Parsing Speech Data:
The Automatic Grammatical Annotation of the C-ORAL-Brasil Corpus

(1000-word abstract, anonymous)

1. Introduction

1.1. Background

While linguistic interest in transcribed speech corpora has grown considerably in recent years, accessibility is often hampered by the lack of standardized markup and systematic searchability. Optimally, the necessary annotation should include not only phonetic issues, prosody, discourse structure etc, but also traditional morphosyntactic annotation. In this paper we will focus on how to integrate the latter with the former, and discuss the question whether and how a tagger-parser primarily designed for written language can be adapted to handle transcribed speech data. The work was carried out in the research context of the C-ORAL speech corpus project for Brazilian Portuguese (Raso & Mello 2010), where morphosyntactic annotation was to be added automatically on top of an existing meta-annotation in the face of non-standard orthography and the absence of punctuation, preserving in-text speech flow markers etc.

Using automatic annotation, either on its own or as a pre-step for manual revision, is an obvious choice for a corpus this size (~ 300.000 words). Thus, previous European C-ORAL sister projects employed statistical part of speech taggers for this task, such as the PiTagger system (Moneglia et al 204) for the Italian section, which had access to a lexicon-based analyzer, a standard lexicon (107.00 lemmas), a training corpus (50.000 words) and a special pre-dictionary covering about 2000 non-standard and dialectal forms. For the European Portuguese section, the Brill tagger (Brill 1993) was used, trained on a written Portuguese corpus of 250.000 words. While no higher-level, syntactic annotation was attempted in the European C-ORAL, other speech corpus projects have opted for full treebank annotation, such as the Arabic treebank describe by Maamouri et al. (2010), which combined manual selection of analyzer suggestion, followed by an automatic syntactic parsing stage.

1.2. Constraint Grammar parsing environment

For our own work we used 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, but has demonstrated great robustness in the face of genre variation - as, for instance, in the Linguateca project (linguateca.pt) and the CorpusEye corpora (corp.hum.sdu.dk). With lexical adaptation and various filter programs, the parser has also been used for non-standard language varieties, such as historical texts (Bick & Módolo 2005). The Constraint Grammar paradigm (Karlsson 1995) can be described as both a robust, modular disambiguation methodology for NLP, and a linguistic-descriptive convention, 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). Several speech annotation projects have made use of these advantages, such as Müürisep & Uibo (2006) for Estonian. In the Nordic Dialect Corpus (Bondi et al. 2009), CG output was used to train a DTT tagger (Schmid 1994). In the European C-ORAL context, the Spanish section employed CG-

inspired rules for part-of-speech disambiguation of morphological output from the GRAMPAL system (Moreno 2003), and for the Palavras parser itself, Bick (1998) reports early experiments with a Constraint-Grammar-only solution in connection with the morphosyntactic annotation of the Brazilian NURC corpus (Castilho 1993).

1. Project methodology

Given the rule-based and lexicon-dependent architecture of PALAVRAS, three challenges can be identified with regard to its application to oral data, affecting lexical recall on the one hand (2.2) and contextual disambiguation on the other (2.1 & 2.3). In many ways, the problems are similar to the ones encountered in the annotation of historical language data (Bick & Módolo 2005).

2.1. Text flow normalization

In order to maintain corpus meta information from other annotation layers, while still providing "running text" input to the PALAVRAS-analyzer, in-text markup for turn-taking, speaker overlap and retractions was turned into <....> meta tags reminiscent of xml tags but without the projectivity restrictions of xml-trees. Tokenization was also standardized, for instance by resolving non-standard contractions *(prum, naquea)* into propositions and pronouns, allowing the parser to match ordinary np and pp constraints.

