Prosodic, syntactic, semantic guidelines for topic structures across domains and corpora

Ana Isabel Mata[1], Helena Moniz[1,2], Telmo Móia[1], Anabela Gonçalves[1], Fátima Silva[3], Fernando Batista[2,4], Inês Duarte[1], Fátima Oliveira[3], Isabel Falé[1]

[1]CLUL/FLUL – Universidade de Lisboa
[2]INESC-ID Lisboa
[3]CLUP – Universidade do Porto
[4]ISCTE – Instituto Universitário de Lisboa
{aim, A.Goncalves, iduarte}@fl.ul.pt, {helena.moniz, fernando.batista}@inesc-id.pt,
{tmoia, foliveir}@netcabo.pt, mhenri@letras.up.pt; imsfale@gmail.com

Abstract

This paper presents the annotation guidelines applied to naturally occurring speech, aiming at an integrated account of contrast and parallel structures in European Portuguese. These guidelines were defined to allow for the empirical study of interactions among intonation and syntax-discourse patterns in selected sets of different corpora (monologues and dialogues, by adults and teenagers). In this paper we focus on the multilayer annotation process of left periphery structures by using a small sample of highly spontaneous speech in which the distinct types of topic structures are displayed. The analysis of this sample provides fundamental training and testing material for further application in a wider range of domains and corpora. The annotation process comprises the following time-linked levels (manual and automatic): phone, syllable and word level transcriptions (including co-articulation effects); tonal events and break levels; part-of-speech tagging; syntactic-discourse patterns (construction type; construction position; syntactic function; discourse function), and disfluency events as well. Speech corpora with such a multi-level annotation are a valuable resource to look into grammar module relations in language use from an integrated viewpoint. Such viewpoint is innovative in our language, and has not been often assumed by studies for other languages.

Keywords: speech annotation, topic structures, European Portuguese

1. Introduction

Studies on prosody-syntax-discourse interface relations based on naturally occurring speech are gaining growing interest. Corpora annotated with all these levels of linguistic information are not very common in general (e.g., Calhoun *et al.*, 2010 and references therein). In European Portuguese, in particular, a first attempt was done for a subset of the CORAL corpus, collected by Trancoso *et al.*, 1998 and Viana *et al.*, 1998. However, multi-level annotations were not time-aligned with the speech signal, and consequently the corpus remains almost unexplored in what regards all the above mentioned linguistic interfaces.

Under the auspices of a National Project (COPAS), a multidisciplinary team, formed by linguists from each of the linguistic areas involved (prosody, syntax, discourse) and speech processing engineers gathered to discuss and create guidelines to better describe a small set of structures that challenge the architecture of grammar generally assumed in Theoretical Linguistics – namely structures involving "discourse-driven" activation of the peripheries (left and right dislocations; clefts) – and to automatically process all the time-linked levels of information. Our approach also innovates insofar as its baseline is spontaneous speech instead of constructed examples or experimentally-induced speech alone.

In this paper we present a multi-level annotation scheme for left periphery structures using a small sample of highly spontaneous speech in which the distinct types of topic structures are represented. The analysis of this sample provides fundamental training and testing material for further applications in a wider range of domains and corpora.

This paper is organized as follows: section 2 introduces the corpora used in our study. Section 3 presents an overview of the annotation scheme for prosody, syntax and semantic layers, summarizing the main sets of features and values for each linguistic level. Section 4 describes the automatic tasks involved in corpora processing. Section 5 presents a first analysis of the multi-level annotations from a small sample of the CPE-FACES corpus. Finally, section 6 presents our final remarks and trends for future work.

