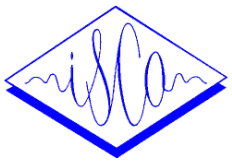


ITRW on Disfluency in
Spontaneous Speech (DiSS'01)
Edinburgh, Scotland, UK
August 29-31, 2001

Annotation and analysis of disfluencies in a spontaneous speech corpus in Spanish

L.J. Rodríguez, I. Torres, A. Varona

Departamento de Electricidad y Electrónica. Facultad de Ciencias. UPV/EHU.
Apartado 644. 48080 Bilbao. SPAIN.
{luisja,manes,amparo}@we.lc.ehu.es

Abstract

A new database consisting of 227 dialogues in Spanish was annotated with disfluencies. Then a detailed analysis of the annotations was carried out. The database had been recorded according to the well known Wizard of Oz paradigm. Seventy-five speakers were given each one three different scenarios to make queries about timetables, prices and other conditions of train travels between two spanish cities. The notion of disfluency was relaxed to include any acoustic, lexical or syntactic feature that distinguises spontaneous from read speech. A specific XML annotation scheme was developed. A simple text editor was used to insert marks, and a specific parser was implemented to find errors in annotations. The analysis of annotations revealed that disfluencies were not uniformly distributed among either user turns or speakers. Most disfluencies were grouped into certain user turns, especially the first one. On the other hand, some speakers were remarkably more prone to hesitate, repeat or correct fragments of speech than others.

1. Introduction

In the mid nineties large vocabulary continuous speech recognition technology achieved the big goal of translating read speech to text with word error rates of around 10%. This technology is now being used as a core component of broadcast news transcription systems, speech-to-speech translating systems and especially dialogue systems [1, 2, 3]. In this context the great challenge is to deal with spontaneous and somewhat unconstrained speech. This will require the acquisition and detailed annotation of generic and application specific databases for many languages. Also new modeling assumptions should be applied and more powerful algorithms should be developed.

Here we present the first milestone, which is the annotation of acoustic, lexical and syntactic disfluencies for an application specific database in Spanish language, which will serve as benchmark to study and to model this kind of phenomena. Unlike the rapidly growing number of spontaneous speech databases for English [4, 5, 6, 7], no corpus with annotation of disfluencies is available for Spanish, so this work can be considered as a pioneering effort.

The rest of the paper is organized as follows: the main features of our database are shown in Section 2; Section 3 presents the concept of disfluency applied in this work and briefly describes the inventory of speech events classified under such category; Section 4 presents the annotation format defined specifically for this work, and some details about the annotation process; statistics of disfluencies are shown and discussed in Section 5. Finally, Section 6 summarizes the conclusions of this work.

This work was partially supported by the Spanish CICYT, under project TIC98-0423-C06-03.

2. The spontaneous speech database

Our spontaneous speech database –which henceforth we will call *OZ1*– consists of the speech signals and the ortographic transcriptions of 227 Spanish dialogues, recorded at 8 kHz across telephone lines applying the well known *Wizard of Oz* mechanism: a human operator simulated the behaviour of the dialogue system, including recognition and/or understanding errors, so that users could think they were interacting with a real system [8, 9]. It must be said that the so called *users* were in fact 75 recruited volunteers, which were given three scenarios with dates, timetables and other conditions for a travel by train between two spanish cities. Actually, to adequately design the scenarios and to clarify what should be the system capabilities, a preliminary database was recorded with dialogues between real users and RENFE[1] information service operators. This preliminary database had been transcribed to plain text but not used for the adverse recording conditions [10]. Recruited users could get as much information as they wanted from the dialogue system, doing it in a natural manner, just as they would in a real call. However, some users still tended to hyperarticulate or even insert pauses between words, whereas others enlarged their turns with unnecessary explanations and often interrupted the system answers. This resulted in a great variability both in spontaneity and turn durations, these latter ranging from 0.5 to 50 seconds. The database includes 1657 non-empty user turns, lasting about 150 minutes. This gives an average of 7.3 non-empty user turns per dialogue, each one lasting an average of 5.4 seconds.

3. The inventory of disfluencies

We apply a wide definition of disfluency as any acoustic, lexical or syntactic feature that distinguishes spontaneous from read speech. In fact, we should better refer to them as spontaneous speech events. To define the inventory of disfluencies two key requirements were posed: coverage and coherence. Therefore, among all possible disfluencies, only those with enough number of samples in our database, plus some others considered significant, were included in the inventory and annotated. Before exploring the kind and frequency of the disfluencies that appear in *OZ1*, a tentative set was defined covering all the spontaneous speech events we could expect in human-machine communications, leaving aside some others which can be only expected in human-human dialogues. Starting from these considerations, a representative subset of 40 dialogues was used to validate the

[1] RENFE is the spanish public railway transportation system.

