Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 65 (2014) 20–35

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Speaking style effects in the production of disfluencies

Helena Moniz [a,b,*], Fernando Batista [a,c], Ana Isabel Mata [b], Isabel Trancoso [a,d]

[a] L2F – INESC-ID, Lisboa, Portugal
[b] FLUL/CLUL, Universidade de Lisboa, Portugal
[c] ISCTE-IUL, Instituto Universitário de Lisboa, Portugal
[d] Instituto Superior Técnico, Universidade de Lisboa, Portugal

Received 25 October 2013; received in revised form 1 May 2014; accepted 21 May 2014
Available online 5 June 2014

Abstract

This work explores speaking style effects in the production of disfluencies. University lectures and map-task dialogues are analyzed in order to evaluate if the prosodic strategies used when uttering disfluencies vary across speaking styles. Our results show that the distribution of disfluency types is not arbitrary across lectures and dialogues. Moreover, although there is a statistically significant cross-style strategy of prosodic contrast marking (pitch and energy increases) between the region to repair and the repair of fluency, this strategy is displayed differently depending on the specific speech task. The overall patterns observed in the lectures, with regularities ascribed for speaker and disfluency types, do not hold with the same strength for the dialogues, due to underlying specificities of the communicative purposes. The tempo patterns found for both speech tasks also confirm their distinct behaviour, evidencing the more dynamic tempo characteristics of dialogues. In university lectures, prosodic cues are given to the listener both for the units inside disfluent regions and between these and the adjacent contexts. This suggests a stronger prosodic contrast marking of disfluency–fluency repair when compared to dialogues, as if teachers were monitoring the different regions – the introduction to a disfluency, the disfluency itself and the beginning of the repair – demarcating them in very contrastive ways.

Keywords: Prosody; Disfluencies; Lectures; Dialogues; Speaking styles

1. Introduction

This paper explores speaking style effects in the production of disfluencies in university lectures and map-task dialogues. In both corpora speech is edited on-line. However, they vary in the ways speakers adjust to communicative contexts. Therefore, distributional patterns, speech and articulation rates and prosodic disfluency/fluency repair strategies are targeted in a broader comparison of inter-corpora styles.

The expression *speaking style* is complex to define. The literature (Biber, 1988; Eskénazi, 1993; Blaauw, 1995; Barry, 1995; Biber and Conrad, 2009; Hirschberg, 2000) has documented the role that multiple dimensions of variation play in style changes, contributing to a more comprehensible view of speaking style. For Eskénazi (1993), there are three essential axes of variation: the degree of intelligibility required by the situation, the familiarity between speaker and listener(s), and the social strata of the communicative participants. The effect that the speaker intends to have on the listener is also another dimension to consider, as evidenced by Barry (1995).

Prosodic analysis has been proved to be very informative for differentiating between speaking styles (e.g., Blaauw (1995); Hirschberg (2000)). The role of disfluencies

* Corresponding author at: L2F – INESC-ID, Lisboa, Portugal. Tel.: +351 917868533.
E-mail addresses: helenam@l2f.inesc-id.pt (H. Moniz), fmmb@l2f.inesc-id.pt (F. Batista), aim@fl.ul.pt (A.I. Mata), isabel.trancoso@l2f.inesc-id.pt (I. Trancoso).

http://dx.doi.org/10.1016/j.specom.2014.05.004
0167-6393/

in this discrimination, however, has typically been rather limited: the presence/absence of disfluent events predicts speech as either spontaneous or read. But recent studies are gradually focusing on *varia* (para)linguistic properties of such events. Either *per se* or combined with other features, disfluencies have been shown to characterize social and emotional behaviour (Gravano et al., 2011; Benus et al., 2012; Ranganath et al., 2013; Schuller et al., 2013). Studies are, thus, moving much beyond the classical view of presence/absence of disfluencies for the classification of spontaneous *vs.* read speech, embracing a diverse set of domains (*e.g.,* speed-dating, Supreme Court hearings, etc.).

The aim of this paper is to characterize prosodic parameters and disfluencies' distributions in European Portuguese for a discrimination of speaking styles and for a contribution on cross-language comparisons of prosodic parameters in different domain. Along this paper we will report trends that point out to a central point: one cannot draw generic conclusions about the distributional and prosodic patterns of disfluencies without taking speaking style into account.

The paper is organized as follows: Section 2 presents the related work. The data used in this study is described in Section 3. Section 4 describes the disfluency annotation. The results of the inter-corpora distributional patterns are reported in Section 5. The inter-corpora prosodic analysis of disfluencies is conducted in Section 6. Conclusions and future work are presented in Section 7.

2. Related work

Disfluencies, *e.g.*, filled pauses, prolongations, repetitions, substitutions, deletions, insertions, characterize spontaneous speech and play a major role in speech structuring (Levelt, 1983; Allwood et al., 1990; Swerts, 1998; Clark and Fox Tree, 2002). There are two main perspectives in the literature to describe disfluencies: (i) as speech errors that disrupt *the ideal delivery of speech* or (ii) as fluent linguistic devices used to manage speech. For a survey on these perspectives, *vide* Kowal and O'Connell (2008). Disfluencies may be used for different purposes related to, *e.g.*, speech structuring (Clark and Fox Tree, 2002), introducing new information (Arnold et al., 2003) and producing fluent strategies in second language learning (Rose, 1998). The fluent component of these phenomena is still rather controversial, even though Heike (1981) and Allwood et al. (1990) have already pointed out the benefits of disfluencies for communicative purposes, and their contribution for on-line planning efforts.

Although the word *disfluencies* still exhibits the depreciating connotation linked to error, this term will be used for sake of terminological simplicity and for a contribution for direct comparisons with other studies. For an overview of the historical perspective of the terminological aspects associated with positive/negative connotations of the terms and of the realms of linguistic studies see Erard (2007).

It is commonly recognized that disfluencies have a specific structure: *reparandum*, *interruption point*, *interregnum*, and *repair* of fluency (Levelt, 1989; Nakatani and Hirschberg, 1994; Shriberg, 1994). The *reparandum* is the region to repair. The *interruption point* is the moment when the speaker stops his/her production to correct the linguistic material uttered, ultimately, it is the frontier between disfluent and fluent speech. The*interregnum* is an optional part and it may have silent pauses, filled pauses (uh, um) or explicit editing expressions (I mean, no). The *repair* is the corrected linguistic material.

It is known that each of these regions has idiosyncratic acoustic properties that distinguish them from each other (Hindle, 1983; Levelt and Cutler, 1983; Nakatani and Hirschberg, 1994; Shriberg, 1994, 2001; Liu et al., 2006). There is in fact an edit signal process (Hindle, 1983), meaning that speakers signal an upcoming repair to their listeners. The edit signal is manifested by means of repetition patterns, production of fragments, glottalizations, co-articulatory gestures and voice quality attributes, such as jitter (perturbations in the pitch period) in the *reparanda*. Sequentially, it is also edited by means of significantly different pause durations from fluent boundaries and by specific lexical items in the *interregnum*. Finally, it may be edited via pitch and energy increases in the repair.

The possible connections between the *reparandum* and the repair have been explored from different perspectives in the literature. Since Levelt and Cutler (1983) there is a binary tendency towards the classification of the prosodic properties of (certain) disfluencies as either copying the pitch contour of the *reparandum* or contrasting the onset of fluency in the repair with the *reparandum*, by means of increasing f_0 and energy. The first strategy is classified as a parallelism between the two regions and is mainly related to appropriateness (involving, for instance, repetition and insertion), whereas the second is classified as contrast marking and is productive with error corrections (mostly substitutions). The literature is not consensual about this dichotomy. For Plauché and Shriberg (1999), repetitions *per se* can behave as parallelistic prosodic structures (copying the pitch contour of the *reparandum*) and also have some degree of contrast (a rising pattern in the repetition is related to an emphasis in the new unit), although not the one reported by Levelt and Cutler (1983). For Savova and Bachenko (2003a,b), distinct categories, such as repetitions and substitutions seem to copy the patterns of their counterparts in the *reparandum*. Moreover, for the authors there is only partial support for the contrastive nature of substitutions when this is manifested by a higher pitch range. Cole et al. (2005) sustains the parallelistic nature of both repetitions and error corrections and considers parallelism the most frequent strategy.

The contrast and parallelism strategies may also be regarded from a comprehension perspective (Levelt, 1983, 1989; Levelt and Cutler, 1983). In comprehension tasks, the information available in disfluencies can help listeners compensate for disruptions and delays in spontaneous

utterances (Brennan and Schober, 2001). Cues are not exclusively the presence of a (certain type of) disfluency, but also the linguistic properties of the structured regions of a disfluent event (Hindle, 1983; Nakatani and Hirschberg, 1994; Shriberg, 1994, 1999, 2001), namely the transition to the repair of fluency, which is of crucial importance for the process of understanding a message. However, the literature does not focus on how those cues may vary accordingly to speaking style due to underlying situational contexts and communicative purposes. The current study aims at filling this gap.

3. Corpora

This work will focus on university lectures and dialogues to discriminate speaking style effects in the production of disfluencies. The choice of the corpora was influenced by the availability of large amounts of (highly spontaneous) transcribed data in European Portuguese for these two domains.

The university lectures corpus was collected within the *LECTRA* national project (Trancoso et al., 2008), aiming at the production of multimedia contents for e-learning applications, and also at enabling hearing-impaired students to have access to recorded lectures. The corpus includes seven 1-semester courses: Production of Multimedia Contents, Economic Theory I, Linear Algebra, Introduction to Informatics and Communication Techniques, Object Oriented Programming, Accounting, and Graphical Interfaces. Six courses were recorded in the presence of students, and only one course was recorded in a quiet environment, targeting an internet audience. Most classes are 60–90 min long. All 7 speakers are native Portuguese speakers and only one course was given by a female speaker. The initial set of 21 h orthographically transcribed was recently extended to 32 h (Pellegrini et al., 2012) in the scope of the Multilingual European Technology Alliance project (META-NET). The corpus was divided into 3 different sets: train (78%), development (11%), and test (11%). The sets include portions of each one of the courses and follow a temporal criterion, meaning the first classes of each course were included in the training set, whereas the final ones were integrated into both development and test sets. In the scope of this paper, only the training portion is being analyzed, but this is a first stage towards automatic clustering and classification tasks.