2. Corpora

The **CPE-FACES corpus** consists of spontaneous and prepared unscripted speech from 25 teenagers (14-15 years old) and 3 adults, all speakers of Standard European Portuguese (Lisbon region), totaling approximately 16h. The corpus was collected in the last year of compulsory education (9th grade), in three Lisbon public high schools. In the spontaneous situation, teenagers and adults were unexpectedly asked to relate a (un)pleasant personal experience. The prepared situation corresponds to typical school presentations, about a book the students must read following specific programmatic guidelines. For students, a variety of presentations on Ernest Hemingway's "The Old Man and the Sea" and on Gil Vicente's "Auto da Índia" was recorded. As for the teachers, all prepared presentations are related to the study of "Os Lusíadas" by Luís de Camões, and two address the same episode - the

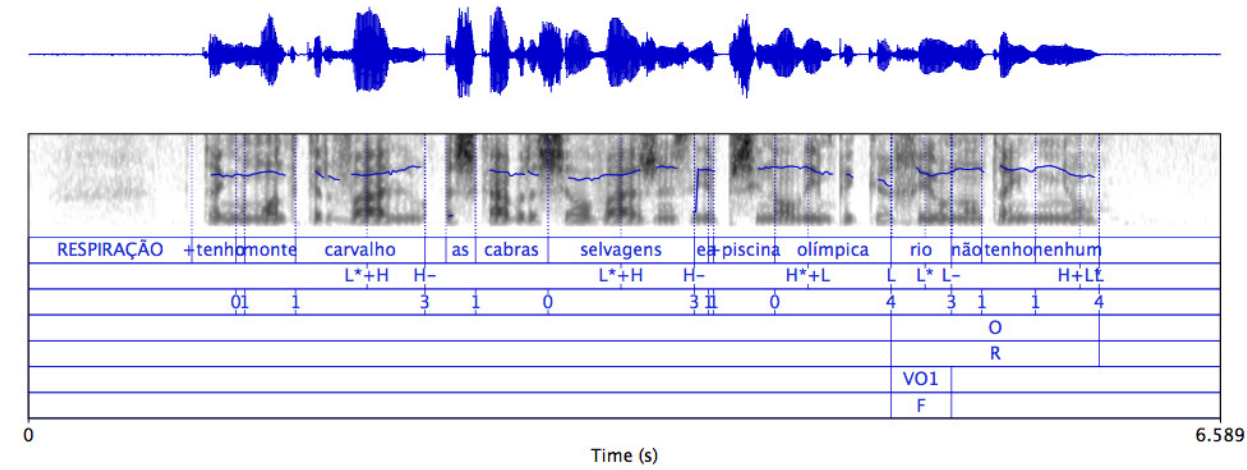

Figure 1. L* L- used with a familiar topic marked as O.

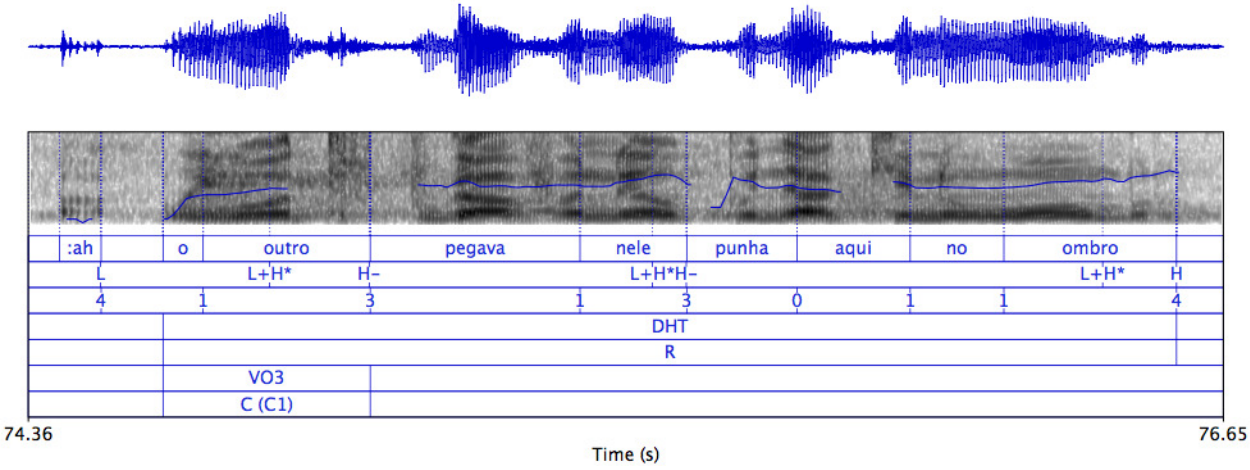

Figure 2. L+H* H- used with a contrastive topic marked as DHT.

lyric-tragic episode of Inês de Castro (for more detailed information on CPE-FACES *vide* Mata, 1999; Mata *et al.*, these proceedings).

