Spontaneous Speech Events in Two Speech Databases of Human-Computer and Human-Human Dialogs in Spanish

Luis J. Rodríguez, M. Inés Torres

University of the Basque Country, Spain

Key words

disfluencies

spanish

spontaneous speech recognition

Abstract

Previous works in English have revealed that disfluencies follow regular patterns and that incorporating them into the language model of a speech recognizer leads to lower perplexities and sometimes to a better performance. Although work on disfluency modeling has been applied outside the English community (e.g., in Japanese), as far as we know there is no specific work dealing with disfluencies in Spanish. In this paper, we follow a data driven approach in exploring the potential benefit of modeling disfluencies in a speech recognizer in Spanish. Two databases of human-computer and human-human dialogs are considered, which allow the absolute and relative frequencies of disfluencies in the two situations to be compared. The rate of disfluencies in human-human dialogs is found to be very close to that found for similar databases in English. Due to setup factors, the rate of disfluencies found in human-computer dialogs was remarkably higher than that reported for similar databases in English. In any case, from the point of view of speech recognition, the high frequencies of disfluencies and the distinct features of the acoustic events related to them support the need for explicit acoustic models. The regularities observed in the distribution of filled pauses and speech repairs reveal that including them in the language model of the speech recognizer may be also helpful. The extent to which the number of events depends on utterance length and on the speaker is also explored. Statistics are shown that follow previous studies for English, and a sizeable space is devoted to comparing our results with them. Finally, various possible cues for the automatic detection of speech repairs—a key issue from the point of view of speech understanding—are explored: silent pauses, filled pauses, lengthenings, cut off words and discourse markers. As previously observed for English, none of them was found to be reliable by itself. More information, especially at the acoustic-prosodic level, is no doubt needed to reliably detect speech repairs.

Acknowledgments: This work was partially supported by the University of the Basque Country, under Grant, 9/UPV-00224310-15900/2004, and the Spanish CICYT, under projects TIC2001-2812-C05-03 and TIN2005-08660-C04-03.Thanks to the reviewers and the editorial staff of *Language and Speech*, very especially to the Associate Editor Dr. Elizabeth Shriberg, who greatly helped with their comments to focus the paper and improve the presentation of data.

Address for correspondence. Luis J. Rodríguez, Departamento de Electricidad y Electrónica, Facultad de Ciencia y Tecnología, Universidad del País Vasco, Apartado 644, 48080 Bilbao, Spain; e-mail: <luisja@we.lc.ehu.es>.

1 Introduction

Current applications of speech recognition and understanding technology (information systems based on spoken dialog interfaces, speech-to-speech translation, etc.) must deal with spontaneous, unconstrained and most times noisy speech. Under these conditions, the performance of speech recognizers degrades. Word error rate (WER) in conversational speech is higher than 30% (Colthurst, Kimball, Richardson, Shu, Wooters, Iyer, & Gish, 2000; Hain, Woodland, Evermann, & Povey, 2000; Ljolje, Hindle, Riley, & Sproat, 2000; Sundaram, Ganapathiraju, Hamaker, & Picone, 2000), whereas in large vocabulary read speech (dictation) applications WER's of 10% or even lower are attained (Bahl, Balakrishnan-aiyer, Bellegarda, Franz, Gopalakrishnan, Nahamoo, Novak, Padmanabhan, Picheny, & Roukos, 1995; Ney, Welling, Ortmanns, Beulen, & Wessel, 1998; Riley, Ljolje, Hindle, & Pereira, 1995). The presence of spontaneous speech events such as hesitations and speech repairs, usually referred to as *disfluencies*, is an important source of errors, not only in speech recognition but also in speech understanding. Note, for instance, that in spoken dialog interfaces an accurate string of words is not so important as an accurate understanding of the user's requirements. So, detecting speech repairs and recovering the intended sequence of words is a key issue in speech understanding.

