Teenage and adult speech in school context: building and processing a corpus of European Portuguese

Ana Isabel Mata[1], Helena Moniz[1,2], Fernando Batista[2,3], Julia Hirschberg[4]
[1]FLUL/CLUL – Universidade de Lisboa
[2]INESC-ID Lisboa
[3]ISCTE -Instituto Universitário de Lisboa
[4]Columbia University
aim@fl.ul.pt, {helena.moniz, fernando.batista}@inesc-id.pt, julia@cs.columbia.edu

Abstract

We present a corpus of European Portuguese spoken by teenagers and adults in school context, CPE-FACES, with an overview of the differential characteristics of high school oral presentations and the challenges this data poses to automatic speech processing. The CPE-FACES corpus has been created with two main goals: to provide a resource for the study of prosodic patterns in both spontaneous and prepared unscripted speech, and to capture inter-speaker and speaking style variations common at school, for research on oral presentations. Research on speaking styles is still largely based on adult speech. References to teenagers are sparse and cross-analyses of speech types comparing teenagers and adults are rare. We expect CPE-FACES, currently a unique resource in this domain, will contribute to filling this gap in European Portuguese. Focusing on disfluencies and phrase-final phonetic-phonological processes we show the impact of teenage speech on the automatic segmentation of oral presentations. Analyzing fluent final intonation contours in declarative utterances, we also show that communicative situation specificities, speaker status and cross-gender differences are key factors in speaking style variation at school.

Keywords: Teenage and adult speech, oral presentations, European Portuguese

1. The Corpus

The CPE-FACES corpus consists of spontaneous and prepared unscripted speech from 25 students (14-15 years old) and 3 teachers (2 female and 1 male), all speakers of Standard European Portuguese (Lisbon region), totaling approximately 16h. The corpus was designed to represent some of the speech tasks that are common in school context and it was collected in the last year of compulsory education (9th grade), in three Lisbon public high schools. It was recorded in a natural setting – the speakers classroom of Portuguese as L1 – in different communication situations: two pair dialogues (both spontaneous) and two oral class presentations (one spontaneous and another one prepared, but unscripted). In the spontaneous presentation, students and teachers were unexpectedly asked to relate a (un)pleasant personal experience. It was assumed that the involvement of speakers on topics related to their personal interests and day-to-day life would manifest in the naturalness and spontaneity of their talks (Labov, 1976). 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 lyric-tragic episode of Inês de Castro – thus allowing for a comparison between speakers skills and discursive strategies when talking about the same (or similar) topics.

Basically, spontaneous and prepared presentations differ in the degree of planning involved, the type of information communicated, the speakers' attention to the speech task and effort to speak clearly. In spontaneous presentations, the speakers can talk freely about any topic of their choice; they can change topic and move on to another topic whenever they feel like it. As far as typical (prepared) oral presentations at school are concerned, it was argued before that "more than talking about a pre-determined theme, an oral presentation presupposes the capacity to individually produce a greater amount of utterances, organizing the information that is given to the public in a clear structured form" (Mata, 1999).

The metadata collected for each student includes: age and gender, birthplace, area of residence and number of years in the area of residence, proficiency level in the discipline Portuguese (L1) during the last Cycle of Education (*i.e.* the last three years), as well as some background indicators (level of education and profession of parents). Additionally, information about the teachers' qualifications and professional experience was also collected.

The recordings of the two female teachers and all the students were done with an UHER 400 Report Monitor recorder with a BASF LPR 35 magnetic tape and a SENNHEISER MD 214 V-3 worn suspended from the neck microphone. These recordings were latter digitized at 44.1 kHz, using 16 bits/sample and afterwards downsampled to 16 kHz. CPE-FACES was recently extended with the recordings of a male teacher, using a TASCAM HD-P2, a Portable High Definition Stereo Audio Recorder, and a head-mounted microphone Shure, a Sub-miniature Condenser Head-worn Microphones, model Beta 54. The sound was recorded in mono, with 16-bit at samples rates of 44.1kHz, and afterwards downsampled to 16 kHz.

