A quantitative study of disfluencies in formal, informal and media spontaneous speech in Spanish

Antonio Moreno Sandoval[1], Leonardo Campillos Llanos[1], Doroteo T. Toledano[2]

[1]Laboratorio de Lingüística Informática (LLI), Universidad Autónoma de Madrid,
28049 Madrid, SPAIN
{antonio.msandoval,leonardo.campillos}@uam.es

[2]ATVS Biometric Research Laboratory, Universidad Autónoma de Madrid,
28049 Madrid, SPAIN
doroteo.torre@uam.es

Abstract. A descriptive study of the prevalence of different types of disfluencies (fragmented words, restarts and vocalic supports) in spontaneous Spanish is presented based on a hand-annotated corpus. A quantitative account of differences among three types of registers (formal, informal and media) and several subtypes of text for each register is provided to analyze the importance of each disfluency class for a given register.

Keywords: Fragmented words, restarts, fillers, Spanish.

1 Introduction

The study of disfluencies is a key topic in current research. On the one hand, descriptive and psychological investigations are focused in the nature of the phenomena [2, 4-9, 13, 27, 28]. On the other hand, applied research in spontaneous speech processing is trying to detect and handle disfluencies [18]. For instance, the National Institute of Standards and Technology (NIST) has launched a series of competitive evaluations under the name Rich Transcription [19] in which spontaneous speech processing, and particularly disfluency detection, is one of the goals. In fact, disfluencies are one of the main problems in ASR [11].

Most of the research on disfluencies focuses in English language, although there are also studies for French, English, German, Italian, Japanese, Estonian or Mandarin Chinese [20-24] and also for European Portuguese [16, 29]; furthermore, there is some research on fillers performed with multilingual data [4]. Spanish has not been one of the languages with more research in this area, and few groups are conducting research in the field of spontaneous speech processing [25, 26]. Analogously, there are a few corpora available for spontaneous speech in Spanish. This study is based on the Spanish data from the C-ORAL-ROM corpus [17]. Previous research on disfluencies in this corpus has focused on acoustic-phonetic decoding of spontaneous speech [30] or has discussed the difficulties for the human transcribers [10]. The main con-

clusion of this last work [10] was that disfluencies are not the hardest problem for human transcribers, but the interaction features such as overlapping or speed rate. Another result was that media recordings are easier to transcribe than formal speech (talks in public context such as conferences, lectures, etc.). The hardest register to transcribe is, unsurprisingly, the informal, private speech. Another piece of research based on C-ORAL-ROM used also data from the MAVIR corpus [14], a collection of recordings and transcriptions gathered from professional conferences on language technologies and corporate presentations; in consequence, this research concentrated only in the formal register [3].

The goal of this paper is not a linguistic nor an acoustic description such as [3, 10, 30], but a quantitative analysis of three types of disfluencies (fragmented words, fillers and restarts) which may be of interest for ASR tasks. The choice of these types of phenomena is due to way they are encoded in the corpus, which makes them suitable for this brief analysis. Section 2 describes the C-ORAL-ROM corpus and each type of disfluency analyzed; and section 3 explains the results for every phenomenon.

2 Description of the C-ORAL-ROM Corpus

C-ORAL-ROM is a multilingual linguistic data bank that comprises four romance languages: Italian, French, Portuguese and Spanish. In this work only the Spanish sub-corpus [17], which contains around 300.000 spoken words, has been used. From a sociolinguistic point of view, speakers are characterized by their age, gender, place of birth, educational level and profession. From a textual point of view the corpus is divided into the parts shown on Table 1 [17]. This subdivision follows the design criterion of the C-ORAL-ROM corpus, which was mainly performed on the basis of register (formal and informal) [15]. A remarkable feature of the corpus design is the importance of the informal register, which means more than half of the corpus size.