CORAL (Viana et al., 1998; Trancoso et al., 1998; Caseiro et al., 2002) is a corpus of map-task dialogues. One of the participants (giver) has a map with some landmarks and a route drawn between them; the other (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. The names of the landmarks were chosen to allow the study of some connected speech phenomena in European Portuguese (*e.g.*,sequences of plosives formed across word boundaries or sequences of obstruents formed within and across word boundaries). The 32 speakers were divided into 8 quartets and in each quartet organized to take part in 8 dialogues, totaling 64 dialogues. Given the reduced number of speakers, they were chosen to achieve an adequate balance of sexes, but were restricted in terms of age (under-graduate or graduate students) and accent (Lisbon area). Speakers were chosen in pairs who know each other, so that half of the conversations took place between friends and half between people who did not know each other. The corpus has 9 h (46k words) and was divided into train (75% corresponding to quartets 1–6) and test sets (remaining 25%, quartets 7 and 8).

Both corpora were manually annotated with multilayer labels. For a full report on the annotation schema shared by both corpora, *vide* Moniz (2006) and Trancoso et al. (2008). Our in-house speech recognition system (Neto et al., 2008) was used to produce force aligned transcriptions. The reference data was then provided to the aligned transcription using the NIST SCLite tool (http://www.nist.gov/speech).

Table 1 presents the overall characteristics of the training subsets of both corpora, where values from the last 4 rows correspond to averages. The higher alignment error

Table 1
Overall characteristics of the training subsets.

	Lectures			Dialogues		
		%	p/min		%	p/min
Time (h)	24:28			9:41		
Alignment error	1.0			0.2		
Sentence-like units (SU)	10,576		7.2	7187		12.4
Disfluent SUs	3772	35.7	2.6	1817	25.3	3.1
Words outside disfluencies	176,853		120.5	42,034		72.3
Disfluent words	14,357	8.1	9.8	3850	8.4	6.6
Disfluent sequences	7382	7.5	5.0	2257	8.8	3.9
Disfluent sequences	0.70			0.31		
Words between disfluencies (words/bd)	30.63			22.50		
Time/bd (s)	11.31			7.49		
Useful time/bd (s)	7.74			5.98		

in the lectures is due to a high frequency of computer jargon, acronyms, anglicisms, and a variety of fillers, *e.g.*, several linguistic structures in their weak forms. The alignment error of the dialogues is constantly very low, with the exception of a dialogue with an impressive 65%. The reason for this very high rate is related to the speakers being identical twins and, since the synchronization between them is immediate, the follower only needs 8 turns to conclude the dialogue. It is the smallest dialogue of the corpus, full of laughs produced in simultaneous with backchannels (affirmative answers, either assertive grunts, such as "hum", or "very well"), very hard stretches to process automatically, even in a forced alignment mode.

Different measures have been used in the literature to indicate disfluent rates. Table 1 displays percentages, average and per minute (p/min) values in both corpora. The first point to highlight is that both corpora display disfluency percentages in line with human–human interactions reported in the literature 5–10% accordingly to Shriberg (2001). If we interpret results considering exclusively percentages, we would consider that lectures have more disfluent SUs and comparable percentages of disfluent words. However, when considering the per minute rates, lectures and dialogues are quite differentiable. The number of sentence-like units (either fluent or with disfluencies) is higher per minute in dialogues than in lectures, clearly supported on the fact that dialogues have fewer words in both SUs. A possible explanation for this is the fact that dialogue turns are quite more dynamic than lectures. The interactions between interlocutors in a dialogue motivate a faster deliver of information and more frequent feedback mechanisms. This previous characterization of the corpora is the mot for the remaining sections. Along the paper other dimensions of the inter-corpora variation will be explored.

4. Disfluency annotation

4.1. Typology

As in other areas, terminology regarding disfluent events is rather diverse. However, in the last decades, since the influential work of Shriberg (1994), there is a common-ground typology that speech scientists have been using, promoting direct comparisons of the results achieved in different areas. Shriberg's typology encompasses the following set of disfluent categories: filled pauses (schwa-like quality vowel and/or nasal murmur for European Portuguese); repetitions (linguistic material repeated); substitutions (linguistic material replaced); deletions (abandoned linguistic material, correspond to a complete refresh); insertions (linguistic material inserted, usually with repetitions to clarify an idea); editing expressions (overt expressions regarding on-line message editing); word fragments (linguistic material truncated or incompleted); complex sequences (linguistic material comprising distinct disfluent categories); and mispronunciations (linguistic material pronounced in an erroneous way).

With the work of Eklund (2004), an overview of prolongations in Sweden and in other languages is described, observing regularities in the segmental properties of the elongated lexical material, which provides evidence for another category *per se* – prolongations. Two contributions were taken from the mentioned study. Besides the category *prolongation*, this study will also consider the disfluent events index system proposed by Eklund (2004), establishing correlations between the material to be corrected and the correction itself and the order in which the linguistic material is uttered. Segmental prolongations are elongated segmental linguistic material. Procedurally, prolongations can be measured and compared with linguistic material in other locations. In EP, prolongations in the sense of management of speech are often related with specific lexical items, *e.g.*, functional words with elongated vowels in a context where we would expect reduction or elision of those vowels. In our previous studies we have found that we may also have lexical words elongated with two effects: prolongation affecting more than the last syllable of the word and final lengthening corresponding to an interval of more than 1 s.

4.2. Disfluency tier

The annotation of disfluencies is provided in a separate tier (annotation file), closely following Shriberg (1994) and basically using the same set of labels. This annotation schema is based on Levelt's model (1983), and has been successfully used, *e.g.*, to train methods for the identification and automatic removal of disfluencies, in order to produce clean readable texts. It appears to be also the most adequate from a point of view of linguistic research. In spite of some divergences, it is widely accepted that disfluencies have an internal structure and three different regions need to be considered in their analysis: (i) the *reparandum*; (ii) the *interregnum*; and (iii) the *repair* itself. The *reparandum* is right delimited by an interruption point, marking the moment in time in which an interruption is visible in the surface form. Following a suggestion of Eklund (2004), disfluent items are indexed, as shown in the following example:

```
<tem um número tem um número,  não.  > tem um elemento.
  rl  r2  sl.   rl  r2  sl   el    rl  r2  sl
(<  it has a number it has a number, no. > It has an element.)
```

Table 2
Labels used in the disfluency tier.

Labels	Description	Examples	Annotation
⟨⟩	Auto-corrected	Sequences of disfluencies	⟨...⟩
.	Interruption point	Moment when the speaker interrupts to repair his/her speech	⟨**n. n**⟩
f	Filled pauses	*ou pode estar* ⟨***%aa***⟩ *trancada* (or it can be ⟨% uh⟩ closed)	⟨f.⟩
lm	Segmental prolongations	***de*** = (of=) pronounced as [d**i:**]	⟨lm1.⟩
r	Repetitions	e ⟨***vocês sabem***⟩vocês sabem que (and ⟨you know⟩ you know that)	⟨r1 r2.r1 r2⟩
s	Substitutions	*são* ⟨***os***⟩ *o conjunto dos* ~*X*, ~*Y* (they are ⟨the⟩the set ~X, ~Y)	⟨s1.s1⟩
d	Deletions	vai haver uma série de resultados, ⟨***vamos chamar***⟩ portanto, nós tínhamos a noção de ~R there will be a series of results, ⟨let's call⟩ therefore, we had the notion of ~R	⟨d1 d2.⟩
i	Insertions	⟨***em + que*** **é** ***que***⟩ **em que** ***medida*** **é que** o padrão é útil? in what way is the pattern useful?	⟨r1 r2. r1 i1 r2⟩
e	Editing expressions	⟨*parou* ***quer dizer*** *%aa*⟩ acabou o tempo ⟨stopped I mean⟩ time ran out	⟨s1 e1 e2 f. s1⟩
–	Word fragments	⟨***comp-***⟩ *complementar* (⟨addi-⟩ additional)	⟨r1-.r1⟩
,	Mispronunciations	pode-nos ⟨**servir**~⟩ servir (can ⟨serve~⟩ serve us) pronounced as [ʃɨr′nir] instead of [sɨr′vir]	⟨r1~.r1⟩

Such a solution appears to be less prone to errors than the complex bracketing used by Shriberg, in order to account for the nested structure of long disfluency sequences. Unlike Eklund, however, all items are indexed for a more direct access to eventual changes in word order and to the different strategies that may be used by speakers. The set of labels used in the disfluency tier are shown in Table 2.

5. Inter-corpora disfluency analysis

This section will encompass other dimensions of inter-corpora variation, focusing on disfluency behaviour. First, an overall characterization of fluent sentences and sentences containing disfluencies is given. In this respect, sentences containing disfluencies are further subdivided into fluent and disfluent parts, supported on psycholinguistic studies (e.g., Brennan and Schober (2001)), which show that sentences containing disfluencies are either more complex or associated with more complex tasks. Furthermore, speaker variation in lectures and in dialogues are also described in this section. Finally, inter-corpora disfluency distribution is discussed.

5.1. Overall characterization

In Section 3, we said that dialogues are more dynamic than lectures, since dialogues have more SUs with fewer words and the information is delivered faster. Now we concentrate on other measures contributing to those differences. Table 3 presents the overall characteristics of the

Table 3
Overall characteristics of lectures and dialogues.