Speakers		Useful Time (min)		#Words		#Disfluent Sequences		#SUs (,/./:/;/?/!)		%Alignment Error	
		Prep.	Spont.	Prep.	Spont.	Prep.	Spont.	Prep.	Spont.	Prep.	Spont.
Boys	B1	4.35	3.32	910	571	8	4	60	42	0.0	19.9
	B2	4.16	2.94	698	477	18	9	28	29	0.0	0.0
	B3	4.33	3.98	923	840	15	8	39	46	1.7	1.0
Girls	G1	3.81	0.82	508	119	12	7	56	24	0.0	0.0
	G2	1.75	2.40	393	574	10	4	61	58	8.2	0.0
	G3	15.26	2.41	3756	539	17	15	29	48	1.1	10.4
Teachers	A1	27.28	3.83	4356	622	89	14	1018	130	0.0	0.0
	A2	30.77	1.76	5087	315	204	16	1042	42	0.0	0.0
	A3	58.50	5.48	9564	928	334	85	1867	122	0.3	0.1
Total		150.21	26.94	26195	4985	707	162	4200	541		
		177.2 (3h)		31180		869		4741		1.5	

Table 1. Overall characteristics of the selected subset of 9 speakers (A3 is the male adult)

2. Transcripts

Manual transcripts include word-by-word time aligned orthographic transcripts, enriched with punctuation marks, disfluency and paralinguistic events (clicks, breathing, cough, laughs…). Punctuation marks were added to the corpus by a teacher of Portuguese, following the guidelines of Duarte (2000). Disfluencies were annotated according to Moniz (2006; 2013), adapted from Shriberg (1994) and Eklund (2004). As for paralinguistic events, it was applied the inventory reported in Mata (1999).

Additionally, a subset of the corpus was recently annotated with the ToBI prosodic system (Silverman *et al.*, 1992) adapted to European Portuguese (*Towards a P_ToBI* by Viana *et al.*, 2007), in order to conduct experiments on automatic ToBI-labeling in European Portuguese (Moniz *et al.*, 2014), as part of an ongoing project (COPAS – PTDC/CLE-LIN/120017/2010).

The subset used comprises 9 spontaneous presentations and 9 prepared unscripted presentations, from 6 students (3 boys and 3 girls) and 3 teachers (2 female and 1 male). Table 1 summarizes the subset characteristics in terms of useful time (measured in minutes, silences not included), number of words, number of delimiting punctuation marks and disfluent sequences, and percentage of alignment error. This subset comprises 3h of useful speech, around 31k words and 900 disfluent sequences, about 4700 sentence like-units (SUs) – in the sense of Jurafsky and Martin (2009) – and 1.5% of overall alignment error.

In the eighteen oral presentations, a total of 44.2 minutes (27.2 minutes of useful speech) was prosodically annotated. Since students' presentations vary from about 1-15 minutes and teachers' from about 2-58 minutes, the shortest presentations were fully annotated. As for the longer ones, only the initial part was included (approximately 1 minute for teenagers and 5 minutes for adults).

The inventory of pitch accents and final boundary tones that were used in this subset is illustrated in Figure 1.

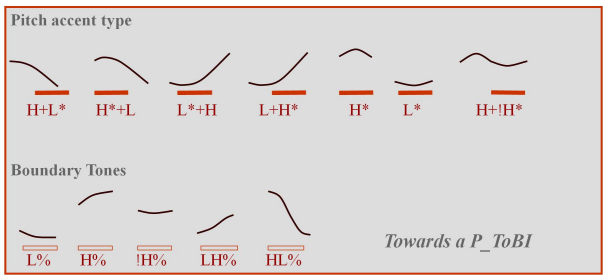

Figure 1. Schematic F0 contours for pitch accents and boundary tones, taking into account *Towards a P_ToBI.* (Thick lines ▬ indicate the stressed syllable.)

ToBI break indices 4 and 3, accounting for two levels of intonational phrasing in European Portuguese (Frota, 2000; Viana *et al.*, 2007), as well as the corresponding tonal labels for boundaries, "%" and "-", were applied to CPE-FACES in order to allow for a comparison of prosodic patterns across different break strengths. Disfluent phrase breaks (corresponding to filled pauses, prolongations, repetitions, substitutions, deletions, insertions) were marked with 'p', following ToBI guidelines. (For more details about the prosodic annotation guidelines, see Mata *et al.*, these proceedings.)