In this study, in order to have a more balanced distribution between the main registers a large subset of the Spanish corpus has been selected, excluding telephone conversations and informal speech in public context (for example, narrations on a topic between unfamiliar speakers). Table 1 shows the subclasses considered for each of the three registers, along with the number of words in each of them. For the informal speech, the classification is based on the dialogic nature, and formal speech is divided between natural context (i.e. discourse recorded in a public setting: conferences, teaching, presentations, etc.) and media speech (or broadcast news speech). We have to point out that recordings from the media may not always be strictly classified in the formal register, since discourse from the sports programs and the talk shows may contain a quite informal speech.

An interesting feature of the C-ORAL-ROM corpus is that several types of disfluencies are annotated by hand. Trained linguists manually transcribed every recording, and along the orthographic transcription they marked disfluency phenomena following rigorous conventions [15]. Table 2 below summarizes and explains (with examples) the conventions used to mark every type of disfluency phenomena studied in this article.

Table 1. Distribution of words in C-ORAL-ROM (sum of words in all the files of each type).

Register		Type of texts	Words
FORMAL	Natural context	Preaching	12941
		Law	6203
		Political debate	6055
		Conference	12275
		Teaching	12353
		Prof. explanation	12063
		Business	9034
		Total	**70924**
	Media	Weather news	1591
		Interview	7640
		Reportage	35190
		Science	6108
		News	17335
		Sports	9330
		Talk-show	19976
		Total	**97170**
INFORMAL		Conversation	23495
		Dialogue	63590
		Monologue	41229
		Total	**128314**

Table 2. Transcription marks (with examples) for the disfluencies analyzed.

Mark	Disfluency	Meaning	Example
[/]	Simple retracting or restart.	Repetition or retrace.	no te hablo de [/] de [/] de equipos (emedsp02) ('I am not speaking to you of [/] of [/] of teams') esto facilita la acción [/] el éxito de la acción ('this facilitates the action [/] the success of the action') (enatbu03)
[///]	Retracting or restart.	Syntactic reformulation.	para &se [///] que te siga llamando (emedsp02) ('to &ke [///] for him to keep on calling to you')
&	Before a fragmented or truncated word.	A non-complete word (self-correction).	trabajamos mucho para &Sudamer [/] Sudamérica ('we work a lot for South &Am [/] South America') (enatpe01)
&eh &ah &mm	Vocalic support or filler.	The speaker uses it to keep his / her turn.	saben / &eh / cómo / puede mejorar su producto ('they know / er / how / their product can improve') (enatbu03) es algo / &mm / muy interesante (enatps01) ('it is something / um / very interesting')

Although an inter-annotator agreement test was not performed, every transcription was reviewed by another linguist to guarantee the accuracy of the data.

3 Results

In order to estimate the prevalence of the three types of disfluency in each register and subclass, the procedure used has been:

1. Count the number of marks for each kind of disfluency in every text.
2. Calculate the ratio between number of words in each text and the number of occurrences of a given mark in the text (i.e. the average number of words by disfluency in each recording).
3. Calculate the average ratio for the texts in the same subclass and register.

Results are shown numerically on Table 3 and graphically in Figures 1 to 3. The prevalence measure used is inversely related to the frequency of the phenomenon in that subclass: a higher number means that the disfluency is less frequent and vice versa.

Table 3. Average number of words between occurrences of each type of disfluency analyzed.

Register		Type of texts	Truncated	Supports	Rep./Restart
FORMAL	Natural context	Preaching	468,69	146,54	471,33
		Law	240,49	130,04	97,53
		Political debate	206,06	70,36	123,13
		Conference	138,93	134,29	91,82
		Teaching	129,42	63,74	45,90
		Prof explanation	97,27	68,99	35,78
		Business	118,27	40,54	45,33
		Average	**199,88**	**93,50**	**130,12**
	Media	Weather news	0,00	554,00	259,50
		Interview	155,56	31,00	44,85
		Reportage	605,34	79,02	118,77
		Science	202,94	55,23	78,45
		News	619,35	233,42	224,88
		Sports	124,06	102,86	58,80
		Talk-show	187,80	172,14	59,61
		Average	**270,72**	**175,38**	**120,69**
INFORMAL		Conversation	339,87	492,55	50,89
		Dialogue	133,17	406,14	42,64
		Monologue	190,10	120,85	40,01
		Average	**221,05**	**339,84**	**44,51**

Figure 1: Distribution of the **average number of words per fragment** in the different types of texts.

