palavras' ordinary disambiguation and syntax modules:

(a) between a noun or a nominative pronoun or a conjunction to the left, and a finite verb to the right, a prosodic /-marker is treated as <pause> (S+V case);

(b) prosodic /-markers between a noun and another np are treated as <break> (appositions);

(c) prosodic /-markers between a prenominal and its head are treated as <pause> (np cohesion, e.g. ART+N).

Contextually disambiguating the function of prosodic breaks allowed us to strike a balance between simply ignoring such markup on the one hand, and syntactic over-segmentation on the other.

6. Evaluation

In order to evaluate the modified parser on our data, one transcription file (bfamdl15) was chosen at random, automatically analyzed and hand-corrected. We then used the Constraint Grammar evaluation tool *eval_cg* to compare the raw analysis file with the revised version. In an ordinary CG setup, meta-markup and punctuation would align 100%, but in our case, matters were complicated by the fact that “commas” had been disambiguated as either break or pause, and in the latter case replaced with a meta-tag. On the one hand, this caused alignment problems for the evaluator, on the other hand, differences had to be identified and counted as recall errors. Other mismatches, caused by faulty splitting or non-splitting of ambiguous MWE's, were also counted as recall errors, e.g in the case of *“primeiro=que”* (conjunction vs. adjective/numeral + relative). Including

"punctuation" tokens, the file contained 1895 word tokens.

	Recall	**Precision**	**F-Score**
Syntactic function	95.3	94.9	95
PoS / Word class	98.5	98.7	98.6
Morphology	98.4	98.6	98.5
Base form	98.6	99.4	99

Table 1: Performance evaluation.

It can be seen from these figures that the easiest task was lemmatization (base forms), while syntactic function was the most difficult. The difference between recall and precision for syntax is a measure of remaining ambiguous tags. For word class and morphology, only one reading was allowed, so the precision-recall differences are entirely due to differences in matching differences between break markers (commas).

In order to judge the effectiveness of using prosodic break markers as punctuation, we compared the standard run (with pause/break disambiguation) with a no-break run (/-marks ignored), a no-utterance run (both /, + and // ignored), and an all-break run (all /-marks turned into commas, without disambiguation). The following table shows the global F-score obtained for the evaluation of each run.

	no sentence	**no break**	**all break**	**pause/ break**
Syntactic function	86.2 R: 86.5 P: 86.1	90.7 R: 91.0 P: 90.6	93.7 R: 93.3 P: 93.6	**95.0** R:95.3 P: 94.8
PoS	98.3	98.8	99.3	**99.4**
Morphology	98.1	98.6	**99**	98.7
Base form	99	99.1	**99.4**	**99.4**

Table 2: Segmentation-based performance.

Clearly, exploiting prosodic break markers did improve performance at all levels. However, the effect was much more marked for syntax than for part of speech, lemmatization and morphology, reflecting the wider contextual scope of syntactic tags and the ensuing greater need for precise and correct segmentation. Interestingly, while syntactic performance can be further increased by pause/break disambiguation, this is not obvious for the more local tag categories. Thus, for inflexion tags (morphology), all-break performance was higher than for the pause/break run, and only for part of speech a slight improvement was observed.

7. Conclusion

While the C-ORAL Brasil annotation project has shown that a standard written-language parser (PALAVRAS) can be used to assign morphosyntactic tags to transcribed speech data, it also demonstrated that for optimal performance, certain adaptations should be made to both the system and the data, comprising some orthographical normalization and lexicon extensions, as well as syntactic segmentation. The latter proved especially important for syntax, and was achieved by exploiting prosodic break markers as "punctuation", enhaced by a rule-based distinctions between pause and break functions. Under optimal conditions, the modified parsing system achieved correctness rates (F-scores) of 98.6% for part of speech, 95% for syntactic function and 99% for lemmatization. Because of the non-uniform distribution of errors across category types, manual linguistic revision has been performed on some segments of the corpus, and the identified error patterns will be used to calibrate Palavras' contextual rules to achieve both better consistency and a better error balance in future reruns.

The implemented annotation scheme does preserve the original prosodic-transcriptional information, including speech flow, retractions, overlaps, turntaking etc., encoded as meta-tagging alongside the morphosyntactic tags, but it remains a future task to figure out an integrated search formalism (GUI interface) that would allow the user to work with both these two different levels of annotation at the same time. Finally, we foresee the addition of higher, semantic levels of annotation, such as case roles or anaphora (both in principle available for written-language Palavras parses), as well as the integration of the latter with the ongoing manual information-structural tagging.

8. References

Bick, E. (2000). *The Parsing System Palavras - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus: Aarhus University Press.

Bick, E. (1998). Tagging Speech Data - Constraint Grammar Analysis of Spoken Portuguese. In *Proceedings of the 17th Scandinavian Conference of Linguistics* (Odense 1998).

Bick, E. and Módolo, M. (2005). Letters and Editorials: A grammatically annotated corpus of 19th century Brazilian Portuguese. In C. Pusch, J. Kabatek & W. Raible (Eds.) *Romance Corpus Linguistics II: Corpora and Historical Linguistics* (Proceedings of the 2nd Freiburg Workshop on Romance Corpus Linguistics, Sept. 2003). Tübingen: Gunther Narr Verlag, pp. 271--280.

Cresti, E. and Moneglia, M. (2005). *C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam/Philadelphia: John Benjamins.

Karlsson, F.; Voutilainen, A.; Heikkilä, J. and Anttila, A.

(1995). *Constraint Grammar, A Language-Independent System for Parsing Unrestricted Text*. Berlin: Mouton de Gruyter.

Mello, H. and Raso, T. (2009). Para a transcrição da fala espontânea: o caso do C-ORAL-BRASIL. *Revista Portuguesa de Humanidades - Estudos Linguísticos*, 13(1), pp. 153--178.

Panunzi, A.; Picchi, E. and Moneglia, M. (2004). Using PiTagger for Lemmatization and PoS Tagging of a Spontaneous Speech Corpus: C-ORAL-ROM Italian. In M.T. Lino et al. (Eds.), *Proceedings of the 4th LREC Conference*, vol. 2. Paris: ELRA, pp. 563--566.

Raso, T. and Mello H. (2010). The C-ORAL BRASIL corpus. In: M. Moneglia & A. Panunzi (Eds.), *Bootstrapping Information from Corpora in a Cross-Linguistic Perspective*. Firenze: FUP, pp. 193--213.

Raso, T. and Mello, H. (2012). *C-ORAL-BRASIL I: Corpus de referência do português brasileiro falado informal*. Belo Horizonte. Editora UFMG.

Raso, T. and Mittmann, M. (2009). Validação estatística dos critérios de segmentação da fala espontânea no corpus C-ORAL-BRASIL. *Revistas de Estudos da Linguagem*, 17(2), pp. 73--91.