3. Inter-annotator agreement

In order to calculate the inter-annotator agreement, we compared the results of 57 files (with 900 break index marks) from two annotators. The agreement was measured in terms of Fleiss' kappa (Fleiss, 1971). 71.8% agreement was achieved for both pitch accents and boundary tones, and 93% for break indices, which compares well with ToBI inter-transcriber consistency for other languages (see Escudero *et al.*, 2012, and references therein). According to the scale proposed by Landis and Koch (1977), *by no means* universally accepted, but commonly used by the scientific community (see, for example, Escudero *et al,* 2012), there is a *substantial agreement* for pitch accents and boundary tones and an *almost perfect agreement* for break indices. Table 2

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 2. Agreement between two annotators.

Concerning break indices, Figure 2 shows the confusion matrix between the two annotators (A and B). As expected, due to the high agreement, the diagonal of the matrix is well defined and the most salient values relate to breaks 1 and 4. However, the matrix also reveals that annotator B marked 10 cases of break 1 as break 3 and replaced 10 cases of break 3 with break 4.

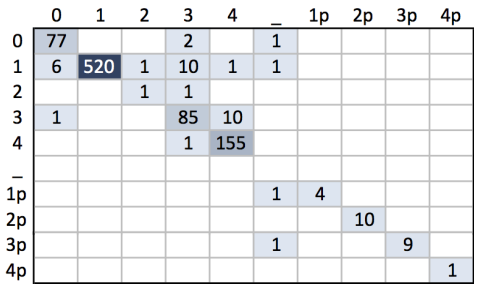

Figure 2. Confusion matrix for breaks.

The confusion matrix for tones is represented in Figure 3. The most salient disagreement is related to the number of tones that were annotated by A and not by B (indicated by _ in the first vertical column of the figure).

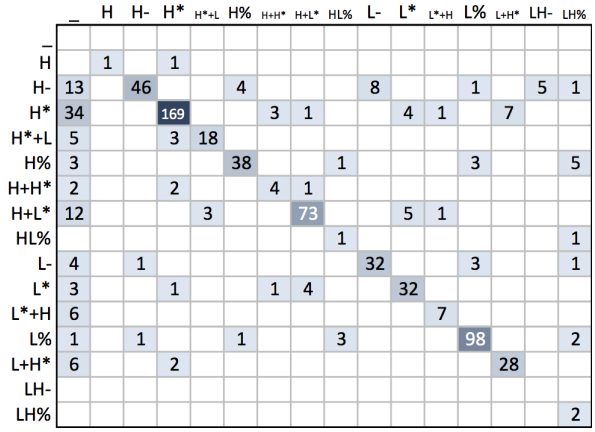

Figure 3. Confusion matrix for tones.

4. Processing Tasks

Up to now we have concentrated our attention on the presentations subset of the data. The automatic prosodic feature extraction process involves synchronizing the manual prosodic labels with complex sets of acoustic features. The synchronization process relies on information coming from the automatic speech recognition system (ASR) output, from manual transcripts, and from the signal itself. After producing the ASR transcripts, all manual annotations are transferred to the ASR transcripts, including all the prosodic labels (tones and breaks), by means of word alignments. With this method, relevant prosodic information is assigned to different units of analysis, including intonational units, words, syllables and phones. Durations of words, syllables and phones were derived from the ASR output. Whereas information regarding pitch (f_0) and energy (E), not available in the ASR pipeline when this study started, has been directly extracted from the speech signal, using the Snack toolkit (Sjölander *et al.*, 1998).

Acoustic-phonetic parameters of segmental and supra-segmental units were automatically extracted to study intonational events in EP. Organizing such information into hierarchies, meaning, into the smallest unit of analysis (phones or even sub-phone units) up to higher order constituents was crucial to the experiments conducted, making it possible to automatically extract a complex set of acoustic measures. At this point, the information extracted encompassed phones, syllables, words, intonational phrases, and speech-acts.