The **CORAL corpus** (Viana et al, 1998; Trancoso *et al.*, 1998) has 64 dialogues in Map-Task format by 32 speakers, totaling 9 hours (61k words). One of the participants (the giver) has a map with some landmarks and a route drawn between them; the other (the follower) has also landmarks, but no route and consequently must reconstruct it. In order to elicit conversation, there are small differences between the two maps: one of the landmarks is duplicated in one map and single in the other; some landmarks are only present in one of the maps; and some are synonyms (*e.g. curvas perigosas* /dangerous curves vs. *troço sinuoso* /sinuous stretch). In the 16 different maps, the names of the landmarks were chosen to allow for the study of connected speech phenomena. Speakers were chosen to achieve an adequate gender balance, but were restricted in terms of age (under-graduate or graduate students) and variety (Standard European Portuguese, Lisbon region). Furthermore, they were chosen in pairs who knew each other, so that half of the conversations took place between friends and half between people who did not know each other.

In order to have a more comprehensive view of the grammatical phenomena under study, other corpora available online – not annotated within project COPAS – are being used. The **Corpus CETEMPúblico** 1.7 v. 7.2, in http://www.linguateca.pt/acesso/, stands out among those. It contains around 180 million words from one of the main Portuguese daily newspapers (*Público*), and provides a solid anchor for comparison, inasmuch as it portrays standard relatively formal registers of contemporary written language.

3. Guidelines

Our working corpus of left/right dislocations and clefts currently comprises around 700 occurrences, selected from subsets of the CPE-FACES and CORAL corpora encompassing 19 spontaneous and 19 prepared unscripted presentations (from 3 adults and 16 teenagers) and 20 spontaneous dialogues. As previously mentioned, for the purpose of the present study, we will focus on topic structures from a representative sample of such structures taken from CPE-FACES. They were multilayered annotated for prosody, syntax, and semantics, after performing the forced alignment of data (phone, syllable,

and orthographic word levels). Thus, the outcome of this multilevel annotation comprises: 1 orthographic tier, plus 2 prosodic tiers (tone; break-index) and 5 syntax-discourse annotation tiers (construction type; construction position; syntactic function; discourse function), all time-linked with the orthographic tier. See Figure 1 (with the excerpt: *tenho o monte carvalho, as cabras selvagens e a piscina olímpica. RIO, não tenho nenhum/* I have the Oak hill, the wild goats and the Olympic pool. RIVER, I don't have any, extracted from the CORAL corpus) and Figure 2 (with the excerpt: *O OUTRO, pegava nele, punha aqui no ombro/* THE OTHER, I would take it and put it here on my shoulder, extracted from the CPE-FACES corpus).

All annotations rely on orthographic transcripts and on listening to contextualized speech examples using Praat (Boersma & Weenink, 2013). Team members of each area independently label intonation and syntactic-discourse information. Furthermore, in order not to bias the results, prosodic and syntactic-discourse annotations are done in separate files. When all files are completed, they are merged into single TextGrids.

The applied guidelines will be described in the next sections.

3.1 Prosody

The prosodic guidelines comprise two main tiers: tones and breaks. The tone tier displays intonation contours decomposed into high (H) and low (L) tones, stemming from Pierrehumbert (1980) and Beckman & Pierrehumbert (1986). The break tier displays the analysis of perceived disjuncture between words, building upon the work of Price *et al.* (1991). The tone tier, as established for MAE ToBI (Silverman *et al.*, 1992), consists of pitch accents (associated with accented syllables) and boundary tones (associated with phrase boundaries). Phrase boundaries correspond to two types: minor phrases (marked with the diacritic "-") and intonational phrases (marked with the diacritic "%"). Pitch accents can either be simple or bitonal (*e.g.*, L*, H*, L+H*, L*+H). The star * diacritic marks the tone associated with the accented syllable and the diacritic "!" is used whenever the H pitch range is compressed, resulting in a !H label.