Table 1: Inventory of disfluencies, XML marks, attribute values, simplified marks and appearing counts for the database *OZ1*.

Category	XML	*source*/type	Simplified	Counts
Noises	n	*world*/generic	nw	661
		speaker/air	na	1404
		speaker/lips	nl	600
		speaker/cough	nt	9
Lengthenings of sounds	a	-	a	1019
Silence pauses	p	-	p	753
Filled pauses	f	a	fa	93
		e	fe	546
		m	fm	179
		trash	fb	210
Lexical disfluencies	l	unfinished	lu	95
		mispronounced	lm	105
Abandoned sentences	b	-	b	70
Retracings	r	repetition	rr	292
		substitution	rs	141
		insertion	ri	37
		deletion	rd	5
Discourse markers	d	open	do	150
		close	dc	189
		accept	da	78
		reject	dr	45
		explain	de	71
		request	dq	92
		fill	df	225
		exclaim	dx	15

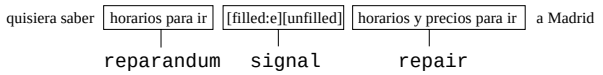
inventory of disfluencies, which evolved from the initial set to the following:

Acoustic disfluencies: this category included noises, lengthenings of sounds, silence –or unfilled– pauses and filled pauses. Noises were included because, though not disfluencies in the strict sense, they seldom appear in read speech, but are pervasive in spontaneous speech. With regard to filled pauses, various acoustic realizations were found in Spanish language, either vowels ('a', 'e') or nasalizations ('m').

Lexical disfluencies: spontaneous speech is far more relaxed than read speech, so a high number of popular or familiar expressions can be found, as well as pronunciation variants –contractions, misarticulations, non-canonical acoustic realizations of phonemes, etc.– due to dialectal or speaker specific features, high speech rates, etc. We defined lexical disfluencies as not properly –or not canonically– pronounced words; for the sake of completeness, cut or unfinished words were also included in this category.

Syntactic disfluencies: among the wide range of them that can be found in spontaneous speech (false starts, repetitions, reformulations, unfinished sentences, sentences completing a previous one, missing words, lacks of concordance, etc) we only considered two categories: *abandoned sentences* (most times false starts) and *retracings*, these latter accounting for repetitions, substitutions and reformulations with insertion or deletion of words.

We applied the structure of retracings shown in [11]: a segment to be repaired –*reparandum*–, a segment marking the correction –*signal*– which may include filled or unfilled pauses and some editing phrases like 'sorry' or 'I mean', and a third segment –*repair*– giving the replacing material, which can be a repetition, a substitution or a more complex reformulation with insertion or deletion of words, as shown in the following example, taken from *OZ1*[2]:



Discourse markers: here we consider very usual words or phrases without any specific meaning but carrying out a metalinguistic function, as opening ('hello', 'good morning'), closing ('thanks', 'good bye', 'that's all'), emphasizing ('please', 'come on'), filling ('well', 'you know'), editing ('sorry', 'I mean'), etc. Although discourse markers cannot be classified as disfluencies, but as pragmatic elements of spoken language, they were annotated to allow the definition of specific categories for them in the language model, which could improve the recognition of spontaneous speech.

4. The annotation scheme

After an exhaustive review of the formats and tools for the annotation of linguistic corpora listed by LDC [12], especially the guidelines given by the european project MATE [13], a specific XML annotation scheme was designed for disfluencies, which –as a first approach– accounted only for disfluencies happening in human-machine communications, and more particularly in *OZ1*, as explained in Section 3. The annotation scheme was accompanied by the corresponding manual [14]. Annotations could refer to instantaneous events, then they were simply inserted in the corresponding place of the ortographic transcription: <mark attribute=value/>, or could refer to a time interval, then affecting some amount of text: <mark attribute=value> TEXT </mark>. Marks were one-letter codes. Some marks needed no attributes, others required one or more attributes. For the database *OZ1* three attributes were defined: *type*, used to give a more detailed description of the disfluencies, *source*, used only for noises, and *word*, used only to supply the canonical version of a word in lexical disfluencies.