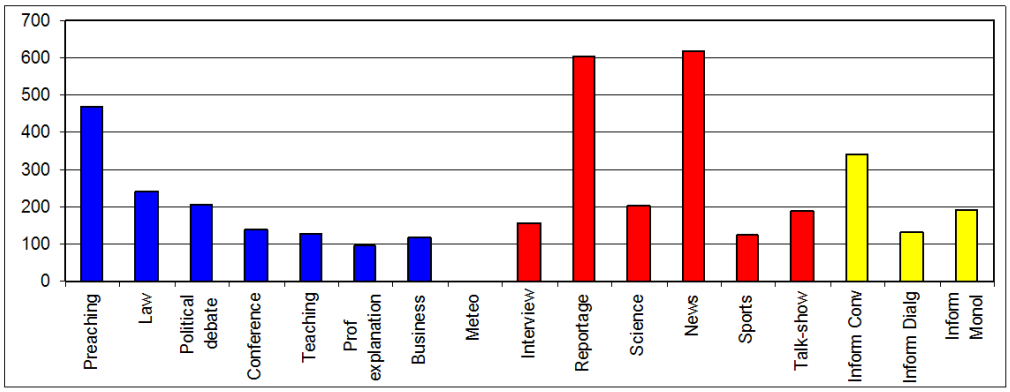


Figure 2: Distribution of the **average number of words per restart** in the different types of texts.

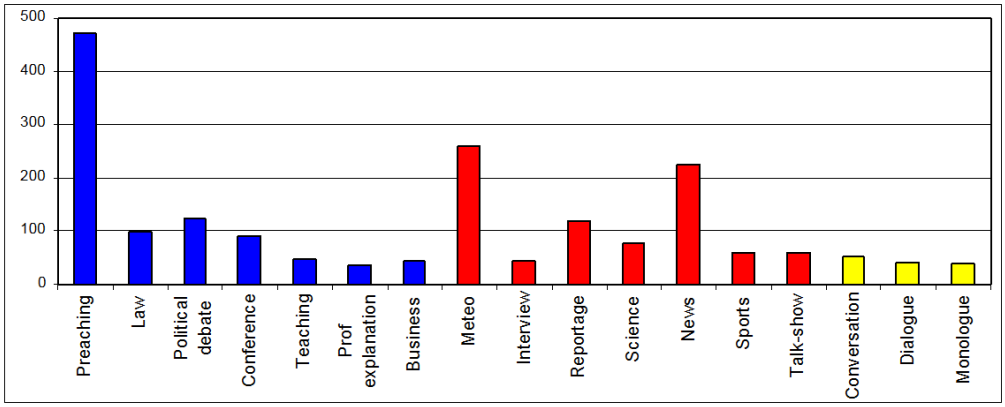
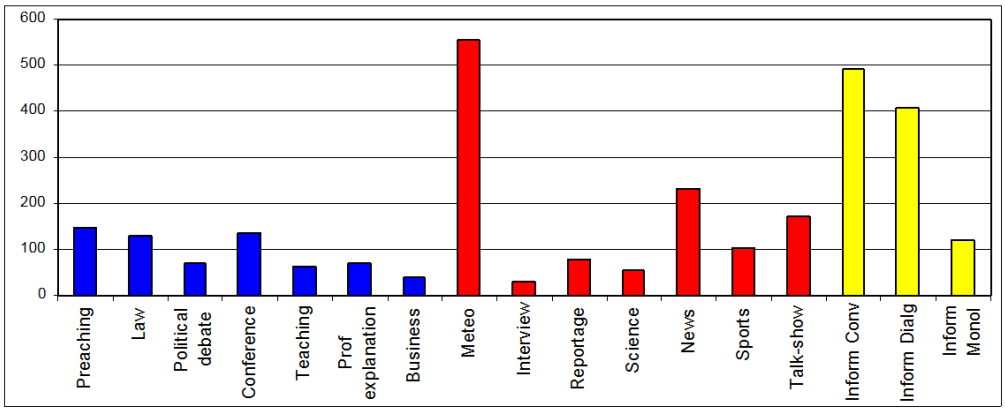