The target structures selected from CPE-FACES and CORAL corpora were annotated with the ToBI prosodic system adapted to European Portuguese (Towards a P_ToBI by Viana *et al.*, 2007). All the pitch accents (H+L*, H*+L, L*+H, L+H*, H*, L*, H+!H*) and the final boundaries (L%, H%, !H%, LH%, HL%) that are covered in that proposal were used. (Schematic F0 contours for pitch accents and boundary tones are presented in Mata *et al.*, these proceedings.)

In the break tier, break indices range from 0 to 4. The level 0 corresponds to the strongest link between words and it marks a high co-articulation between two consecutive words, *e.g.*, in European Portuguese it could be the index for a sequence like ['tESt 6'gOr6] (test know) with the ellipsis of the schwa vowel, instead of ['tESt@ 6'gOr6][1]. The level 1 is the common index between two connected words within a phrase. The level 2 stands for dubious interpretations (either perceived as a break 1, but displaying tonal and lengthening cues; or perceived as 3 or 4, but without a tonal boundary). Break levels 3 and 4 represent a minor phrase and an intonational phrase boundary, respectively. Additionally the diacritic "p" is also used for marking disfluent disjuncture between words.

3.2 Syntax

The (mainly) syntactic part of the annotation involves three sets of values: (i) construction type (*i.e.* the type of construction where the left dislocated constituent appears), (ii) construction position (concerning the root or embedded position of the construction, in the context where it appears), and (iii) syntactic function (*i.e.* the syntactic function of the left dislocated constituent in the relevant construction). As for the construction type, 5 major structurally distinct types are considered, following descriptions for Portuguese in the literature (namely, Duarte, 1987; 2003). These are: HT (hanging topic), DHT (left-dislocated hanging topic), DC (clitic left-dislocation), T (topicalization) and WT (wild topicalization) – cf., among others, Cinque (1977; 1983; 1990), Ross (1967), Zribi-Hertz (1986). A sixth possibility is envisaged – marked as O (other) – for structures that (to a greater or lesser extent) diverge from these five core types. As for the syntactic function, 7 distinctions are regarded, including subject and several types of complements and modifiers. For the special case where the left dislocated constituent does not have a (clear) functional role in the sentence, the additional value N (none) is included. Furthermore, for subjects and verb complements, we distinguish whether they appear in (or are associated with) embedded positions – attaching the prefix E to the syntactic function label.

3.3 Semantics

With regard to the semantic annotation, it involves one set of values: the discursive functions of the left dislocated constituents at stake on the analysis. In order to set the relevant discursive functions, two main features regarding information structure were considered, the dichotomies topic/focus and given/new, both considered from a semantic point of view, based on the proposals of several authors (see overall summary of these works in Frascarelli & Hinterhölzl, 2002). Four main discursive functions, labeled as topics, were established, being two of them divided into two subtypes: i) continuing topic – CONT (Givón, 1983), refers to the continuity of a certain topic throughout conversation after being introduced in the previous discourse; ii) familiar topic – F (Chafe, 1987), which bears some resemblance to the previous topic, consists of a topic that is recognized by the interlocutor because it is accessible in the discourse in spite of not

[1] Transcriptions are given in SAMPA.

being previously introduced; iii) shifting topic – SHIFT (Gívon, 1983) occurs when there is a shifting of the current topic or the introduction of a new topic, occurring according two modalities: Rough-shifting topic -SHIFT(RSHIFT), when the speaker brings momentarily to the discourse a new entity, and Smooth-shifting topic -SHIFT(SSHIFT), when the speaker introduces a new entity with the intention to keep it active in the following segment of discourse; iv) contrastive topic – C (Kuno, 1976; Büring, 1999), when there is a contrastive value or the presence of a feature that induces an alternative, being specified through two subtypes: contrastive 1 – C1 (Kuno, 1976; Büring, 1999), when the contrast is done by the opposition of two elements, *e.g.*, x ou y; constrastive 2 – C2 (Calhoun, Nissin, Steedman & Brenier, 2005), when contrast is obtained on the basis of the choice of a set of options available in the context or in speakers' knowledge.