Marks were added by hand, using a simple text editor. To make easier such a tedious process, a simplified format was also defined. Each simplified annotation consisted of a short mark, usually two letters encoding both the mark and the value of the attribute *type*, enclosed between parentheses and affecting some text. The XML and the simplified annotations for the example shown above would be:

XML

```
quisiera saber
<r type="insertion">
<m> horarios para ir </m>
<s> <f type="e"/> <p/> </s>
<c> horarios y precios para ir </c>
</r>
a Madrid
```

Simplified

```
quisiera saber (ri (m horarios para ir) (s (fe)(p)) (c horarios y precios para ir)) a Madrid
```

Since machine answers were automatically generated from a predefined set of templates, only user turns were annotated –after careful listenings of the speech signals. To help the detection and correction of annotation errors, a very simple parser was implemented, which accounted not only for the parentheses and marks, but also the correctness of their contents. The parser was iteratively applied to the annotated dialogues, and

[2] translated to English as: *I would like to know timetables to go [filled:e][unfilled] timetables and prices to go to Madrid.*

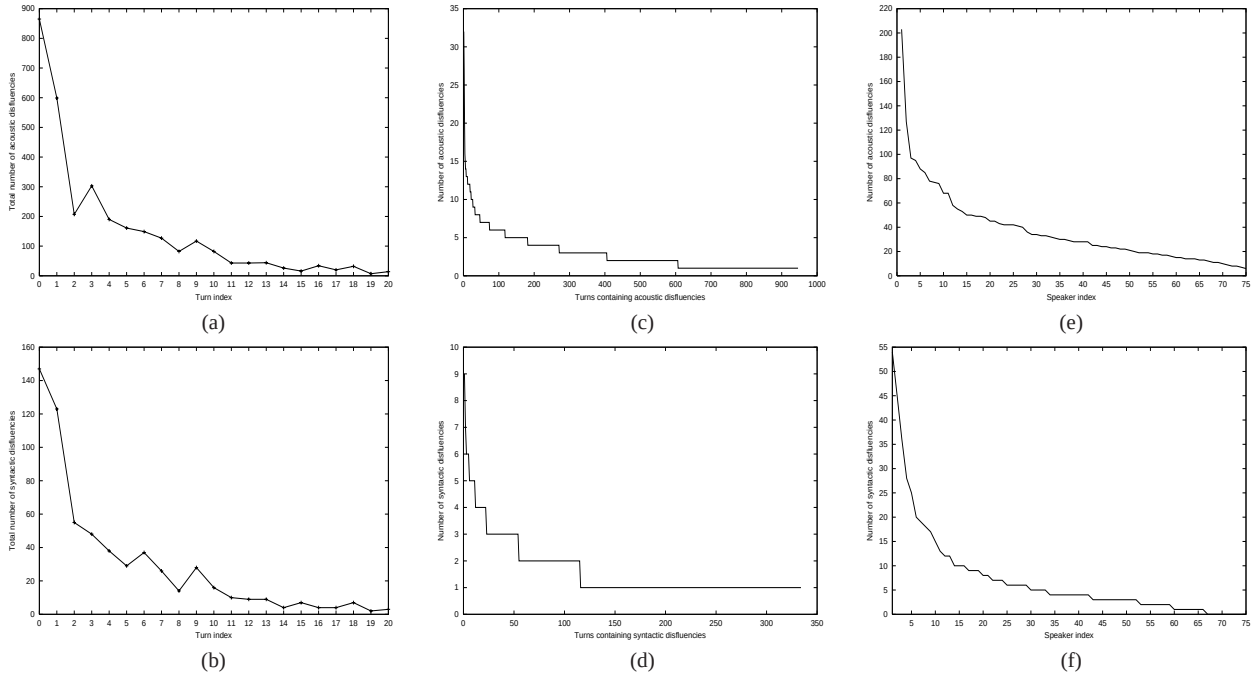


Figure 1: Graph (a) gathers acoustic disfluencies appearing in user turns with the same index, and shows the sums for the first 21 turns; graph (c) shows the counts of acoustic disfluencies for each user turn, putting them in decreasing order and leaving aside turns with no acoustic disfluencies; graph (e) shows the counts of acoustic disfluencies for each speaker, in decreasing order. Graphs (b), (d) and (f) show the same for syntactic disfluencies.

these corrected until no errors were found. By slightly modifying its source code, the parser was easily adapted to other tasks, like translating annotations from simplified format to XML, or producing various kinds of enhanced ortographic transcriptions.