Figure 3: Distribution of the **average number of words per support** in the different types of texts.



3.1 Fragmented words

A fragmented (or truncated) word is an incomplete pronunciation of a given lexical item. Truncated words abound in spontaneous speech since speakers are thinking while they are speaking, and they retrace and self-correct when a segment is already uttered. According to a still unpublished research performed in a phonological lexicon of over 400,000 items, in Spanish the most frequent number of syllables for a full inflected word is 4, followed by 5 and 3. The number of monosyllables and bi-syllables in the same Spanish lexicon is less than 20,000 (only 5% of the words). That could explain that this disfluency is relatively common in Spanish.

Results for fragmented words are shown in both Table 3 and Figure 1. Fragmented words are a very rare in media texts, especially in news and reportages, probably because of their almost-read nature. In the meteorological reports there are no occurrences. The homilies (non-read commentaries on the Holy Writings during the mass) are also a subclass in the formal register, where the truncated words are very infrequent. In contrast, they abound in several formal subclasses (professional explanations, business, teaching and conferences) as well as in informal dialogues.

3.2 Restarts and repetitions

The retracting is the most frequent fragmentation phenomenon in spontaneous speech; in the scientific literature, they have also been called *retrace and repair sequences* [1], *restarts* or *false starts* [15], or *repairs* [18]. Actually, the term *reparandum* is frequently used to refer to the sequence of syllables or words (or even the beginning of an utterance) that is deleted, while the replacing fragment is called *repair* [12, 27]:

un / interesante debate / que también **vas** [/] **vamos** a tratar (enatte01)
REPARANDUM REPAIR
(‘an / interesting debate / which you are also going to [/] we are going to discuss’)

The speaker hesitates while trying to find the best way to express himself and retracts his speech before choosing between two alternatives. The retracting phenomenon is almost always accompanied by the repetition (complete or partial) of the linguistic material and clearly causes a loss of the informational value of the retracted material, which is abandoned by the speaker in favor of the chosen alternative [15].

In our analysis, we counted together both the restarts and the repetitions (or repeats, as they are also referred to in some articles [7]). The most frequent repetitions and retraces are short items (usually just one word) and mainly mono- or disyllabic (as a matter of fact, prepositions and determiners the majority of the times) [3].

As in the previous case, restarts and repeats are very rare in preaching, weather forecast and broadcasting news. The retracting appears very often in informal speech, but also in several subclasses of formal and media registers: professional explanation, teaching and business, on formal speech; and interviews, sports and talk shows, on the media domain. In all those subclasses, its interactive nature of the process could explain why speakers hesitate so frequently.

3.3 Vocalic supports

Vocalic supports (or filled pauses) are a paralinguistic element used for helping the speaker to think about the way the next utterance must be expressed. In the literature, they are also called *fillers* [6, 7], *pause fillers* or *hesitators* [1]. In ASR of spontaneous discourse, fillers tend to be included in the grammar module along with full lexical words [12]; thus, it is important is to model filled pauses in a way that they would not be erroneously recognized. For example, Spanish *eh* may be mistaken with the conjunction *e* (used as a variant of *y* before a syllable beginning with *i-*/*hi-*) or *ah* may be interpreted as the preposition *a* (and vice versa). According to previous research in Spanish formal speech [3], the most frequent filler is *eh,* whereas *ah* is scarce.