The speech recognizer (Neto *et al.*, 2008), trained for the broadcast news domain, is unsuitable for the oral presentations domain and even more for teenage speech. Therefore the ASR was used in a force alignment mode, in order not to bias the study with the bad results obtained with an out-of-domain recognizer. The force aligned transcripts still contain about 1.5% of unaligned words (mainly due to overlapping speech and to phonetic-phonological processes at phrase-final position, further detailed in the next section).

The features were extracted for intonational phrases, either break 3 or 4, involving the final word itself and the adjacent contiguous words or the syllables within those units. Features involving a single word/syllable encompass: pitch and energy slopes; ASR confidence score; word/syllable duration; number of syllables and phones. Features involving two consecutive words/syllables encompass: pitch and energy slopes and shapes; pitch and energy differences; word durations, pitch medians, and energy medians. Pitch slopes were calculated based on Semitones rather than Hertz. Slopes in general were calculated using linear regression. Pitch and energy shapes are based on slope values and expand to 9 binary features, assuming one of the following values {*RR, R-, RF, -R, --, -F, FR, F-, FF*}, where *F*=Fall (if the slope is negative), -=plateau (if the slope is near zero), *R*=Rise (otherwise). For a more detailed analysis vide Batista *et al.* (2012).

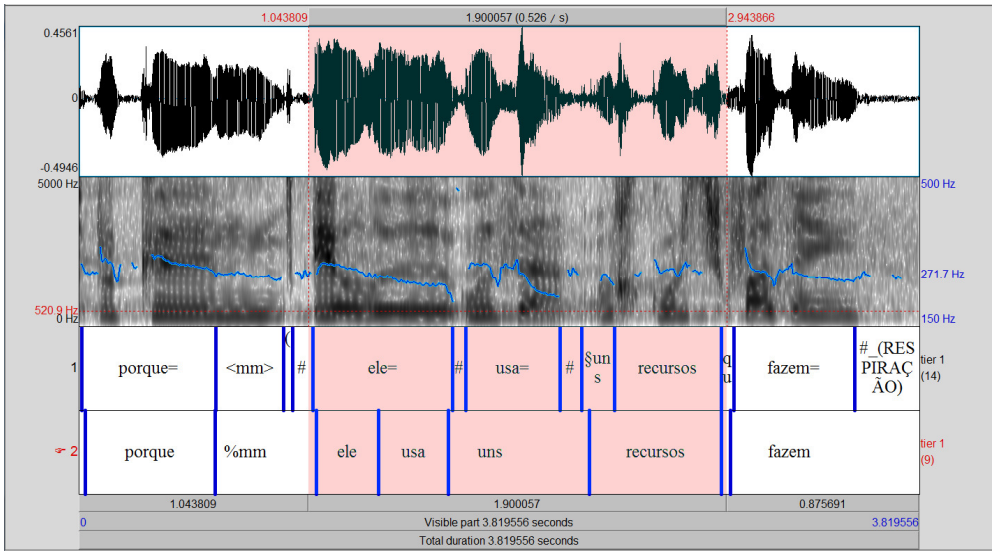

Figure 4: Example of erroneous segmentation of a sequence with disfluencies. The first row corresponds to the manual segmentation whereas the second is the automatic one.