Finally, to guarantee the coherence of the annotations, one single expert reviewed all of them. At the same time, speech signals corresponding to user turns were re-segmented according to the annotated phenomena, so that speech signals and their ortographic counterparts became completely coherent. The resulting database was composed of 227 text files with very reliable annotations of disfluencies and 1657 binary files containing the speech signals corresponding to coherently segmented user turns.

The categories, XML marks, attribute values and simplified marks specified in the annotation manual, along with the appearing counts for the database *OZ1* are shown in Table 1. For a lack of space, the more specific marks reserved for the *reparandum* (m), the *correcting signal* (s) and the *repair* (c) in retracings are not showed.

5. Discussion

In this Section we will try to analyze the disfluencies appearing in the database *OZ1*. This will help to identify those features that make spontaneous speech so difficult to recognize, and give ideas about which elements of the recognition software should be improved.

As shown in Table 1, the most common events were noises. This was due to the high degree of detail of annotations. Although most times speaker aspirations and speaker lips were hardly audible, we wanted detailed annotations to allow the recognition of various kinds of *silence*, which could improve the segmentation of speech signals, thus yielding more accurate acoustic models. After noises, acoustic disfluencies: lengthenings, silence pauses and filled pauses, were the most common events. It must be said that, although a very wide range of silence pauses was observed, we considered a difficult task to assign them a duration attribute, so the annotation of silence pauses did not include duration information. The same was applied to filled pauses and lengthenings. It was left to recognition algorithms the correct alignment of such events.

It was found a sizeable amount of retracings, especially repetitions and substitutions, which denotes the importance of modeling this kind of disfluencies, even when speakers are not real users but recruited volunteers. Significant data about the distribution of acoustic –graphs (a), (c) and (e)– and syntactic disfluencies –graphs (b), (d) and (f)– are shown in Figure 1: graphs (a) and (b) gather disfluencies appearing in user turns with the same index, and show the sums for the first 21 turns; graphs (c) and (d) show the counts of disfluencies for each user turn, putting them in decreasing order and leaving aside the turns with no disfluencies; finally, graph (c) shows the counts of disfluencies for each speaker, putting them in decreasing order.

A detailed inspection of the annotations –see graph (b) of Figure 1– revealed that most retracings, accompanied by a remarkable number of acoustic disfluencies acting as correcting signals –as shown by graph (a) of Figure 1– were grouped into certain turns, especially the first one, where users showed a hesitating behaviour. In fact, this behaviour can appear at any time in the dialogues, but it's more probable at the beginning, when the user has not defined his needs yet and does not know the system capabilities. So a high variability can be observed in the distribution of disfluencies, with many turns showing a few or no disfluencies –most times these turns consisted of a few words like *"yes, please"*, *"no, thanks"*, *"on Tuesday"*, etc– and a reduced set of turns gathering most of them, as shown in graphs

(c) and (d) of Figure 1.

A second study was made by counting disfluencies for each speaker. As shown in graph (e) and (f) of Figure 1, there was a high variability in the distribution of acoustic and syntactic disfluencies in the set of speakers. A few speakers gathered most disfluencies. This study was detailed by considering six general categories: noises (N), silence pauses (P), filled pauses and lengthenings of sounds (F), lexical disfluencies (L), syntactic disfluencies (S) –putting together abandoned sentences and retracings– and discourse markers (D). Mean and deviation values for the whole set of speakers, and counts for 10 especially selected speakers are shown in Table 2. Some speakers were remarkably more prone to hesitate, repeat or correct fragments of speech than others, yielding generally much longer dialogues (speakers 9, 11 and 30), whereas others produced very short dialogues with a few disfluencies (speakers 25 and 31). As shown in Table 2 long dialogues show a high number of disfluencies, but the amount of disfluencies was not always correlated with the length of the dialogues: speakers 4, 19, 22, 45 and 69 show very similar times but the amount of disfluencies ranges from a total number of 36 to 120. This reveals that some speakers are intrinsically more *disfluent* than others.

Table 2: Full duration of user turns and counts of disfluencies for the three dialogues carried out by each of 10 speakers selected from the database *OZ1*. The symbol N stands for noises, P for unfilled pauses, F for filled pauses, L for lexical disfluencies, S for syntactic disfluencies and D for discourse markers. Mean and standard deviation values over the whole set of speakers are shown too.

