Annotation of disfluencies in Spanish dialogues*

Luis Javier Rodríguez, Inés Torres, Amparo Varona

Departamento de Electricidad y Electrónica. Facultad de Ciencias.
Universidad del País Vasco / Euskal Herriko Unibertsitatea (UPV/EHU).
Apartado 644. 48080 Bilbao. Spain.
{luisja, manes, amparo}@we.lc.ehu.es

Abstract

This paper presents the inventory of disfluencies and the annotation procedure for a set of 227 Spanish dialogues recorded according to the well known *Wizard of Oz* paradigm. A XML-like annotation scheme was designed and used, which accounted only for disfluencies happening in the dialogues. A first draft of the manual for annotators was written and iteratively tested, corrected and augmented over a representative subset of dialogues. Finally the whole set of dialogues was annotated with acoustic, lexical and syntactic disfluencies, as well as discourse markers, using the ultimate version of the manual. Only user turns were annotated, because *WoZ* turns were automatically synthetized according to a collection of rules and templates. A very simple parser was implemented, which helped to locate most errors in annotations. A detailed inspection of the annotations revealed that most disfluencies were grouped into certain user turns. Statistics show that acoustic phenomena: noises produced by user, lengthening of sounds, silence pauses and filled pauses, were the most common disfluencies. On the other hand, disfluencies were not uniformly distributed among speakers. Some speakers were remarkably more prone to hesitate, repeat or correct fragments of speech than others.

Keywords: Spontaneous Speech, Disfluencies, Linguistic Annotation.

1 Introduction.

In the mid nineties large vocabulary continuous speech recognition technology achieved the big goal of translating read speech to its text correlate 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. We have recently opened the line of spontaneous speech recognition, with special emphasis on Spanish language. This will require the acquisition and detailed annotation of generic and application specific databases, new modeling assumptions and more powerful

*Work partially supported by the Spanish CICYT, under project TIC98-0423-C06-03.

algorithms. Here we present the first milestone, which is the production of an application specific database in Spanish language, serving as benchmark to study and to model acoustic, lexical and syntactic disfluencies. This database -which we call *OZ1*- is composed of 227 dialogues where users asked for timetables, prices and some specific features of train travels between two spanish cities. Unlike the rapidly growing number of spontaneous speech databases for English [4][5][6][7], as far as we know no corpus with annotation of disfluencies is available for castilian Spanish, so this work can be considered as a pioneering effort.

The rest of the paper is organized as follows: Section 2 gives some background by defining the concept of disfluency and describing the speech events classified under such category. Section 3 presents the annotation format defined specifically for this work. The main features of our database, the inventory of disfluencies, some details about the annotation process and statistics of disfluencies are shown and discussed in Section 4. Finally, Section 5 gives some conclusions and perspectives for future research.

2 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. Under the category of acoustic disfluencies we consider noises, speech rate variations, lengthening of sounds, silence -or unfilled- pauses and filled pauses. Noises are a good example of events that, though not disfluencies in the strict sense, do seldom appear in read speech, but are pervasive in spontaneous speech due to recording conditions. Noises could be classified into more detailed categories, attending to the source and type: speaker aspirations, lips, coughs or laughs, background typing, computing, ringing or traffic, etc. or attending to their duration: short -usually isolated- and long -usually overlapping speech- noises. Unlike noises, filled and unfilled pauses and lengthening of sounds carry out specific functions in spontaneous speech, usually to give the user time to plan what he is going to say, or also when hesitating, to mark a correction. Various acoustic realizations can be found for filled pauses, either vowels (’a’, ’e’) or nasalizations (’m’, ’n’). Additionally, all kind of pauses and lengthenings can be assigned a duration parameter.

Lexical disfluencies account for not properly -or not canonically- pronounced words, and more generally for new dictionary entries, not found in written language. In fact, both filled and unfilled pauses should be processed as new dictionary entries. Spontaneous speech is far more relaxed than read speech, so a high number of popular or familiar expressions will 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 or simply errors in pronunciation.