2.2. Lexical and orthographic normalization

While maintaining the oral transcription forms as tokens, modified word forms were fed to the analyzer module where transcriptional orthography deviated from the written norm *(emitivi, ladim, estudemo)*. For the cases where systematic rules could not be used, we created a both a standardization dictionary (used by a preprocessor), and a list of ready-analyzed non-standard forms (used by a post-processor). In both cases, both multi-word expressions and regular inflexional variation was covered on top of individual word forms.

2.3. Syntactic segmentation

A serious problem for the automatic analysis of transcribed speech is the lack of syntactic surface structure encoded as punctuation. To solve this problem, and to provide the parser with syntactic windows for its rules (such as the uniqueness principle), we exploited C-ORAL's prosodic markup, inserting sentence breaks and "comma candidates" instead. Contextual CG rules were used to determine if the latter should be treated as a (syntactically neutral) pause, or as a real syntactic break.

3. Evaluation

Under optimal conditions, measured against a randomly chosen and manually corrected 2000-token gold-standard section of the corpus, the modified parser achieved correctness rates (F-scores) of 98.6% for part of speech, 95% for syntactic function and 99% for lemmatization. Experiments with unsegmented input showed that the use of prosodic breaks reduced syntactic error rates by two thirds, and PoS by half. However, the added effect of pause/break disambiguation affected only syntactic tags, not PoS tags, reflecting the two tag types' unequal dependence on long-distance contexts.

References

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

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

Bick, Eckhard & Marcelo Módolo. 2005. Letters and Editorials: A grammatically annotated corpus of 19th century Brazilian Portuguese. In: Claus Pusch & Johannes Kabatek & Wolfgang Raible (eds.) *Romance Corpus Linguistics II: Corpora and Historical Linguistics (Proceedings of the 2nd Freiburg Workshop on Romance Corpus stics, Sept. 2003)*. pp. 271-280. Tübingen: Gunther Narr Verlag.

Brill, Eric. 1992. A simple rule-based part of speech tagger. In: Proceedings of the workshop on Speech and Natural Language. HLT '91, Morristown, NJ, USA: Association for Computational Linguistics, pp.112–116

Castilho, Ataliba de (ed.), 1993. Gramática do Português Falado, vol.3, Campinas: Editora da Unicamp.

Johannessen, Janne Bondi, Joel Priestley, Kristin Hagen, Tor Anders Åfarli, and Øystein Alexander Vangsnes. 2009. The Nordic Dialect Corpus - an Advanced Research Tool. In Jokinen, Kristiina and Eckhard Bick (eds.): Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009. NEALT Proceedings Series Volume 4

Karlsson, Fred & Voutilainen, Atro & Heikkilä, Juka & Anttila, Arto. 1995. Constraint Grammar, A Language-Independent System for Parsing Unrestricted Text. Berlin: Mouton de Gruyter.

Maamouri, Mohamed et al. 2010. From Speech to Trees: Applying Treebank Annotation to Arabic Broadcast News. In: Proceedings of LREC 2010, Valletta, Malta, May 2010.

Moreno, A. & J.M. Guirão. 2003. "Tagging a spontaneous speech corpus of Spanish". In: Proceedings of the International Conference on Recent Advances in Natural Language Processing.). Borovets, Bulgaria, 2003. p. 292-296.

Müürisep, Kaili and Uibo, Heli (2006). "Shallow Parsing of Spoken Estonian Using Constraint Grammar". In: P.J.Henriksen & P.R.Skadhauge, Proceedings of NODALIDA-2005 special session on treebanking. Copenhagen Studies in Language #33/2006

Moneglia, M., A. Panunzi, E. Picchi, 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, ELRA, Paris, pp. 563-566.

Raso, Tommaso & Heliana Mello. 2010. The C-ORAL BRASIL corpus. In: Massimo Moneglia & Alessandro Panunzi (eds): Bootstrapping Infromation from Corpora in a Cross-Linguistic Perspective. Universitá degli studi di Firenze, Biblioteca Digitale.

Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of the International Conference on New Methods in Language Processing 1994. pp. 44-49.