3.4 Inter-annotator agreement

In order to calculate the inter-annotator agreement in terms of prosody, 57 files were annotated by two annotators. A Fleiss' kappa (Fleiss, 1971) of 71.8% was achieved for both pitch accents and boundary tones, and 93% for break indices. These results compare well with consistency metrics evaluated for other languages (see Escudero *et al.*, 2012 and references therein). According to the table proposed by Landis and Koch (1977), there is a *substantial agreement* for tones and an *almost perfect agreement* for break indices. Table 1 presents more detailed statistics on this process (http://dfreelon.org/utils/recalfront/recal3/).

	breaks	tones
n cases	900	729
average pairwise percent agreement	95.78%	76.13%
Fleiss' kappa	92.97%	71.78%
FK observed agreement	95.78%	76.13%
FK expected agreement	39.91%	15.41%
average pairwise Cohen's kappa	92.98%	71.95%
Krippendorff's alpha	92.98%	71.80%

Table 1. Agreement between two annotators for prosody.

	Ctype	Cpos	SFunc
n cases	25	25	25
avg pairwise agreement	100.0%	100.0%	68.0%
Fleiss' kappa	100.0%	100.0%	61.8%
FK observed agreement	100.0%	100.0%	68.0%
FK expected agreement	24.5%	78.9%	16.2%
avg pairwise Cohen's kappa	100.0%	100.0%	62.2%
krippendorff's alpha	100.0%	100.0%	62.6%

Table 2. Agreement between annotators for syntax.

We also evaluated the inter-annotator consistency in terms of syntax, comparing 25 files from two annotators. Table 2 presents the details concerning the three different tiers, namely: construction type, construction position, and syntactic function. The latter achieves a Fleiss' kappa of 61.8%, showing that this information is considerably less consistent than the other ones, which achieve 100%.

4. Corpus processing

Along with the orthographic word level, an automatic speech recognition (ASR) system (Neto *et al.*, 2008) was used for producing force aligned transcripts. The motivation for using the ASR in force aligned mode instead of a fully automatic speech recognition concerns to the fact that current ASR models were trained for the Broadcast News domain, and the poor results obtained with an out-of-domain recognizer would not be suitable for our study. Force aligned transcripts are in everything similar to fully automatic transcripts, but without recognition errors, and provide useful additional information, such as phones and syllables.

The manual annotation is being performed using Praat (Boersma & Weenink, 2013), a multi-platform open-source tool, which incorporates a vast number of features for speech analysis, labeling and annotation. The segmentation from the orthographic tier serves as a starting point for the remainder annotation tiers. The prosodic, syntactic and semantic annotations, being performed by different groups, are all time aligned with the speech signal, making it possible to produce a final database where all the information can be related.

5. Sample analysis

Using on a small sample of the CPE-FACES corpus, syntactic analysis has focused on 116 syntactically annotated structures. The first results show that HT, DHT, DC and WT are relatively infrequent. The predominant constructions are O (58/116), corresponding either to structures where left-periphery is indicated merely by intonation (with no effects on basic word order) or to structures where (intonationally marked) relatively high adjuncts are placed in sentence-initial position, and T (36/116).

As for the syntactic function, the left dislocated constituent corresponds more frequently to subject (45/116); in most of these cases involving subjects, only prosodic aspects seem to be at stake (whence the structure has been annotated as O, wrt construction type). Modifiers occupy a left-peripheral position in 32 out of 116 structures (essentially in T and O constructions). Verb objects are comparatively less frequent (22/116), with most occurrences concerning direct objects (13/116), mainly in T.