As shown in Figure 3, the filled pauses are typical of the formal speech, where the speaker wants to be precise in his or her speech. On the other side, vocalic supports are scarce in informal speech, where the speed of the interaction and a more relaxed situation do not favor the filled pauses. In media recordings, those subclasses closer to the informal speech (talk shows and sport) have less average number of vocalic supports. Weather and broadcast news subclasses are the ones with less number of filled pauses, mainly because their near-read nature. Some research based on Spanish data from formal register [3] has stated that the frequency of filled pauses in every subregister may be influenced by other factors different from the dialogic style or the genre: for instance, the speaker's speaking style, the difficulty related to the degree of complexity of the contents or the anxiety induced by the communicative situation.

4 Conclusions

The present empirical study on the prevalence of tagged disfluencies in a spontaneous speech corpus of Spanish shows interesting quantitative data about the frequency and distribution among subclasses and registers of three important types of disfluencies. Retractings and vocalic supports are more frequent than fragmented words in all the types of recordings. These results, based on data from informal and formal register, match up with the results obtained from formal data in previous research [3]. Attending only disfluency type and prevalence, instead of register, it would be more accurate to divide the corpus into the following classes that share similar disfluency rates:

- Professional explanation, teaching, business, interviews, sports and talk shows present a high frequency of retractings.
- Talk shows and sport news present a similar, high rate of vocalic supports to that in informal speech.
- Preaching speech, weather and broadcast news do not abound with truncated words

The results of this study could be interesting in the design of ASR tasks, since they provide a measure of the disfluency rates per type of disfluency found in Spanish for different types of speech. This will allow researchers in the field of ASR to focus on the most important types of disfluency for the particular type of speech they have to process.