		Features	Lectures	Dialogues	*z* (Wilcoxon)
Overall		Words per sentence	18.1	6.4	−55.089
		Disfluent words per sentence	1.4	0.5	−17.885
Fluent SUs		Words	10.0	4.6	−40.473
		Syllables	18.4	8.5	−40.849
		Phones	38.7	17.9	−40.406
		Speech rate	7.8	7.2	−15.377
		Articulation rate	9.2	8.1	−25.942
		Phonation ratio	83.0	89.6	−21.631
		Duration with silences (s)	6.1	1.9	−50.258
		Duration without silences (s)	4.3	1.5	−49.055
Disfluent SUs	Fluent	Fluent words	28.9	9.5	−37.925
		Fluent syllables	56.8	18.1	−37.925
		Fluent phones	120.6	38.6	−37.696
	Disfluent	Disfluent sequences	2.0	1.2	−20.104
		Disfluent words	3.8	2.1	−18.360
		Disfluent syllables	5.4	3.0	−17.044
		Disfluent phones	9.6	5.7	−12.564
		Duration of disfluency	1.3	0.7	−14.439
	Overall	Speech rate	6.0	6.1	Not significant
		Articulation rate	7.3	6.7	−10.412
		Phonation ratio	82.2	90.5	−24.967
		Duration with silences (s)	10.7	3.2	−36.597
		Duration without silences (s)	7.8	2.6	−36.152
		Words between disfluencies	30.6	22.5	−3.008*
		Time between disfluencies	11.8	7.5	−3.598
		Useful time between disfluencies	8.1	6.0	−2.935*

training sets of both corpora for sentence units (SU), either fluent or containing disfluencies. The latter are further subdivided in fluent and disfluent parts. When measuring or assessing fluency, the articulation and speech rates as well as the phonation ratio are of crucial importance. Those measures were calculated based on Grojean (1980) and on Cucchiarini et al. (2002). In the latter the units targeted are phones, whereas in the former they are syllables. In the present study both measures will be given. Thus, *articulation rate* corresponds to the number of phones or syllables divided by the duration of speech without utterance internal silences. *Speech rate* is based on the number of phones or syllables divided by the duration of speech including utterance internal silences. As for the phonation ratio it corresponds to 100% times the duration of speech without utterance internal silences divided by the duration of speech including utterance internal silences. Statistical significance values are displayed for each feature and all features are statistical significant with $p < 0.001$, except those marked with "$*$", corresponding to $p < 0.01$.

A general statement regarding the information displayed in the table is that dialogues, in all the features analyzed, are characterized by the production of fewer words in consequently faster times than lectures. Even though more disfluent SUs are produced per minute in the dialogues (cf. Table 1), this does not slow speakers down, since no significant differences are found in the speech rate of disfluent SUs of both corpora. Thus, an evidence more of the dynamic workflow of a dialogue at this level as well. We interpret the outlined differences as being linked to underlying distinctions between dialogic *vs.* (essentially) monologic communication. In a dialogue, sentences have fewer words and are shorter than the sentences produced by a teacher in expository lectures. The on-the-fly editing process in a map-task dialogue implies a straight cooperative process between two interlocutors under strict temporal constraints, which are totally different from the production circumstances of an university lecture.

It is also evident that for both corpora fluent parts of a sentence-like unit containing disfluencies have more words than fluent SUs, thus supporting the claims that sentences containing disfluencies are more complex (considering number of words as a measure of complexity) than fluent SUs.

With respect to the average number of words uttered per sentence, a comparison can be made with previous studies for European Portuguese. Batista et al. (2012a) reports an average number of 22 and 21 words per sentence in corpora of Portuguese and English broadcast news, respectively. A similar result is reported by Ribeiro and de Matos (2011) and Amaral and Trancoso (2008) for Brazilian Portuguese newspapers (21 words). Moreover, Batista (2011) points out an average number of 29 words in the European Parliament Proceedings Parallel Corpus (Europarl, Koehn (2005)). When analyzing a high-school lecture, Mata (1999) reports an average number of 17 words produced by a teacher within an intonational utterance. As for the study of child–adult dialogues, Mata and Santos (2010) present an average number of 3 words per sentence in the questions made by adults to young children. Comparing the results just described, the latest are the ones closer to the averages of both lectures and dialogues analyzed, as represented in Table 4.

Table 4
Average number of words per sentence in distinct corpora.

Corpora	#Words
Child-directed speech	3
Map-task dialogues	6
High-school lecture	17
University lectures	18
Portuguese broadcast	22
European parliament	29

Table 4 shows a clear distinction between dialogues and the remaining corpora. Dialogues are build upon interactions between interlocutors, as previously mentioned. It is thus comprehensible that fewer words are produced per sentence. Academic presentations are associated with the need to explain in detail several concepts. To do so, teachers often use paraphrases, explicative sentences, examples to illustrate theoretical concepts, etc. The European parliament presentations, as an oratory domain, are mostly related to a clear structured presentation of arguments, being, therefore, the most verbose. What it is also interesting to note is that words uttered between disfluencies in the university lectures (30.6) are closer to the parliament presentations.

5.2. Speaker variation in lectures

Figs. 1–3 show average values per lecture across speakers, where S1–S7 represent the speakers and each bar corresponds to a lecture. Fig. 1 compares useful time (measured in seconds, silences not included) and total time (useful time and silences) between disfluencies. Fig. 2 shows the average number of words uttered between disfluencies. Finally, Fig. 3 shows the total number of fluent/disfluent words. Results show that all measures are subject to speaker and lecture variations. For instance, the average number of words uttered between disfluent events ("/bd") ranges from a maximum of 59.3 words for speaker 3 to a minimum of 12.4 words for speaker 7. Systematically, those two speakers contrast in the average number of words uttered between disfluent events and, as expected, they maintain the same tendencies regarding the time spent speaking and the actual useful time (without silent pauses) used. We performed a Kruskal–Wallis H test to assess significant differences (expressed by H and the degrees of freedom within parentheses). When accounting for all the speakers, there are significant differences with $p < 0.001$ regarding all measures: words/bd ($H(6) = 26.783$), time/bd ($H(6) = 27.463$), and useful time/bd ($H(6) = 28.174$). Speaker 3 is the only female speaker and the most different regarding the useful time and the average number of words

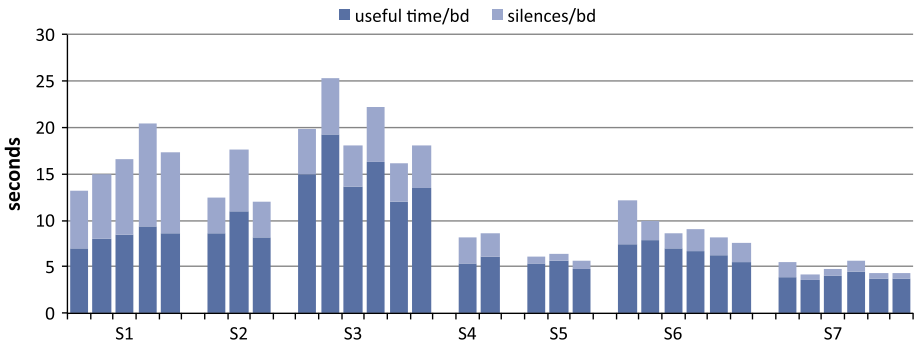


Fig. 1. Total time and useful time (s) between disfluencies (/bd), per lecture and speaker.

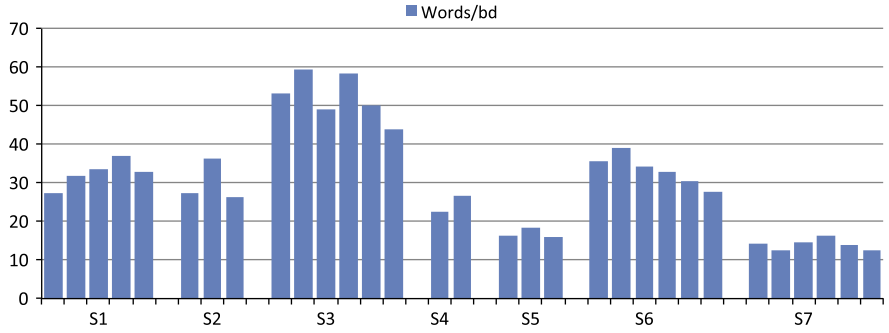


Fig. 2. Average number of words uttered between disfluencies (/bd), per lecture and speaker.

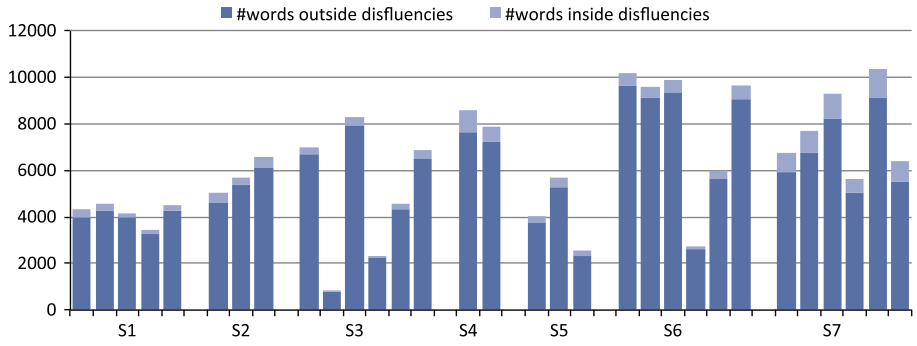


Fig. 3. Total words and disfluent words, per lecture and speaker.

between disfluencies. Even when speaker 3 is removed from the grouping variable, those differences still stand: words/bd ($H(5) = 19.413$), time/bd ($H(5) = 21.871$), and useful time/bd ($H(5) = 21.536$). However, when analyzing exclusively speakers 1, 2 and 4 there are no significant differences in all the measures: words/bd ($H(2) = 4.331, p = 0.115$), time/bd ($H(2) = 5.076, p = 0.079$), and useful time/bd ($H(2) = 5.079, p = 0.079$). The same applies to speakers 5 and 7 regarding words produced between disfluencies ($H(1) = 3.267, p = 0.071$).