Spontaneous speech shows a very wide range of syntactic disfluencies: false starts,

unfinished sentences, sentences completing a previous one, missing words, lacks of concordance, etc. and as a separate category the so called *retracings*, which account for repetitions, substitutions and reformulations with insertion or deletion of words. Some references [8] apply a very restrictive definition of disfluency, leaving aside noises, lexical disfluencies and most syntactic deviations from written language. In fact, they do not clearly distinguish between retracings and disfluencies, handling filled and unfilled pauses as sub-categories depending on retracings. The structure of retracings is always the same: a segment to be repaired -called *reparandum*-, a segment marking the correction -called *signal*- which may include filled or unfilled pauses and some editing phrases like 'sorry' or 'I mean', and a third segment -called *repair*- giving the replacing material, which can be a repetition (1), a substitution (2) or a more complex reformulation with insertion (3) or deletion (4) of words, as shown in the following examples, taken from [8]:

1. show me flights from boston on um on monday
2. show me flights from boston uh baltimore on monday
3. show the flights um the early morning flights to boston
4. show me the which morning flights go to boston

Some spontaneous speech events are difficult to classify. Two good examples happening in dialogues are overlapping of turns -two speakers talking simultaneously- and third party conversations -one of the speakers talks to a third person in the room. Finally, we will refer to *discourse markers* , very usual words or phrases without any specific meaning but carrying out a meta-linguistic 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, they are very closely related to spontaneous speech. In fact, they seldom appear in written language.

3 Annotation format.

After an exhaustive revision of almost all the existing formats and tools for dialogue annotation [9], a XML-like annotation scheme was designed, which accounted only for disfluencies happening in *OZ1*, as explained in Section 4. The annotation scheme was accompanied by the corresponding manual [10]. 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, some of them taking a finite set of values, others taking integer or real values. For the database *OZ1* three attributes were used: *type*, *source* -only for noises- and *word* -only for lexical disfluencies. We did not find

any adequate graphical annotation tool, so we decided to add marks by 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 -most times two letters encoding both the mark and the value of attribute *type* -enclosed between parentheses and affecting some text. The following example shows the XML-like annotation (1) and the simplified annotation (2) of the fragment 'show me flights' with external background noise:

1. <n source="world" type="generic">show me flights</n>
2. (nw show me flights)

The XML marks and attribute values and the corresponding simplified marks used for the database *OZ1* are shown in Table I.

4 Annotating disfluencies in the database *OZ1*.

4.1 The database *OZ1*.

Our spontaneous speech database consists of 227 Spanish dialogues, recorded at 8 kHz across telephone lines applying the well known *Wizard of Oz* (*WoZ*) mechanism[1]: a human operator which simulates the behaviour of the dialogue system, including recognition and/or understanding errors, so that users could think they were interacting with a real system. 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 information service operators[2]. Recruited users were told to 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, preferably using short sentences. 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 as they would with a real operator. The database includes 1657 user turns, lasting about 150 minutes. Although we consider this database large enough to study spontaneous speech and to carry out preliminary experimentation, a larger and hopefully more spontaneous set of dialogues will be recorded when a first version of the system be available. Those recordings should help to model more accurately the dialogue task and the disfluencies happening in that context.

[1]Recordings were made at the *Universidad Politécnica de Cataluña*, and preliminary ortographic transcriptions created by contracted annotators at the *Universidad de Zaragoza*.

[2]This preliminary database was transcribed to plain text but not used for the adverse noise conditions and the complexity of the interactions, which included many speech overlaps, third party conversations, etc.

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

Category	XML	type	source	word	Simplified	Counts
Noises	n	generic	world	-	nw	661
		air	speaker	-	na	1404
		lips	speaker	-	nl	600
		cough	speaker	-	nt	9
Lengthening 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	-	*canonical version*	lu	95
		mispronounced	-	*canonical version*	lm	105
False starts		-	-	-	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

4.2 The inventory of disfluencies.

Coverage and coherence were the key requirements for the inventory of disfluencies. Therefore, among all possible disfluencies, only those with enough number of samples in *OZ1*, plus maybe some others considered significative, should be annotated and modeled. To verify the kind and frequency of the disfluencies that appear in *OZ1*, a tentative set of dis-

fluencies was defined covering all the disfluencies we could expect in human-machine communications and leaving aside many others which can be expected only in human-human dialogues. Starting from these considerations, the inventory of disfluencies was dinamically modified as the process of annotation itself was configured, using a validation set of 10 representative dialogues. Three annotators processed separately these dialogues applying a first draft of the manual which was iteratively discussed, augmented and corrected. Version 3.0 of the manual was definitive and included the categories and marks shown in Table I.

4.3 The annotation process.

Once established the inventory of disfluencies and the annotation format, the whole set of dialogues was annotated with acoustic, lexical and syntactic disfluencies, as well as discourse markers, using the ultimate version of the manual. Only user turns were annotated. As said above, a simple text editor was used to add the marks. This task was accomplished by the same three experts that developed the manual, starting from the speech signal (each dialogue was stored in a separate file) and the original ortographic transcriptions. 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 applied iteratively until no errors were found. Besides locating errors, other tasks could be carried out by slightly modifying the source code of the parser, like generating a new annotation file with XML marks starting from the file in simplified format, or producing various kinds of ortographic transcriptions. Finally, we obtained 227 text files with very reliable annotations of disfluencies and 1657 binary files containing the speech signals corresponding to correctly segmented user turns. Appearing counts for each annotation mark are shown in Table I.