Semantic analysis as also focused on 116 semantically annotated discourse segments. The results of this exploratory study show: 19 continuing topics, which corresponds to approximately 16.4% of the overall examples; 28 familiar topic, which corresponds to approximately 24.1%; 25 shifting topic, corresponding to 21.6%, from which 14 were smooth-shifting topic and 11, rough-shifting topic; 44 contrastive topic, corresponding to 37.9% of the corpus, divided in 33 contrastive 1 topic,

and 11 contrastive 2 topic. This classification shows that speakers organize the discourse information following different strategies that allow them to pursue their communicative goals.

Focusing on the same data set, the analysis of intonation patterns shows that the majority of topic structures (83.6%) are phrased independently from the rest of the utterance and that H+L* L% - the canonical contour of declaratives in EP – is fairly infrequent in these structures. Furthermore, topics are not associated to a single intonation pattern. L+H* (virtually absent from lab speech in EP) and H* (unusual in nuclear position in lab speech, in standard EP) frequently appear in T/DHT constructions associated with a contrastive value (see Figure 2, for L+H* H- in a contrastive topic marked as DHT). H*+L (associated with focus in lab speech), even though less common in the target structures examined, also appears in some examples associated to contrastive and shifting topics. As for O constructions, a large contour variation is observed (see Figure 1, for L* L- used with a familiar topic marked as O). A thorough grammatical analysis of these structures, from a prosodic, syntactic and semantic point of view, is currently under investigation.

6. Final remarks and future work

This paper has presented a multi-level annotation process that is being applied across corpora and domains in European Portuguese. All linguistic annotations are time-aligned with the speech signal and all linguistic levels can be analyzed in relation to each other. This is a key development that allows us to investigate prosody-syntax-semantic interactions in European Portuguese.

In this paper we focused on topic structures selected from a small sample of highly spontaneous speech. The exploratory analysis of this sample, coded for syntactic structure, discourse function, prosodic prominence and phrasing in a time-aligned way, provided fundamental training and testing material for further application in a wider range of domains and corpora.

7. Acknowledgments

This work was supported by national funds through FCT - Fundacão para a Ciência e a Tecnologia, under project COPAS – PTDC/CLE-LIN/120017/2010. Fernando Batista is supported by ISCTE – Instituto Universitário de Lisboa.

8. References

Beckman, M. and Pierrehumbert, J. (1986). Intonational structure in Japanese and English. *Phonology Yearbook 3*, Cambridge, CUP, pp. 255-309.

Boersma, P. and Weenink, D. (2013). Praat: doing phonetics by computer [Computer program]. Version 5.3.56, retrieved 15 September 2013 from http://www.praat.org/

Büring, D. (1999). Topic. In Bosch, P. and van der Sandt, R. (eds.). *Focus. Linguistic, Cognitive and Computational Perspectives*. Cambridge: CUP, pp. 142-165.

Calhoun, S., Nissim, M., Steedman, M. and Brenier, J. (2005). A Framework for Annotating Information Structure in Discourse, pp.1-8, available at http://groups.inf.ed.ac.uk/switchboard/infostruc.pdf

Calhoun, S., Carletta, J., Brenier, J. M., Mayo, N., Jurafsky, D., Steedman, M., and Beaver, S. (2010). The NXT-format Switchboard Corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources & Evaluation* (2010) 44, pp. 387–419.

Chafe, W. L. (1987). Cognitive constraints on information flow. In R. S. Tomlin (Ed.) *Coherence and grounding in discourse*. Amsterdam: John Benjamins, pp.21-52.

Cinque, G. (1977). The movement nature of left dislocation. *Linguistic Inquiry* 8(2): 397-411.

Cinque, G. (1977). 'Topic' constructions in some European languages and 'Connectedness'. In Ehrlich and van Riemsdijk (orgs.), *Connecteness in Sentence, Discourse and Text.* Tilburg: KLUB.

Cinque, G. (1990). *Types of A'-dependencies*. Cambridge, Massachusetts: The MIT Press.