Regarding the distribution of disfluencies per sentence, the analysis accounted for several variables measured per sentence and per speaker: average number of words, syllables and phones within (dis)fluent sentences and also within disfluent sequences; duration of (dis)sentences and of disfluent sequences with and without internal silences. Statistical analysis shows that speaker variation is once more reflected at the sentence level, not only at the lecture *per se*. Thus, results show significant differences with $p < 0.001$ in all the measures analyzed. Speaker 5 presents the highest values for the majority of features, whereas speaker 6 presents the lowest. The patterns of both speakers are in line with their performances in class, meaning that speaker 5 is teaching for an internet audience, whereas speaker 6 is often in dialogue with his students. As for disfluent words and disfluency duration, again speaker 6 exhibits the lowest values and speaker 7 the highest disfluency duration and shares with speaker 4 the highest average of disfluent words.

5.3. Speaker variation in dialogues

Figs. 4–6 show average values per dialogue and across speakers, namely the total and useful time between

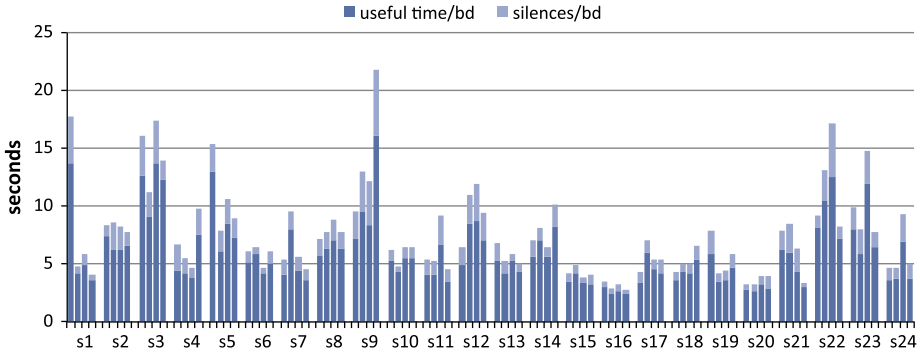


Fig. 4. Total time and useful time between disfluencies, per dialogue and speaker.

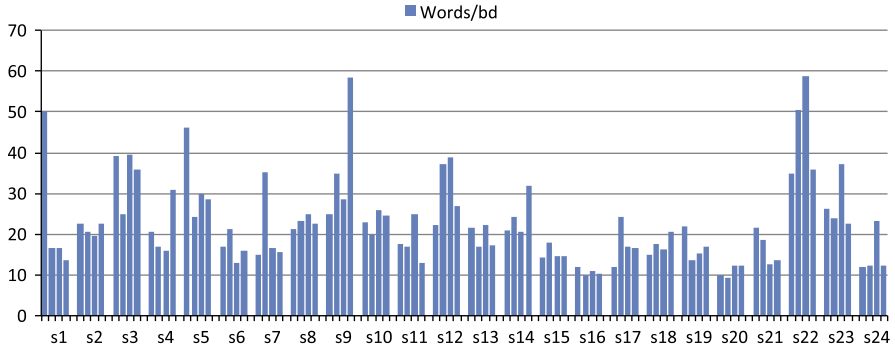


Fig. 5. Average number of words uttered between disfluent sequences, per dialogue and speaker.

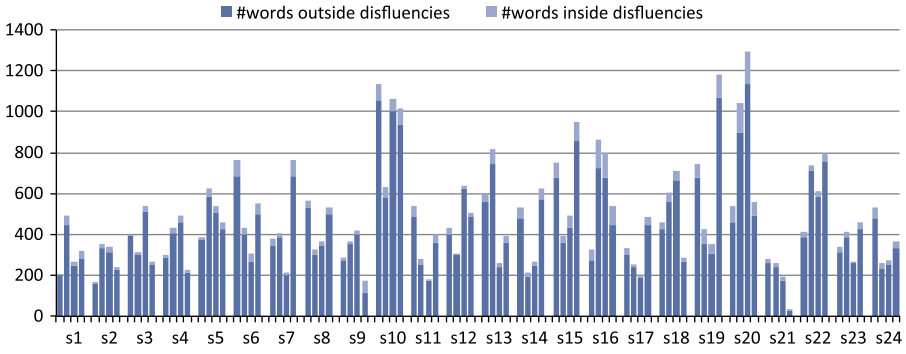


Fig. 6. Total words and disfluent words, per dialogue and speaker.

disfluencies, the average number of words uttered between disfluencies, and the total number of fluent/disfluent words, where S1–S24 represent the speakers and each bar corresponds to a dialogue. Results show that all measures vary per speaker and within speaker per dialogue as well. We performed a Kruskal–Wallis H test to assess significant differences. When accounting for all the speakers, there are significant differences with $p < 0.001$ regarding all measures: words/bd ($H(23) = 68.237$), time/bd ($H(23) = 66.915$), and useful time/bd ($H(23) = 68.472$). Taking into account the number of words uttered between disfluencies, *e.g.*, speaker 22 utters the maximum average of words/bd (58.7), whereas speaker 20 produces the minimum (only 9.3 words). Within speakers, dialogues tend to be quite distinct from each other too (*e.g.*, speakers 1, 5 and 9).

Regarding the distribution of disfluencies per sentence, the main conclusion is that there are fewer sequences and words per sentence than what was observed for the university lectures, ranging from 1 to around 2 sequences and from 1.3 words to a maximum of 3 words. As already pointed out, all measures are comparatively smaller than the ones observed for the university lectures, meaning that the turns are quite small and dynamic. Statistical analysis shows that speaker variation is once more reflected at the sentence level, not only at the dialogue *per se*. Results show significant differences with $p < 0.001$ in all the measures analyzed. Speaker 3 produces more fluent words, syllables

and phones and utters longer fluent sentences, whereas speaker 24 produces more disfluent sequences and words, and also has lengthier disfluent sentences.

5.4. Inter-corpora speaker variation comparison

In the previous sections we have pointed out speaker differences regarding several variables. Beyond speaker differences, there are core characteristics showing distinct inter-corpora properties. As Fig. 7 shows, despite the fact that there is speaker variation, both corpora display significant differences regarding useful time, silent pauses and mean number of words uttered. In dialogues all the measures analyzed are inferior to the ones of the lectures. In the latter there are more differences, for instance, speaker 5 has the highest values and he is the only speaker targeting an internet audience; whereas speaker 6 is the one who most resemble the patterns of the dialogues, being the speaker with more interactions with his audience. Moreover, in lectures there are proportionally more silent pauses than in dialogues, consentaneous with the working flow of a dialogue and with the multifunctionality of silent pauses produced by teachers, *e.g.*, to give-the-floor, to emphasize information, or even to make students think about the topic presented before starting a new one.

Statements on disfluency rates were given in the previous sections. When zooming in and analyzing exclusively SUs containing disfluencies and measuring the percentages of disfluent words over the words produced in the disfluent SU and the duration of the disfluency over the total duration of the disfluent SU, once more the inter-corpora percentages are very distinct, as illustrated in Fig. 8. Dialogues have a higher percentage of disfluencies than lectures. Another contrasting pattern concerns the fact that in dialogues the time spend producing disfluencies is, in the majority of the cases, higher than the percentage of words produced.

Another variable showing inter-corpora differences is the phonation ratio, as displayed in Fig. 9. This variable reinforces the proportions of useful time and silent pauses already described in Fig. 7. The proportion of silent pauses are higher in the lectures than in dialogues, visible in the lower phonation ratios of this domain.

The only characteristic that does not distinguish both corpora is the speech rate (no significant differences were found, as shown in Table 3). Fig. 10 shows that, despite some speaker differences, in both corpora 6 syllables in SUs containing disfluencies are produced per second.

5.5. Disfluency types distribution

Regarding the distribution of disfluent categories, as illustrated in Table 5, *filled* pauses *are* the most frequent type in both corpora, as well as the most frequent type

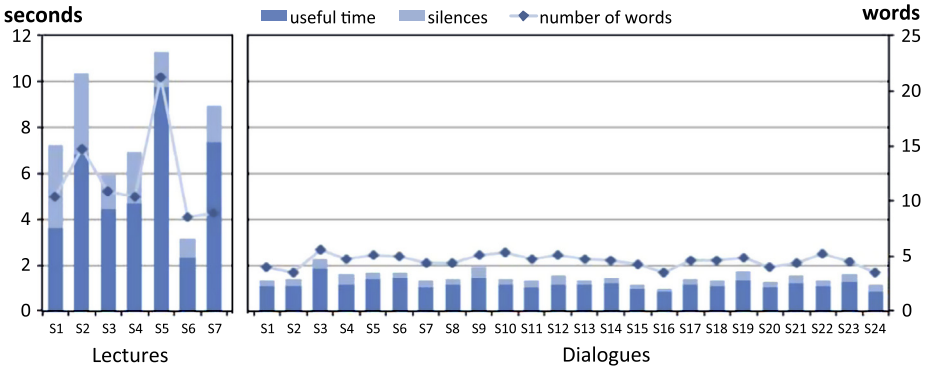


Fig. 7. Comparing fluent SUs.