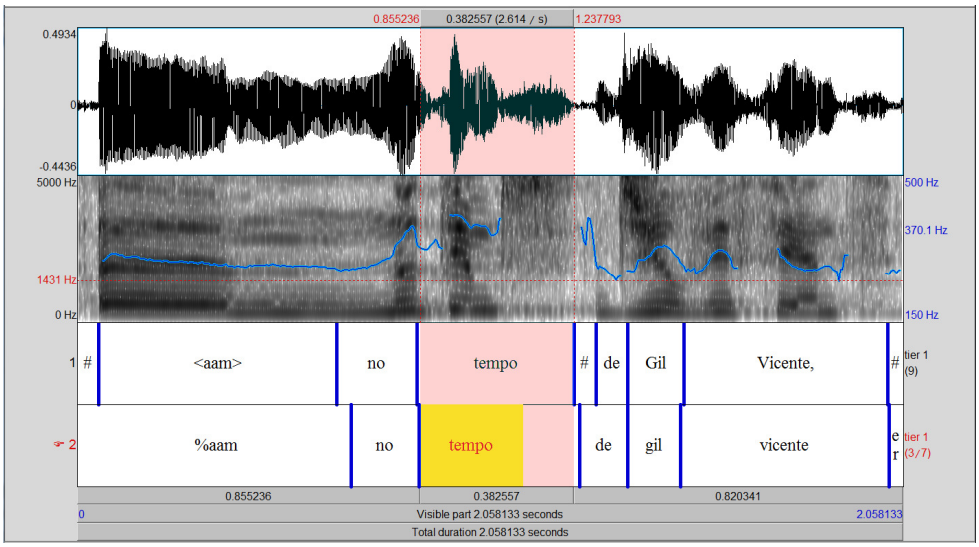

Figure 5: Example of erroneous segmentation of "tempo", pronounced with final vowel elision and plosive frication.

5. Disfluencies and phrase-final effects: impact on automatic segmentation

The most common automatic erroneous segmentation is due to disfluencies, especially excessive prolongations and filled pauses, as illustrated in Figure 4 (with the following example extracted from a prepared presentation by a female teenager: *porque= <%mm> ele usa uns recursos que fazem/* because *um* he uses some resources which make). This figure shows the erroneous segmentation of a sequence with excessive prolongations in word-final position (marked with "=" in the first row) and a filled pause "mm".

Additionally, several phonetic-phonological effects occurring in phrase-final position also trigger erroneous segmentations: i) aspiration; ii) creaky voice; iii) frication of plosives; iv) vowel epenthesis (in EP [@]); v) and truncation of final segmental material. Aspiration and creaky effects are common in the data, especially in teenagers' data.

Regarding frication, Figure 5 (with the excerpt *<%aam> no tempo de Gil Vicente/ um* at Gil Vicente's era, extracted from the same presentation) illustrates the frication of [p] in the word "tempo" – the last syllable in post-stressed position, corresponding to plosive and vowel, occurs with the final vowel elided and the remaining plosive with frication effects (a burst and unvoiced fricative spectra). In our data set, these frication effects seem to consistently occur at the end of intonational phrases and may be a phonetic cue to prosodic phrasing. The epenthetic vowel [@] also causes miss-detections, promoting either the recognition of a new word or even the segmentation of two different words. Regarding truncated segmental material, this process is the most common one and also triggers erroneous identification of the word itself and of the adjacent contexts. The use of alternative pronunciations could be a possible solution for covering the above mentioned

processes.
EP is a language known to extremely reduce and frequently delete unstressed vowels, particularly in post-stressed and phrase-final position. Exploring the prosodically annotated subset of the corpus, the data shows that this effect varies across speech types with apparent inter-speaker differences: it is overall more frequent in teenagers' presentations than in adults' presentations, particularly in the spontaneous situation.

6. Final intonation contours: spontaneous vs. prepared presentations

All the pitch accents (H+L*, H*+L, L*+H, L+H*, H*, L*, H+!H*) and the final boundary tones (L%, H%, !H%, LH%, HL%) that are covered in the *Towards a P_ToBI* proposal were used in the annotated subset of CPE-FACES. The analysis of fluent final intonation contours (N=1552) in declarative utterances shows that different pitch accent and boundary tone combinations occur in the data. Low/falling accents can be combined with high boundary configurations (*e.g.* L* H%, H+L* LH%) and high/rising accents with low boundary configurations (*e.g.* H* L%, L+H* HL%). However, simple combinations of low/falling tones and high/rising tones (*e.g.* (H)+L* L%; (L+)H* H%, cf. Figure 6) are the most common in both spontaneous and prepared presentations.

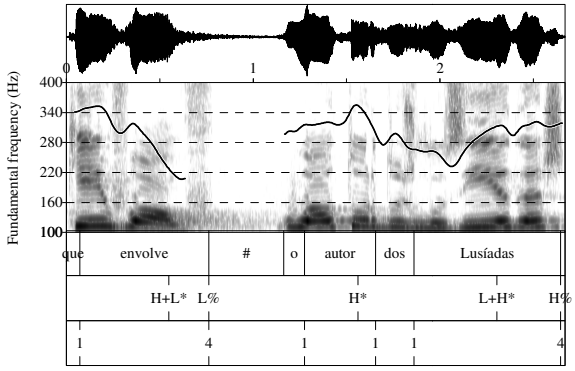

Figure 6. Example of H+L* L% and L+H* H% in the excerpt: *que envolve o autor d'Os Lusíadas*/that involves the author of *Lusíadas* (extracted from a female adult presentation).