Speaker	Full duration (sec)	N	P	F	L	S	D	Total
4	91.24	29	5	3	2	1	6	46
9	378.86	124	16	72	2	12	56	282
11	394.45	135	63	140	4	54	17	413
19	70.73	29	3	3	0	0	1	36
22	89.79	19	15	35	2	12	9	92
25	42.90	14	1	6	1	1	4	27
30	478.15	160	52	75	17	45	21	370
31	38.49	10	1	9	0	0	3	23
45	125.91	14	27	41	11	19	8	120
69	72.98	20	10	18	3	9	11	71
Mean	118.63	35.65	10.04	27.29	2.67	7.27	11.53	94.45
Deviation	75.65	25.98	10.28	23.56	3.42	9.74	9.96	70.02

6. Conclusions

The main features of a spontaneous speech database consisting of 227 dialogues in Spanish were introduced. The speech events considered as disfluencies were described. Both a XML annotation format and a simplified format –to make easier the annotation process– were presented. Also a very simple parser was implemented which helped to locate and correct errors in annotations. Finally, annotation data were shown and discussed, finding that acoustic, lexical and syntactic disfluencies must be all studied and modeled for the recognition of spontaneous speech. Statistics showed that disfluencies were not uniformly distributed in the set of user turns, being more probable at the beginning of dialogues. Also a high dependence on speaker was observed. Our current work concerns two issues: first, to extend this preliminary study by recording and annotating a bigger database for the same application, but with a full dialogue system prototype and real users; and second, to model acoustic disfluencies as a first step towards a more general scheme which will include modelling approaches for lexical and syntactic disfluencies.

7. References

[1] V. Zue, S. Seneff, J. Glass, J. Polifroni, C. Pao, T.J. Hazen, and L. Hetherington, "JUPITER: A telephone-based conversational interface for weather information," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 1, pp. 100–112, January 2000.

[2] L. Lamel, "Spoken language dialog system development and evaluation at LIMSI," in *Proceedings of the International Symposium on Spoken Dialogue*, Sydney, Australia, November 1998.

[3] A.L. Gorin, G. Riccardi, and J.H. Wright, "How May I Help You," *Speech Communication*, vol. 23, no. 1-2, pp. 113–127, October 1997.

[4] Air Travel Information System (ATIS), URL: http://www.ldc.upenn.edu/Catalog/ATIS.html.

[5] Switchboard: Telephone Speech Corpus, URL: http://www.ldc.upenn.edu/Catalog/LDC97S62.html.

[6] TRAINS: Spoken Dialogue Corpus (University of Rochester), URL: http://www.ldc.upenn.edu/Catalog/LDC95S25.html.

[7] The HCRC Map Task Corpus (University of Edinburgh and University of Glasgow), URL: http://www.ldc.upenn.edu/Catalog/LDC93S12.html.

[8] J.B. Mariño and J. Hernando, "Especificación de las grabaciones mediante Mago de Oz," Technical Report BS16AV10, Proyecto TIC98-0423-C06: Sistema de diálogo para habla espontánea en un dominio semántico restringido, Universidad Politécnica de Cataluña, November 1999, URL: http://gps-tsc.upc.es/veu/basurde/Home.htm.

[9] A. Sesma, J.B. Mariño, I. Esquerra, and J. Padrell, "Estrategia del Mago de Oz," Technical Report BS52AV22, Proyecto TIC98-0423-C06: Sistema de diálogo para habla espontánea en un dominio semántico restringido, Universidad Politécnica de Cataluña, December 1999, URL: http://gps-tsc.upc.es/veu/basurde/Home.htm.

[10] A. Bonafonte and N. Mayol, "Documentación del corpus INFOTREN-PERSONA," Technical Report BS14AV20, Proyecto TIC98-0423-C06: Sistema de diálogo para habla espontánea en un dominio semántico restringido, Universidad Politécnica de Cataluña, June 1999, URL: http://gps-tsc.upc.es/ veu/basurde/Home.htm.

[11] E. Shriberg, *Preliminaries to a Theory of Speech Disfluencies*, Ph.D. thesis, University of California at Berkeley, 1994.

[12] Linguistic Data Consortium: Linguistic Annotation, URL: http://www.ldc.upenn.edu/annotation.

[13] Multilevel Annotation Tools Engineering (MATE), URL: http://mate.nis.sdu.dk/.

[14] L.J. Rodríguez, I. Torres, and A. Varona, "Manual para el etiquetado de disfluencias," Technical Report BS12BV30, Proyecto TIC98-0423-C06: Sistema de diálogo para habla espontánea en un dominio semántico restringido, Universidad del País Vasco, May 2000, URL: http://gps-tsc.upc.es/veu/basurde/Home.htm.