Duarte, I. (1987). *A Construção de Topicalização na Gramática do Português: Regência, Ligação e Condições sobre Movimento*. PhD Thesis, University of Lisbon.

Duarte, I. (2003). Frases com tópicos marcados. In Mateus, M. H. M. *et al. Gramática da Língua Portuguesa*. Lisboa: Editorial Caminho. 5.ª edição revista e aumentada, pp. 489-502.

Escudero, D., Aguilar, L., Vanrell, M. del Mar, and Prieto, P., (2012). Analysis of inter-transcriber consistency in the Cat_ToBI prosodic labeling system. *Speech Communication*, volume 54, issue 4, pp. 566-582.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, volume 76, no. 5, pp. 378–382.

Frascarelli, M. and Hinterhölzl, R. (2002). Types of topics in german and Italian. In Winkler, S. and Schwabe, K. (eds.). *On information structure, meaning and form*. Amsterdam/Philadelphia: John Benjamins, pp.1-29.

Givón, T. (1983). *Topic continuity in discourse: a quantitative cross-language study*. Amsterdam: John Benjamins.

Kuno, S. (1976). Subject, theme, and the speaker's empathy – A reexamination of relativization phenomena. In Li, C. (ed.) *Subject and Topic*. New York: Academic Press, pp. 1-29.

Landis, J. and Koch, G., (1977). The measurement of observer agreement for categorical data. *Biometrics*, volume 33, pp. 159–174.

Mata, A. I. (1999). *Para o Estudo da Entoação em Fala Espontânea e Preparada no Português Europeu: Metodologia, Resultados e Implicações Didácticas*. PhD Thesis, University of Lisbon.

Mata, A. I., Moniz, H., Batista, F., and Hirschberg, J. (2014). Teenage and adult speech in school context: building and processing a corpus of European Portuguese. These Proceedings, *LREC 2014*, Reykjavik, Iceland.

Neto, J., Meinedo, H., Viveiros, M., Cassaca, R., Martins, C., and Caseiro, D. (2008). Broadcast news subtitling system in Portuguese. *ICASSP 2008*, pp. 1561–1564.

Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation*. PhD thesis, MIT.

Price, P., Ostendorf, M., Shattuck-Hufnagel, S. and Fong, G. (1991). The use of prosody in syntactic disambiguation. *Journal of the Acoustic Society of America*, 90 (6), pp. 2956–2970.

Ross, J. R. (1967). *Constraints on variables in syntax*. PhD. dissertation. Cambridge, Massachusetts: MIT.

Silverman, K., M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert and Hirschberg, J. (1992). ToBI: a standard for labeling English prosody. In *Proceedings of ICSLP 92*, Banff, volume 2, pp. 867-870.

Trancoso, I., Viana, M. C., Duarte, I., and Matos, G. (1998). "Corpus de diálogo CORAL". In *PROPOR'98*, Porto Alegre, Brazil.

Viana, M. C., Trancoso, I., Mascarenhas, I., Duarte, I., Matos, G., Oliveira, L. C., Campos, H., and Correia, C. (1998). Apresentação do Projecto CORAL - Corpus de Diálogo Etiquetado. In *Workshop I de Linguística Computacional*, Lisboa, Portugal.

Viana, C., Frota, S., Falé, I., Fernandes, F., Mascarenhas, I., Mata, A. I., Moniz, H. and Vigário, M. (2007). Towards a P_ToBI. *PAPI2007. Workshop on the Transcription of Intonation in Ibero-Romance*. University of Minho, Portugal.

Zribi-Hertz, A. (1986). *Relations anaphoriques en Français: esquisse d'une grammaire générative raisonée de la reflexivité et de l'ellipse structurale*. Thèse de Doctorat, Université de Paris-VIII.

Walker, M. A. and Prince, E. F. (2003). A Bilateral Approach to Givenness: A Hearer-Status and a Centering Algorithm. In Fretheim, T & Gundel, J. (eds.). *Reference and referent accessibility.* Amsterdam/Philadelphia: John Benjamins, pp. 291-306.