References

1. Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E.: Longman Grammar of Spoken and Written English. London: Longman. (1999)
2. Boula de Mareüil, P., Habert, B., Bénard, F., Adda-Decker, M., Baras, C., Adda, G., Paroubek, P.: A quantitative study of disfluencies in French broadcast interviews. In Proc. ISCA Workshop on Disfluency in Spontaneous Speech, DiSS 2005 (2005)
3. Campillos, L., Alcántara, M.: Speech Dysfluencies in Formal Context. Analysis based on Spontaneous Speech Corpora. In Proc. Corpus Linguistics Conference 2009. (2009)
4. Candea, M., Vasilescu, I., Adda-Decker, M.: Inter- and intra-language acoustic análisis of autonomous fillers. Proceedings of DiSS'05, Disfluencies in Spontaneous Speech Workshop 2005, Aix-en-Provence (2005)
5. Campione, E., Véronis, J. Pauses and hesitations in French spontaneous speech. Proc. ISCA Workshop on Disfluency in Spontaneous Speech, DiSS 2005. (2005)
6. Clark, H. Speaking in time. Speech Communication, 36, pp. 5-13 (2002)
7. Clark, H. & Wasow, T. Repeating Words in Spontaneous Speech. Cognitive Psychology, 37(3), pp. 201-242 (1998)
8. Ferreira, F., Lau, E. F., Bailey, K. G. D.: Disfluencies, language comprehension, and Tree Adjoining Grammars. Cognitive Sciences, 28, pp. 721-749 (2004)
9. Gilquin, G., De Cock, S.: Errors and Disfluencies in Spoken Corpora. Special issue of International Journal of Corpus Linguistics, 16:2. Amsterdam: John Benjamins (2011)
10. González Ledesma, A., De la Madrid, G., Alcántara Plá, M., De la Torre, R., Moreno-Sandoval, A.: Orality and Difficulties in the Transcription of Spoken Corpora. In Proc. of the Workshop on Compiling and Processing Spoken Language Corpora, LREC, 2004, Lisbon (2004)
11. Huang, X., Acero, A., Hon, H.-W.: Spoken Language Processing: A Guide to Thory, Algorithm and System Development. NJ: Prentice Hall PTR. (2001)
12. Jurafsky, D., Martin, J.: Speech and Language Processing. 2nd edition. Prentice Hall Series in Artificial Intelligence. (2008)
13. Levelt, W.: Speaking: from intention to articulation. MIT Press. (1989)
14. MAVIR Corpus. Distributed by the Fundación General de la Universidad Autónoma de Madrid (FGUAM) (2008)
15. Moneglia, M.: The C-ORAL-ROM resource. In Cresti, E., Moneglia, M. (eds.) p. 27 (2005)
16. Móniz, H., A. I. Mata, C. Viana: On filled pauses and prolongations in European Portuguese. Proc. INTERSPEECH – ISCA, August 27 - 31, pp. 2645-2648. (2007)
17. Moreno, A., De la Madrid, G., Alcántara, M., González, A., Guirao, JM., de la Torre, R.: The Spanish Corpus. In Cresti, E., Moneglia, M. (eds.) C-ORAL-ROM: Integrated reference Corpora for Spoken Romance Languages, pp. 135-161. Amsterdam: John Benjamins (2005)
18. Nakatani, C. H., J. Hirschberg: A Corpus-based study of repair cues in spontaneous speech. Journal of the Acoustic Society of America, 95:3, pp. 1603-1616. (1994)
19. NIST Rich Transcription Evaluation Homepage, http://www.itl.nist.gov/iad/mig/tests/rt/.
20. Proceedings of the ICPhS Satellite meeting on Disfluency in Spontaneous Speech, UC Berkeley DiSS-1999. (1999)
21. Proc. DiSS-2001, Disfluencies in Spontaneous Speech Workshop 2001, August 29-31, Edinburgh, Scotland, UK. http://www.isca-speech.org/archive_open/diss_01/ (2001)
22. Proc. DiSS-2003, Disfluencies in Spontaneous Speech Workshop 2003, September 5-8, Göteborg, Sweden. http://www.isca-speech.org/archive_open/diss_03/ (2003)

23. Proc. DiSS-2005, Disfluencies in Spontaneous Speech Workshop 2005, Aix-en-Provence. http://www.isca-speech.org/archive_open/diss_05/ (2005)
24. Proc. DiSS-2010, 5th Workshop on Disfluency in Spontaneous Speech, Tokyo. (2010)
25. Rodríguez, L. J., Torres, I., Varona, A.: Annotation and Analysis of Disfluencies in a Spontaneous Speech Corpus in Spanish. Proc. ISCA Workshop on Disfluency in Spontaneous Speech, DiSS 2001. (2001)
26. Rodríguez, Luis J., Torres, I.: Spontaneous Speech Events in Two Speech Databases of Human-Computer and Human-Human Dialogs in Spanish. Language and Speech, 49, 3, pp. 333-366 (2006)
27. Shriberg, E.: Preliminaries to a Theory of Speech Disfluencies. U. Cal. Berkeley. Ph.D. Thesis. (1994)
28. Shriberg, E.: Spontaneous Speech: How People Really Talk and Why Engineers Should Care. Proc. EUROSPEECH 2005 - INTERSPEECH 2005, Lisbon, Portugal, pp. 1781-1784. (2005)
29. Silva, A.: Caracterização segmental e prosódica de disfluências em discurso espontâneo. Master's Thesis. Universidade de Aveiro (2006)
30. Torre Toledano, D., Moreno Sandoval, A., Colás Pasamontes, J., Garrido Salas, J.: Acoustic-phonetic decoding of different types of spontaneous speech in Spanish, In Proc. DiSS-2005, Aix-en-Provence, pp. 165-168. (2005)