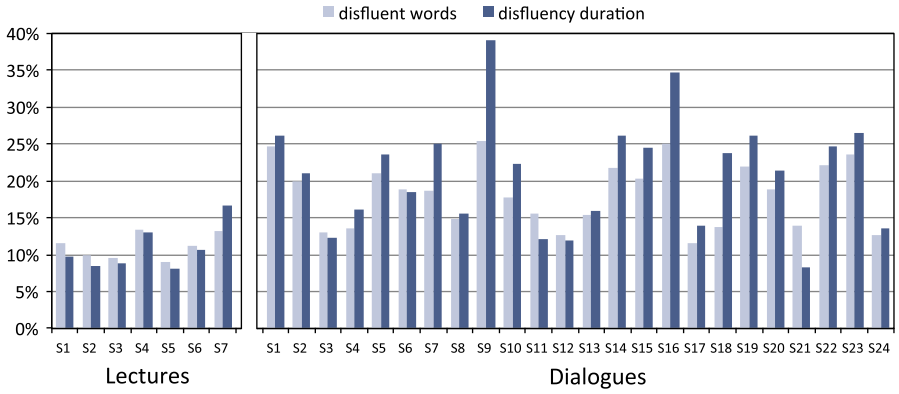


Fig. 8. Comparing disfluency percentages in disfluent SUs.

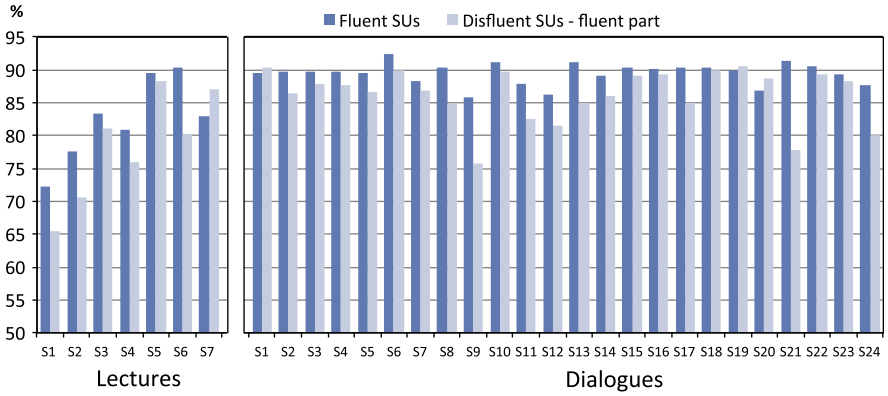


Fig. 9. Comparing phonation ratio between speakers and corpora.

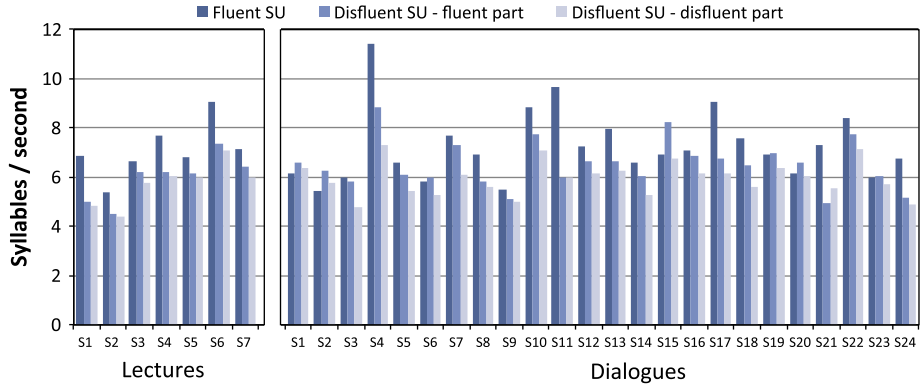


Fig. 10. Comparing speech rate, based on syllables, between speakers and corpora.

reported in the literature (*e.g.*, Shriberg (1994) and Eklund (2004)).*Complex* sequences and *repetitions* are also very frequent in both corpora. However, while lectures display a higher percentage of *complex* sequences (29.1%) than *repetitions* (16%), in dialogues both categories have a similar distribution. Additional differences in the distribution of categories are: dialogues show twice as much *fragments* as lectures and fewer *deletions*. The higher frequency of fragments in dialogues is an evidence more of the strict time constraints of this domain, since speakers interrupt themselves as soon as they notice an error, not preserving the integrity of the word (Levelt, 1989). Speakers rarely choose a *deletion*, since deletions are more complex to process Fox-Tree (1995). A plausible explanation for the above mentioned strategies may be linked to the fact that, unlike dialogue participants, teachers have more time to edit their speech, displaying strategies associated with more careful word choice and careful speech planning.

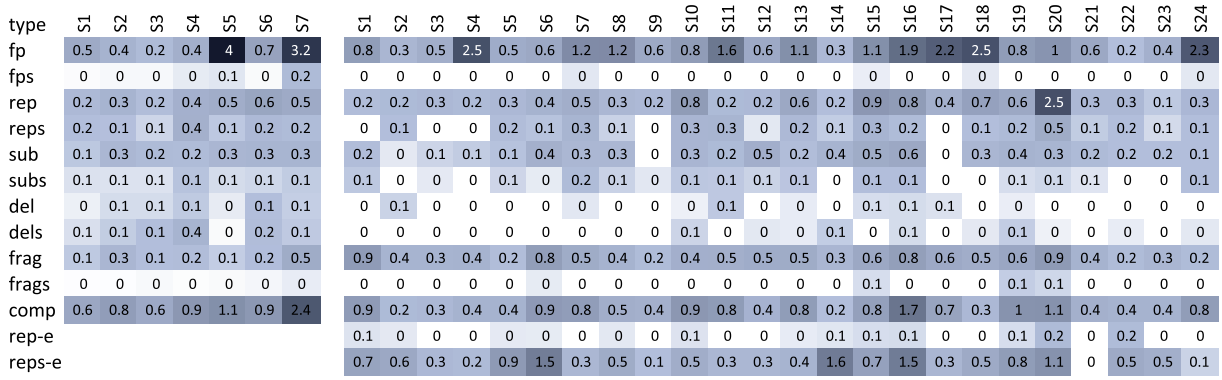
Table 5
Distribution of disfluencies per corpora.

Type	Lectures (%)	Dialogues (%)
Complex	29.1	20.2
Deletions	6.1	1.6
Filled pauses	33.1	31.4
Fragments	6.6	14.9
Repetitions	16.0	22.0
Substitutions	9.1	9.9
Total	100	100

Fig. 11 shows the distribution of disfluency types per speaker, both in lectures and dialogues. Each line represents a different type of disfluency, namely: Complex (comp), Deletion (del), Deletions (dels), Filled pause (fp), Filled pauses (fps), Fragment (frag), Fragments (frags), Repetition (rep), Repetitions (reps), Substitution (sub), Substitutions (subs), emphatic repetition (rep-e), and emphatic repetitions (reps-e).

Concerning lectures, the percentage of disfluencies is distributed almost equally between speakers 1–5, around 8–9%. Although speaker 5 only speaks for 1:37 h and he is the only teacher targeting an internet audience, the production of disfluencies is relatively the same as speakers 1–4, due to a high percentage of *filled pauses*. As for speakers 6 and 7, they spend equivalent speaking time, however the latter utters 40% of all the disfluencies in the corpus, mostly *filled pauses* and *complex* sequences of disfluencies, whereas the former produces 18% and with a more balanced distribution by disfluency type.

In what concerns to dialogues, the percentages range from a minimum of 1.5%, for speaker 9, to a maximum of 12.6%, for speaker 20. The percentage of disfluencies is



type	S1	S2	S3	S4	S5	S6	S7
fp	0.5	0.4	0.2	0.4	4	0.7	3.2
fps	0	0	0	0	0.1	0	0.2
rep	0.2	0.3	0.2	0.4	0.5	0.6	0.5
reps	0.2	0.1	0.1	0.4	0.1	0.2	0.2
sub	0.1	0.3	0.2	0.2	0.3	0.3	0.3
subs	0.1	0.1	0.1	0.1	0.1	0.1	0.1
del	0	0.1	0.1	0.1	0	0.1	0.1
dels	0.1	0.1	0.1	0.4	0	0.2	0.1
frag	0.1	0.3	0.1	0.2	0.1	0.2	0.5
frags	0	0	0	0	0	0	0
comp	0.6	0.8	0.6	0.9	1.1	0.9	2.4
rep-e							
reps-e							

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20	S21	S22	S23	S24
fp	0.8	0.3	0.5	2.5	0.5	0.6	1.2	1.2	0.6	0.8	1.6	0.6	1.1	0.3	1.1	1.9	2.2	2.5	0.8	1	0.6	0.2	0.4	2.3
fps	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
rep	0.2	0.2	0.3	0.2	0.3	0.4	0.5	0.3	0.2	0.8	0.2	0.2	0.6	0.2	0.9	0.8	0.4	0.7	0.6	2.5	0.3	0.3	0.1	0.3
reps	0	0.1	0	0	0.2	0.1	0.3	0.1	0	0.3	0.3	0	0.2	0.1	0.3	0.2	0	0.1	0.2	0.5	0.1	0.2	0.1	0.1
sub	0.2	0	0.1	0.1	0.1	0.4	0.3	0.3	0	0.3	0.2	0.5	0.2	0.4	0.5	0.6	0	0.3	0.4	0.3	0.2	0.2	0.2	0.1
subs	0.1	0	0	0	0.1	0	0.2	0.1	0	0.1	0.1	0.1	0.1	0	0.1	0.1	0	0	0.1	0.1	0.1	0	0	0.1
del	0	0.1	0	0	0	0	0	0	0	0	0.1	0	0	0	0.1	0.1	0.1	0	0	0	0	0	0	0
dels	0	0	0	0	0	0	0	0	0	0.1	0	0	0	0.1	0	0.1	0	0	0.1	0	0	0	0	0
frag	0.9	0.4	0.3	0.4	0.2	0.8	0.5	0.4	0.2	0.4	0.5	0.5	0.5	0.3	0.6	0.8	0.6	0.5	0.6	0.9	0.4	0.2	0.3	0.2
frags	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0	0	0	0.1	0.1	0	0	0	0
comp	0.9	0.2	0.3	0.4	0.4	0.9	0.8	0.5	0.4	0.9	0.8	0.4	0.8	0.2	0.8	1.7	0.7	0.3	1	1.1	0.4	0.4	0.4	0.8
rep-e	0.1	0	0	0	0	0	0	0	0	0.1	0	0	0	0.1	0.1	0.1	0	0	0.1	0.2	0	0.2	0	0
reps-e	0.7	0.6	0.3	0.2	0.9	1.5	0.3	0.5	0.1	0.5	0.3	0.3	0.4	1.6	0.7	1.5	0.3	0.5	0.8	1.1	0	0.5	0.5	0.1

Fig. 11. Disfluency type frequency per minute and per speaker.

again in line with the ones reported in Shriberg (2001), with the exception of speakers 10, 15, 16, 19, and 20 (totaling 20.8% of the speakers). As for the most frequent disfluency types, *filled pauses* are the most frequent types, followed by *complex* sequences (16.8%).