4.4 Discussion.

As shown in Table I, the most common events were noises, partly due to recording conditions and the high degree of detail of annotations. In fact, recording conditions were quite controlled -almost laboratory- and more noisy conditions can be expected in real dialogues. On the other hand, although most times speaker aspirations and speaker lips were nearly silence, 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: lengthening of sounds, 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 -*short, normal, long, very long*-, so the annotation of silence pauses did not include duration information. The same was applied to filled pauses and lengthening of sounds. It was left to recognition algorithms the correct alignment of such events.

With regard to discourse markers, most of them corresponded to opening or closing phrases ('hola', 'buenos días', 'adios', 'gracias', etc.) and filling expressions ('bueno', 'a ver', etc.). The sub-category *Request* included mainly the phrase 'por favor', used to ask the system for information. On the other hand, the words annotated in the sub-category *Explain* appeared mainly as correcting signals in retracings.

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 and instructed volunteers. Also a detailed inspection of the annotations revealed that most retracings were grouped into certain turns, especially the first one, where user showed a hesitating behaviour.

Table II. Durations of user turns and counts of disfluencies for 10 selected speakers of the database *OZ1*. Mean and standard deviation values over the whole set of speakers are shown too.

Speaker id	Duration of user turns (sec)	N	P	F	L	S	D	Total
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
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

A second study was made by counting disfluencies for each speaker. Six general categories were considered: noises (N), silence pauses (P), filled pauses and lengthening of sounds (F), lexical disfluencies (L), syntactic disfluencies (S) -putting together false starts and retracings- and discourse markers (D). Mean and deviation values for the whole set of the speakers, and counts for 10 especially selected speakers are shown in Table II. A high speaker variability in the distribution of disfluencies can be observed. 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). Some speakers were not really interested in extracting information, thus yielding very short dialogues with a few disfluencies (speakers 25 and 31). As expected, the counts of noises did show a clear correlation with the length of the dialogues. However, 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. Note again that the speakers of *OZ1* were not real users, but recruited volunteers. It can be expected that real users show a higher variability.

5 Conclusions and future work.

The speech events considered as disfluencies were described, and both a XML-like annotation format and a simplified format -to make easier the annotation process- were defined. The main features of a spontaneous speech database consisting of 227 dialogues in Spanish were introduced. Then a representative subset of the dialogues was explored to establish a suitable set of disfluencies. 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. On the other hand, statistics showed a high dependence on speaker. Future work will include modeling of acoustic disfluencies, which might imply the use of durational models or the inference of more adequate topologies. This would help to locate disfluencies at higher levels. Also a more flexible vocabulary, augmented with fillers and pronunciation variants of words, will be used.

References

[1] V. Zue, S. Seneff, J. Glass, J. Polifroni, C. Pao, T.J. Hazen, L. Hetherington. *"JUPITER: A Telephone-Based Conversational Interface for Weather Information"*. IEEE Transactions on Speech and Audio Processing, Vol. 8, N. 1, pp. 100-112, January 2000.

[2] L. Lamel. *"Spoken Language Dialog System Development and Evaluation at LIMSI"*. International Symposium on Spoken Dialogue, Sydney, Australia, Nov. 1998.

[3] A.L. Gorin, G. Riccardi, J.H. Wright. *"How May I Help You"*. Speech Communication, Vol. 23, N. 1-2, pp. 113-127. October 1997.

[4] ATIS: http://morph.ldc.upenn.edu/Catalog/ATIS.html

[5] Switchboard: http://morph.ldc.upenn.edu/Catalog/LDC97S62.html

[6] TRAINS: http://morph.ldc.upenn.edu/Catalog/LDC95S25.html

[7] Map Task: http://morph.ldc.upenn.edu/Catalog/LDC93S12.html

[8] E. Shriberg. *"Preliminaries to a Theory of Speech Disfluencies"*. PhD Thesis. University of California at Berkeley. 1994.

[9] L.J. Rodríguez. *"Anotación de corpora para diálogo"*. Technical Report BS12AV02. Proyect TIC98-0423-C06. December 1999.

[10] L.J. Rodríguez, I. Torres, A. Varona. *"Manual para el etiquetado de disfluencias"*. Technical Report BS12BV30. Proyect TIC98-0423-C06. May 2000.