Together, H+L* L, L* L, H* H and L+H* H account for 70% of all phrase-final contours – 70.4% in the spontaneous and 69.4% in the prepared subset. Furthermore, their frequency distributions vary with ToBI break level ($p < .001$): low/falling tones are more frequent with break 4, the predominant level of intonational phrasing across presentations; high/rising tones are more frequent with break 3. Final word pitch slopes and standard deviation measures (cf. Section 4) allow a distinction between these final contours. A statistical analysis confirms that these results are highly significant ($p < .001$, U= -10.414, -8.183, respectively for normalized slopes and standard deviations across L+H* H and H* H; $p < .001$, U= 5.789, -9.437, respectively for normalized slopes and standard deviations across H+L* L and L* L). As Figure 7 shows, although with varying degrees across phrase levels, H* H and L* L differ mainly in pitch slope ($p < .001$); L+H* H and H+L* L contrast both in pitch slope and standard deviation ($p < .001$; $p < .01$). Significantly higher values in both pitch features discriminate H+L* L and L+H* H from L* L and H* H, respectively.

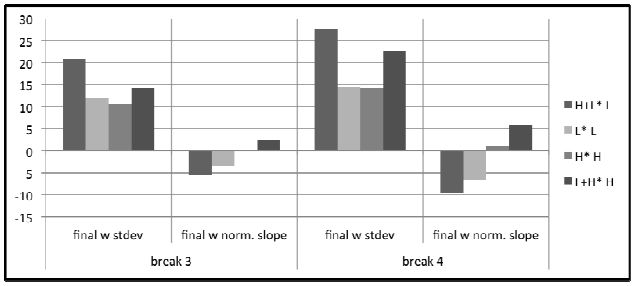

Figure 7. H+L* L, L* L, H* H and L+H* H contours: final word pitch slopes and standard deviations.

The analysis of fluent final intonation contours in declarative utterances also shows that the specificity of oral presentations and the speaker status/age affect the distribution of intonation patterns in phrase-final position: unlike adults (the teachers), teenagers (the students) increase non-low patterns (mainly (L+)H* H) across phrase levels in the prepared presentations; adults differ from teenagers by using more non-low patterns in the spontaneous presentations. This analysis also shows that gender-related differences are a source of variation in the balance of phrase levels across spontaneous and prepared presentations – the contrast between minor phrases and intonation phrases is stronger for boys than for girls; this is also observed for female adults.
In addition, an analysis of the acoustic features introduced in Section 4 points to significant differences regarding to: (i) spontaneous vs. prepared speech; (ii) speakers' status; and (iii) speakers' gender. Pitch slopes, maximum, minimum and standard deviations extracted for the word in phrase-final position are the most salient indicators of these overall differences. Although pitch slopes within the accented syllable are also significant, they are not as consistent as the features extracted at the word level. Overall, prepared speech displays significantly higher values than spontaneous speech in pitch slope, maximum, minimum and standard deviation at the level of the phrase-final word ($p < .01$ for pitch minimum; $p < .001$ for the other features). When accounting for the differences amongst speakers by status/age, the set of features increases substantially for teachers, and energy slopes at the final word level are added to the features listed above (all displaying higher values for teachers than for students; $p < .01$ for energy slopes; $p < .001$ for pitch slopes, maximum, minimum and standard deviations). Finally, with regard to gender differences, pitch minimum and maximum in the phrase-final word are the most

significantly different features ($p < .001$) for boys vs. girls in both speech tasks: values are always relatively higher for girls, although boys increase pitch maximum and minimum in the prepared presentation.