In dialogues, *emphatic repetitions* account for 16.7% of all disfluencies (summing single/multiple *emphatic repetitions*). Emphatic repetitions comprise several structures, being the most productive: (i) affirmative or negative back-channels ("sim, sim, sim."/yes, yes, yes. "não, não, não"/no, no, No.) and (ii) repetition of a syntactic phrase, such as a locative prepositional phrase ("para cima, para cima"/above, above.), used for precise tuning with the follower and for stressing the most important part of the instruction. *Emphatic repetitions* in the map-task are a cue repertoire mainly used to help the follower reaching a location. Thus, for the annotation of the map-task corpus the label emphatic repetition ("rp-e" or rps-e) had to be added, since the occurrence of such events was very notorious and productive, even in the pilot dialogue. Needless to say that *emphatic repetitions* are not disfluencies, either emphatic mechanisms used to highlight information. However, there was an obvious need to include *emphatic repetitions*, to see how in a *continuum*of fluency they would be produced when compared to other events. Since we said in the related work that disfluencies have a fluent component when they are uttered with specific prosodic properties, setting *emphatic repetitions* as an example of fluent mechanisms, selected for making information structure more salient, is a way to check if they have a comparable behaviour with the one of disfluent events.

6. Inter-corpora prosodic analysis

Along this work, we have been describing properties of SUs with and without disfluencies, showing inter-corpora and cross-speaker variation. This section focuses on the characterization of prosodic properties of disfluencies and of their adjacent contexts, in order to verify if the inter-corpora differences are also displayed at the prosodic level as well.

6.1. Feature extraction

Pitch (f_0) and energy (E) are two important sources of prosodic information that can be extracted directly from the speech signal. In our study, the Snack Sound Toolkit (Sjölander et al., 1998) has been used for this purpose, with the default parameters taken from the Wavesurfer tool configuration. Another important information source is the set of durations and confidence scores of phones, words, and interword-pauses, which can be extracted from the recognizer output. Features were calculated for the disfluent sequence itself and also for the two contiguous words, before and after the disfluent sequence. For a complete overview of the prosodic processing, *vide* Batista et al. (2012b). The following set of features has been used for each word in those regions: f_0 and energy raw and normalized mean, median, maxima, minima, differences between units, and standard deviation, as well as POS, number of phones, and durations. Energy and f_0 slopes within the words were calculated based on linear regression. Energy and pitch shapes were also considered, for example, $pslopes : FR_{cw,fw}$ is a shape feature that refers to the pitch slope in the current (cw) and following words (fw), which is **F**alling in *cw* and is **R**ising in *fw*.

6.2. Overall prosodic analysis

In Section 2, two main prosodic strategies from disfluency to fluency repair were discussed – prosodic contrast marking and parallelism. Taking into account that inter-corpora comparisons showed significant differences in almost all the features analyzed, we could also hypothesized that lectures and dialogues could have distinct prosodic strategies regarding disfluency/fluency repairs. Figs. 12–15 show pitch and energy patterns per speaker and disfluency type. As these figures illustrate, in the lectures, pitch and energy increase from the disfluency to the repair region, independently of the speaker and for the majority of the disfluent types (with the exception of sequences of more than a single repetition or deletion). Pitch and energy shapes are, thus, represented by increases in the following word and (mostly) a plateau contour on the preceding word ($pslopes : PR_{cw,fw}$). In the dialogues,

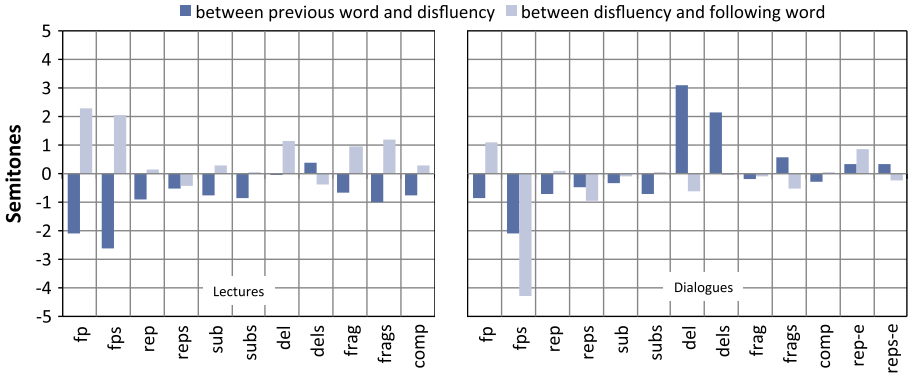


Fig. 12. Pitch differences between units based on the average per type.

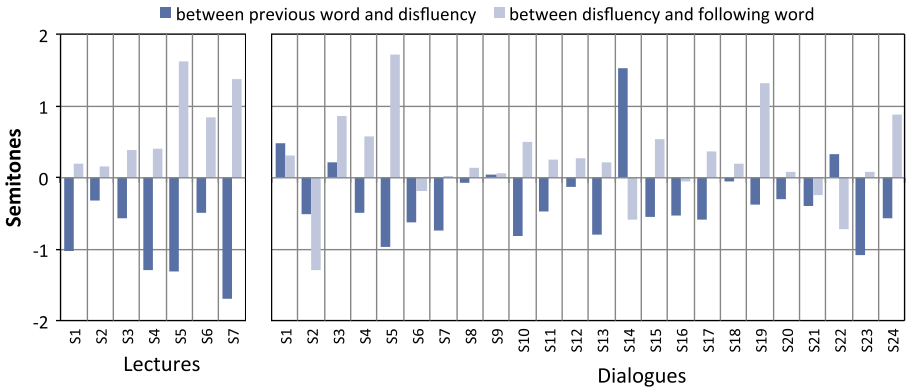


Fig. 13. Pitch differences between units based on the average per speaker.

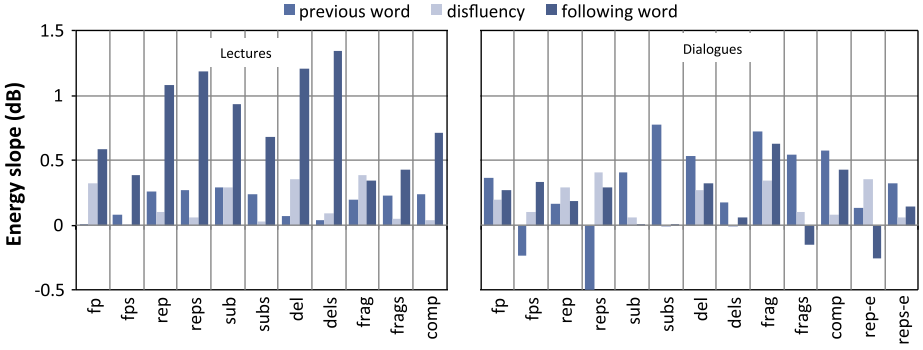


Fig. 14. Energy slopes inside the previous word, disfluency, and in the following word per type.

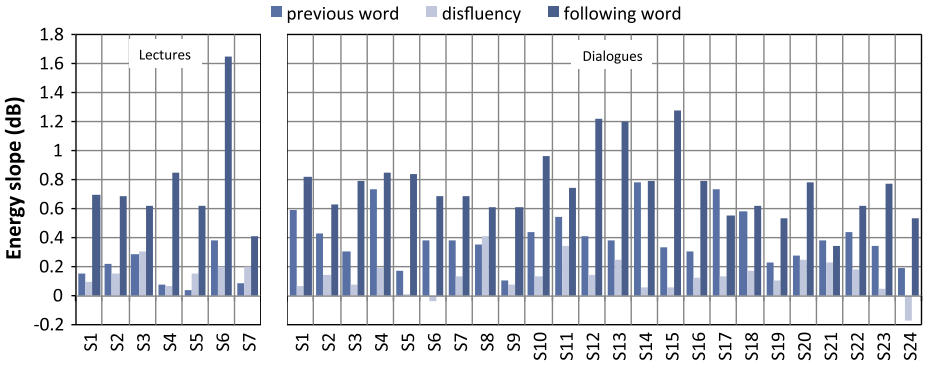


Fig. 15. Energy slopes inside the previous word, disfluency, and in the following word per speaker.

71% of speakers produce the pitch increases, and half of the categories are uttered with subsequent pitch resets. Energy increasing patterns are constant per speaker, however they also vary per disfluency type, *i.e.*, *deletions* and *fragments* do not exhibit an energy gain from the disfluency to the repair. What is interesting to observe is that the disfluent categories with no energy increases have in fact a very striking difference between the disfluency and the repair, meaning, there may not be a gain from the disfluency to the repair, but the differences between those regions are very clear.

Lectures display significant differences at $p < 0.001$ in all units of analysis, even in the context previous to a disfluency; whereas in dialogues such cues for the context previous to a disfluency are not significantly different. Although both corpora display pitch and energy increases, inter-corpora significant differences are found ($p < 0.05$ for energy slopes inside disfluencies and $p < 0.001$ in all the remaining features) in pitch and energy; the only feature with no significant differences is pitch slope inside disfluencies ($p = .171$). The differences are due to the fact that lectures present higher pitch maxima values than dialogues, around a semitone more for both disfluency adjacent contexts. As for energy, dialogues display higher energy maxima values, around 2 dB more in both disfluency adjacent contexts and also within the disfluency region itself.

Inter-corpora prosodic contrast marking strategy of disfluency–fluency repair does not fully agree with the one established by Levelt and Cutler (1983), since there is a cross-speaking style strategy displayed by the majority of the disfluency types and not only by error correction categories, such as substitutions. However, the patterns observed in the lectures, ascribed either for speakers or for disfluency types, do not hold with the same regularity for the dialogues. In university lectures, prosodic cues are being given to the listener both for the units inside disfluent regions, and between these and the adjacent contexts, pointing out to a stronger prosodic contrast marking of disfluency–fluency repair when compared to dialogues.

Pause duration can also be considered as a cue to signal prosodic contrast (Vaissière, 2005). For both corpora, there are 21.2% of disfluent sequences without a previous silent pause (9.6% for dialogues and 11.7 for lectures) and only 3.9% without a subsequent silent pause (3.4% for dialogues and 0.6% for lectures). Fig. 16 illustrates the durations of disfluencies, previous and following lexical contexts, and silent pauses. The disfluency is the longest event, the silent pause between the disfluency and the following word is longer than the previous silent pause, and the *disf+1* word is shorter than the *disf-1* word. Thus, a similar general trend was observed in both corpora. However, two specific properties are crucially different in dialogues: the duration of the silent pause before a disfluency is shorter than *disf+1*, whereas in the lectures they are practically equal; all the averages in the dialogues are shorter than the ones reported for the lectures. Inter-corpora comparisons, conducted with Mann–Whitney–Wilcoxon U test, show that there are significant differences in *disf-1* ($U = -17.099$, $p < 0.001$), previous silent pause ($U = -11.034, p < 0.001$) and *disf* ($U = -2.616, p < 0.01$); whereas no significant difference was found for silent pause after ($U = -.627, p = 0.530$) and *disf+1* ($U = -.792, p = 0.428$). The tempo characteristics of the disfluency and adjacent contexts are an evidence more of the dynamic nature of dialogues.

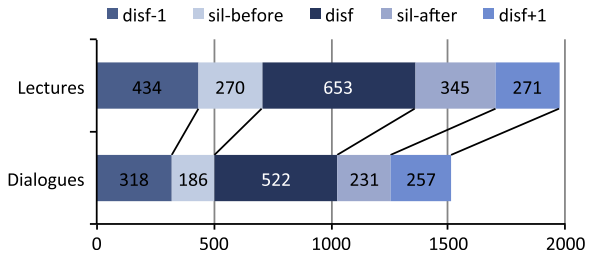


Fig. 16. Duration of the disfluency (in ms), of the adjacent words and silent pauses.

6.3. Speaker and disfluency type variation in lectures

As stated in the previous section, pitch and energy increase from the disfluency to the repair region, independently of the speaker and for the majority of the disfluent types (with the exception of sequences of *repetitions* and of *deletions*). There are, however, degrees in the pitch reset of the next unit. The highest pitch reset is after a *filled pause* or a sequence of *filled pauses* (more than 2 ST), significantly different ($p < 0.001$) from all the other disfluency types. This is, in fact, the disfluency with the subsequent prosodic context that most resemble a full stop. Although *filled pauses* are the events that contribute the most to disfluency/repair pitch increase, even without them pitch and energy resets are still significantly different ($H(9) = 55.130$ with $p < 0.001$; ($H(9) = 178.235$ with $p < 0.001$; respectively). We know that for EP (Moniz, 2006), as for other languages, *filled pauses* tend to occur mainly at major intonational boundaries, therefore pitch and energy resets in the subsequent units are not that surprising. The second highest pitch reset occurs after a single *deletion*. Again, these findings are related to the fact that the unit after a *deletion*, as refreshed linguistic material, is more prone to exhibit an f_0 reset, which is an expected property at the beginning of a major intonational unit.

As for energy,*deletions* and *repetitions* are significantly different ($p < 0.001$) from all the remaining types, with the highest energy slope within the repair. It is worth noting that energy increases from disfluency to the repair with sequences of *repetitions* and of *deletions* are not significantly different from each other ($U = 38630.0$, with $p = 0.062$). Even without *repetitions* and *deletions*, again pitch and energy resets are still significantly different ($H(7) = 629.876$; $H(7) = 262.442$; respectively).

Tempo patterns exhibit significant differences $p < 0.001$ per speaker and disfluency type in the units "disf-1"; "silent

pause before", "disf", "silent pause after", and "dif+1" ($(H(6) = 514.752)$, $(H(6) = 286.032)$, $(H(6) = 334.792)$, $(H(6) = 883.652)$, and $(H(6) = 511.590)$; $(H(11) = 880.179)$,$(H(11)=874.084)$,$(H(11)=2510.487)$, $(H(11)= 243.516)$, and $(H(11)=949.304)$; respectively). Sequences of more than one event are lengthier than single events. The longest disfluency is a*complex* sequence of disfluencies and the smallest a *fragment*. Furthermore, there is a general tendency to produce lengthy silent pauses after a disfluency. However, there is a striking different pattern concerning the production of *filled pauses*, *i.e.*, the previous silent pause is longer (423 ms) than the one after (262 ms). When two or more *filled pauses* occur the adjacent silent pauses are exactly the same (173 ms).

Based on the prosodic parameters analyzed, one may conclude that speakers exhibit different degrees in mastering all the features. Thus, the acoustic correlates of the most proficient speaker (S6) are expressed by means of: (i) the highest energy slope within the repair; (ii) a considerable pitch increase also in the repair; (iii) the smallest disfluency duration; and (iv) the highest articulation and speech rates. The fact that S6 has the smallest duration of speech with and without internal silences is mainly related to the rich dynamics of the interactions with the class. Despite being a theoretical course, the time spent in asking the students to discuss concepts and to give examples of those is substantial. It is interesting to note that, when asked to classify the speakers regarding "likeability", our three annotators were unanimous in stating that speaker 6 is the most "likeable" one. The prosodic correlates of this naive classification may be linked to several distinct features, namely, the highest energy slope within the repair, and also a considerable pitch increase, correlates which have been frequently associated with fluency and with higher level strategies of language use.

6.4. Speaker and disfluency type variation in dialogues

There are two distinct patterns regarding pitch and energy increases from the disfluency to the repair region. Pitch increases are strongly dependent on the disfluency type and on the speaker, whereas energy ones do not vary per speaker, only per disfluency type. 20 speakers (71% of the speakers) produce pitch increases from disfluency/fluency repair and half of the categories are uttered with subsequent pitch increases. The energy increases are constant per speaker, however they also vary per disfluency type, *i.e.*, *deletions* and *fragments* do not exhibit an energy gain from the disfluency to the repair.

The highest pitch reset is after a single *filled pause*, again similar to the university lecture corpus, and an *emphatic repetition*. However, contrarily to the university lectures, in the map-task corpus the pitch reset does not encompasses contiguous sequences of *filled pauses* and the reset is not as striking as in the university lectures. A single *deletion* exhibits a high pitch and energy slope inside the disf+1 word.

Tempo patterns are significant different $p < 0.001$ per speaker and disfluency type in the units "disf-1", "silent pause before", "disf", "silent pause after", and "dif+1" ($(H(23) = 79.005)$, $(H(23) = 74.878)$, $(H(23) = 69.465)$, $(H(23) = 161.392)$, and $(H(23) = 44.599, p < 0.01)$, $(H(12) = 67.384)$, $(H(12) = 171.633)$, $(H(12) = 944.476)$, $(H(12) = 110.527)$, and $(H(12) = 222.460)$; respectively).

As expected, sequences of more than one event are lengthier than single events. The longest disfluency is a sequence of *filled pauses* (for the university corpus it was a *complex* sequence of disfluencies) and the smallest a *fragment*. *Complex* sequences and *filled pauses* as well as *emphatic repetitions* are uttered with previous silent pauses longer than the subsequent ones. Furthermore, *emphatic repetitions* have in fact the biggest comparable difference (almost 100 ms) between the adjacent silent pauses. The patterns observed for silent pauses regarding *emphatic repetitions*, either single or in sequences, distinguish them from repetitions *per se*.

7. Conclusions

Speaking style effects in the production of disfluencies in both corpora are confirmed based on distributional patterns and prosodic properties. Distributional patterns evidenced that the selection of disfluency type is corpus dependent. Excluding filled pauses, the remaining disfluency categories have different distributional patterns. In dialogues, speakers produce more often repetitions and fragments than in lectures. In lectures, teachers prefer complex sequences of disfluencies (mostly repetitions and substitutions used for lexical search). Those strategies were associated with teachers having more time to edit their speech, displaying strategies associated with more careful word choice and speech planning, whereas dialogue participants had stricter time constraints.

Regarding prosodic parameters, although there is a cross corpora prosodic contrast marking between disfluency/fluency repair, there are significant differences in the degrees of contrast made in both corpora. In lectures, prosodic cues are given for the disfluency and the adjacent contexts for all the speakers and for the majority of the disfluency types; whereas in dialogues fewer cues are given. Another striking difference is the fact that lectures exhibit the highest pitch maxima in all units of analysis, whereas dialogues exhibit the highest energy maxima. As for tempo patterns, disfluencies, adjacent contexts and silent pauses are shorter in dialogues than in lectures, evidencing once more the dynamic nature of dialogues.

Focusing on disfluency sequences, prosodic parameters previously explored for the discrimination of speaking styles were analyzed. Thus, speech and articulation rates, pause duration, pitch and energy mean, minima and maxima were studied for sentence-like units with disfluencies and, for comparison sake, also for sentences without disfluencies. Results show that lectures display longer pause durations, higher speech and articulation rates, and higher

pitch maxima values than dialogues. However, when considering only the speech rate in disfluent sequences, no inter-corpora significant differences were found.