7. Concluding remarks

The work presented in this paper, as well as the results reported elsewhere, on intonational variation and on fluency/disfluency contrasts (Mata, 2012; Moniz *et al.*, 2012) show that CPE-FACES is a very useful resource for studying prosody in spontaneous speech and spontaneous *vs*. prepared differences among adults and teenagers, and for research on speaking styles in European Portuguese. At present, the oral presentations data set is being explored within the national project COPAS, for an empirical research on intonation-syntax-discourse interactions (see Mata *et al.*, these proceedings), and also for experiments on extending the AuToBI prosodic annotation system from English to European Portuguese (Moniz *et al.*, 2014). In the future, a part of the prosodically annotated data will be available to the research community via Reciprosody (an open access, shared repository for intonation and prosody resources).

Although spontaneous speech, in general, and teenage speech, in particular, is very hard to automatically process, the experiments conducted so far are very promising. We have shown that errors are strongly patronized and we believe that future processing of the corpus will conduct to the production of training material for further speech recognition experiments.

8. Acknowledgments

We are grateful to all the students and teachers whose generous cooperation allowed creating the CPE-FACES corpus. This work was supported in part by national funds through FCT - Fundação para a Ciência e a Tecnologia, under project PTDC/CLE-LIN/120017/2010 – COPAS. Fernando Batista is supported by ISCTE-IUL.

References

Batista, F., Moniz, H., Trancoso, I., Mamede, N., and Mata, A. I. (2012). Extending automatic transcripts in a unified data representation towards a prosodic-based metadata annotation and evaluation. *Journal of Speech Sciences*, vol. 2, n. 2, pp. 113–136.

Duarte, I. (2000). *Língua Portuguesa. Instrumentos de Análise*. Lisbon, Universidade Aberta.

Eklund, R. (2004). *Disfluency in Swedish Human-Human and Human-Machine Travel Booking Dialogues*. PhD thesis, University of Linköpink.

Escudero, D., Aguilar, L., Vanrell, M. del M. 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

Frota, S. (2000). *Prosody and Focus in European Portuguese*. *Phonological Phrasing and Intonation*. New York, Garland Publishing.

Jurafsky, D. and Martin, J. (2009). *Speech and Language Processing*, 2nd ed. Prentice Hall PTR.

Labov, W. (1976). *Sociolinguistique*. Paris: Minuit

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. (2012). Intonational differences between speech tasks in school context. In A. Botinis (ed.), *Proceedings of the 5th ISEL Conference on Experimental Linguistics*, *ExLing 2012*, Athens, Greece, pp. 81-84.

Mata, A. I., Moniz, H., Móia, T., Gonçalves, A., Silva, F., Batista, F., Duarte, I., Oliveira, F., and Falé, I. (2014). Prosodic, syntactic, semantic guidelines, for topic structures across domains and corpora. These proceedings, *LREC 2014*, Reykjavik, Iceland.

Moniz, H. (2006). *Contributo para a caracterização dos mecanismos de (dis)fluência no Português Europeu*. MA Thesis, University of Lisbon.

Moniz, H., Batista, F., Trancoso, I. and Mata, A.I. (2012). Prosodic context-based analysis of disfluencies. In *Proceedings of Interspeech 2012*, Portland, U.S.A.

Moniz, H. (2013). *Processing Disfluencies in European Portuguese*. PhD Thesis, University of Lisbon.

Moniz, H., Mata, A. I., Hirschberg, J., Batista, F., Rosenberg, A., and Trancoso, I. (2014) Extending AuToBI to prominence detection in European Portuguese. *In Speech Prosody 2014*, Dublin, Ireland.

Neto, J., Meinedo, H., Viveiros, M., Cassaca, R., Martins, C. and Caseiro, D. (2008). Broadcast news subtitling system in Portuguese". *In Proceedings of ICASSP '08*, pp. 1561–1564.

Shriberg, E. (1994). *Preliminaries to a Theory of Speech Disfluencies*. PhD Thesis, University of California.

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, vol. 2, pp. 867-870.

Sjölander, K., Beskow, J., Gustafson, J., Lewin, E., Carlson, R. and Granström, B. (1998). Webbased educational tools for speech technology. *In Proceedings of ICSLP98*, Sydney, Australia, pp. 3217-3220.

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.