Although there are stylistic effects in the production of disfluencies, there is also a considerable range of speaker variation within each speaking style. For instance, the duration of fluent SUs shows clear differences across speaking style regardless of speaker differences. However, for speaking rate, the difference is not pronounced. We intend to conduct further work distinguishing the influence of speaker idiosyncrasies and speaking style. Ideally, this could be done through recording the same speakers in different situations, which is not so easy with the current corpora, but certainly worth exploring.[1]

Future work will also tackle other corpora in order to encompass distinct domains and verify possible speaking style effects in the production of disfluencies. Another trend we are currently following is the impact of speaking styles in automatic disfluency detection.

Acknowledgments

This work was supported by national funds through FCT – Fundação para a Ciência e Tecnologia, under Ph.D Grant FCT/SFRH/BD/44671/2008 and Post-doc fellow researcher Grant SFRH/BPD/95849/2013, projects PEst-OE/EEI/LA0021/2013 and PTDC/CLE-LIN/120017/2010, by European Project EU-IST FP7 project SpeDial under Contract 611396, and by ISCTE-IUL, Instituto Universitário de Lisboa.

References

Allwood, J., Nivre, J., Ahlsen, E., 1990. Speech management: on the non-written life of speech. Nord. J. Linguist. (13), 3–48.

Amaral, R., Trancoso, I., 2008. Topic segmentation and indexation in a media watch system. In: Interspeech 2008. Brisbane, Australia.

Arnold, J., Fagnano, M., Tanenhaus, M., 2003. Disfluencies signal theee, um, new information. J. Psycholinguist. Res. (32), 25–36.

Barry, W., 1995. Phonetics and phonology of speaking styles. In: ICPhS 1995. Stockholm, Sweden.

Batista, F., 2011. Recovering Capitalization and Punctuation Marks on Speech Transcriptions. Ph.D. thesis, Instituto Superior Técnico.

Batista, F., Moniz, H., Trancoso, I., Mamede, N., 2012a. Bilingual experiments on automatic recovery of capitalization and punctuation of automatic speech transcripts. Trans. Audio Speech Lang. Process. (20), 474–485.

Batista, F., Moniz, H., Trancoso, I., Mamede, N., Mata, A.I., 2012b. Extending automatic transcripts in a unified data representation towards a prosodic-based metadata annotation and evaluation. J. Speech Sci. (3), 115–138.

Benus, S., Levitan, R., Hirschberg, J., 2012. Entrainment in spontaneous speech: the case of filled pauses in Supreme Court hearings. In: 3rd IEEE Conference on Cognitive Infocommunications. Kosice, Slovakia.

Biber, D., 1988. Variation Across Speech and Writing. Cambridge University Press.

Biber, D., Conrad, S., 2009. Register, Genre, and Style. Cambridge Textbooks in Linguistics.

Blaauw, E., 1995. On the Perceptual Classification of Spontaneous and Read Speech. Research Institute for Language and Speech.

Brennan, S., Schober, M., 2001. How listeners compensate for disfluencies in spontaneous speech? J. Mem. Lang. (44), 274–296.

Caseiro, D., Silva, F., Trancoso, I., do Céu Viana, M., 2002. Automatic alignment of map task dialogues using WFSTs. In: PMLA – ISCA Tutorial and Research Workshop. Aspen, Colorado.

Clark, H., Fox Tree, J., 2002. Using uh and um in spontaneous speaking. Cognition (84), 73–111.

Cole, J., Hasegawa-Johnson, J., Shih, C., Kim, H., Lee, E., Lu, H., Mo, Y., Yoon, T., 2005. Prosodic parallelism as a cue to repetition and error correction disfluency. In: DISS 2005. Aix-en-Provence, France.

Cucchiarini, C., Strik, H., Boves, L., 2002. Quantitative assessment of second language learner's fluency: comparisons between read and spontaneous speech. J. Acoust. Soc. Am. 111 (6), 2862–2873.

Eklund, R., 2004. Disfluency in Swedish Human–Human and Human–Machine Travel Booking Dialogues. Ph.D. thesis, University of Linköpink.

Erard, M., 2007. Um – Slips Stumbles and Verbal Blunders and What They Mean. Pantheon Books, New York.

Eskénazi, M., 1993. Trends in speaking styles research. In: Eurospeech 1993. Berlin, Germany.

Fox-Tree, J.E., 1995. The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. J. Mem. Lang. (34), 709–728.

Gravano, A., Levitan, R., Willson, L., Benus, S., Hirschberg, J., Nenkova, A., September 2011. Acoustic and prosodic correlates of social behavior. In: Interspeech 2011. Florence, Italy.

Grojean, F., 1980. Temporal variables within and between languages. In: Dechert, H., Raupach, M. (Eds.), Towards a Cross-linguistic Assessment of Speech Production. Lang, Frankfurt, pp. 144–184.

Heike, A., 1981. A content-processing view of hesitation phenomena. Lang. Speech (24), 147–160.

Hindle, D., 1983. Deterministic parsing of syntactic non-fluencies. In: ACL 1983. Cambridge, Massachusetts, USA, pp. 123–128.

Hirschberg, J., 2000. A corpus-based aproach to the study of speaking styles. In: Bruce, G., Horne, M. (Eds.), Theory and Experiments: Studies Presented to Gösta Bruce. Kluwer Academic Publishers, pp. 335–350.

Koehn, P., 2005. Europarl: a parallel corpus for statistical machine translation. In: 10th Machine Translation Summit 2005. Phuket, Thailand.

Kowal, S., O'Connell, D.C., 2008. Communicating with One Another: Toward a Psychology of Spontaneous Spoken Discourse. In: Cognition and Language: A Series in Psycholinguistics. Springer, New York.

Levelt, W., 1983. Monitoring and self-repair in speech. Cognition (14), 41–104.

Levelt, W., 1989. Speaking. MIT Press, Cambridge, Massachusetts.

Levelt, W., Cutler, A., 1983. Prosodic marking in speech repair. J. Semantics (2), 205–217.

Liu, Y., Chawla, N.V., Harper, M.P., Shriberg, E., Stolcke, A., 2006. A study in machine learning from imbalanced data for sentence boundary detection in speech detection in speech. Comput. Speech Lang. 20 (4), 468–494.

Mata, A.I., 1999. Para o estudo da entoação em fala espontânea e preparada no Português Europeu. Ph.D. thesis, University of Lisbon.

Mata, A.I., Santos, A.L., 2010. On the intonation of confirmation-seeking requests in child-directed speech. In: Speech prosody. Chicago, USA.

Moniz, H., 2006. Contributo para a caracterização dos mecanismos de (dis)fluência no Português Europeu. Master's thesis, University of Lisbon.

Nakatani, C., Hirschberg, J., 1994. A corpus-based study of repair cues in spontaneous speech. J. Acoust. Soc. Am. (JASA) (95), 1603–1616.

Neto, J., Meinedo, H., Viveiros, M., Cassaca, R., Martins, C., Caseiro, D., 31 2008–April 4 2008. Broadcast news subtitling system in Portuguese. In: ICASSP 2008. pp. 1561–1564.

[1] The authors acknowledge this valid suggestion from one of our reviewers.

Pellegrini, T., Moniz, H., Batista, F., Trancoso, I., Astudillo, R., 2012. Extension of the lectra corpus: classroom lecture transcriptions in european portuguese. In: GSCP 2012. Brazil.

Plauché, M., Shriberg, E., 1999. Data-driven subclassification of disfluent repetitions based on prosodic features. In: ICPhS 1999. S. Francisco, USA.

Ranganath, R., Jurafsky, D., McFarland, D.A., 2013. Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates. Comput. Speech Lang. 27 (1), 89–115.

Ribeiro, R., de Matos, D., 2011. Centrality-as-relevance: support sets and similarity as geometric proximity. J. Artif. Intell. Res. (42), 275–308.

Rose, R., 1998. The Communicative Value of Filled Pauses in Spontaneous Speech. Ph.D. thesis, University of Birmingham, UK.

Savova, G., Bachenko, J., 2003a. Designing for errors: similarities and differences of disfluency rates and prosodic characteristics acoss domains. In: Interspeech 2003. Geneva, Switzerland.

Savova, G., Bachenko, J., 2003b. Prosodic features of four types of disfluencies. In: DISS 2003. Göteberg, Sweden.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., Narayanan, S., 2013. Paralinguistics in speech and language – state-of-the-art and the challange. Comput. Speech Lang. 27 (1), 4–139.

Shriberg, E., 1994. Preliminaries to a Theory of Speech Disfluencies. Ph.D. thesis, University of California.

Shriberg, E., 1999. Phonetic consequences of speech disfluency. In: International Congress of Phonetic Sciences. San Francisco, pp. 612–622.

Shriberg, E., 2001. To errrr is human: ecology and acoustics of speech disfluencies. J. Int. Phonet. Assoc. 31, 153–169.

Sjölander, K., Beskow, J., Gustafson, J., Lewin, E., Carlson, R., Granström, B., 1998. Web-based educational tools for speech technology. In: ICSLP 1998. Sydney, Australia, pp. 3217–3220.

Swerts, M., 1998. Filled pauses as markers of discourse structure. J. Pragmatics (30), 485–496.

Trancoso, I., do Céu Viana, M., Duarte, I., Matos, G., 1998. Corpus de diálogo CORAL. In: PROPOR'98. Porto Alegre, Brasil.

Trancoso, I., Martins, R., Moniz, H., Mata, A.I., Viana, M.C., 2008. The Lectra corpus – classroom lecture transcriptions in European Portuguese. In: LREC 2008 – Language Resources and Evaluation Conference. Marrakesh, Morocco.

Vaissière, J., 2005. Perception of intonation. In: Pisoni, D., Remez, R. (Eds.), The Handbook of Speech Perception. Blackwell Publishing, pp. 236–263.

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