Previous work in English has revealed that disfluencies follow regular patterns. Various ways of incorporating them into the language model have been explored, leading to lower perplexities (Heeman & Allen, 1997; Siu & Ostendorf, 1996; Stolcke & Shriberg, 1996), and in some cases to a better performance of the recognizer (Heeman, 1999). On the other hand, some works report sizeable error reductions in speech recognition in English just by extending the baseline set of acoustic models with human and nonhuman noises, silent pauses and filled pauses (common in spontaneous speech), and including them in the language model as pseudowords (Liu et al., 1998; Rose & Riccardi, 1999; Schultz & Rogina, 1995).

In this paper, we follow a data driven approach in exploring the potential benefit of modeling disfluencies and other spontaneous speech events in a speech recognizer, rather than treating them as noise. The approach is similar to (Shriberg, 1994), but deals with conversational Spanish instead of conversational English. We are not aware of any study on the distribution of spontaneous speech events in Spanish, and though some work has been done outside the English community (e.g., Japanese: Heeman & Loken-Kim, 1999), we have not found any study on disfluency detection in Spanish either. Besides representing a pioneering effort in Spanish, the data presented in this paper will be useful to those interested in comparing the distribution of spontaneous speech events across different languages. Two databases of human-computer and human-human dialogs are considered. This allows the presence of disfluencies and other spontaneous speech events in the two situations to be compared.

The data presented in this paper are based on manual annotations. A suitable inventory of spontaneous speech events and the corresponding markup scheme are defined, starting from previous works for English, in particular those carried out for ATIS (Air Travel Information System: a corpus of spontaneous human-computer face-to-face dialogs in the air travel planning domain) and Switchboard (a corpus

of spontaneous human-human conversations over the telephone on various topics) (LDC94S19, 1994; Meteer, Taylor, Macintyre, & Iyer 1995). The analysis of counts and statistics of events and their positions inside utterances, which extends the analysis in previous works (Rodríguez & Torres, 2003; Rodríguez, Torres, & Varona, 2001a), allows us to propose what events are worth modeling and suggests ways of improving the recognition of spontaneous speech in Spanish. The high frequencies of speech and nonspeech acoustic events and the distinct features of some of those events support the need for explicit acoustic models. On the other hand, the regularities observed in the distribution of some of those events, filled pauses, speech repairs, and so forth, reveals that including them in the language model may be also helpful. In (Rodríguez, Torres, & Varona, 2001b) a very simple and straightforward approach was presented, in which the inventory of sublexical units is augmented with 12 additional units (accounting for lengthenings of sounds, filled pauses and noises), which are also included in the vocabulary and taken into account in the language model, leading to a better performance of a speech recognition system.

The extent to which the number of spontaneous speech events depends on utterance length is also explored. This is interesting from the point of view of speech understanding, since the more disfluencies usually means the more difficult to get the right interpretation, and reducing the number of disfluencies, for instance, through an adequate design of the spoken dialog interface, should help to better understand speech. Statistics are shown that follow previous studies for English (Oviatt, 1995; Shriberg, 1996) and some space is devoted to comparing our results with them. The number of events seems to depend linearly on utterance length, as was previously found for some databases in English (Shriberg, 1994). The same conclusion can be drawn from speaker statistics. Studying the dependence on the speaker allows to determine whether or not is worth a speaker adaptation strategy in automatic speech recognition at any of the acoustic, syntactic or semantic levels. Adapting the models to each particular speaker should improve the performance of the speech recognition system especially when dealing with highly fluent or highly disfluent speakers.

Finally, we test whether acoustic events, lexical distortions or even discourse markers, as suggested by other authors (Levelt, 1989, ch.12; Shriberg, 1994), can be used as cues to detect speech repairs. Detecting speech repairs, and possibly *correcting* them, is a key issue from the point of view of speech understanding. Note that even when the correct sequence of words is recognized, if the speech repair is not edited the right way, interpretation may be erroneous or at least ambiguous. We find that lexical distortions, in particular cut off words, often appear at the end of the portion of speech to be repaired, whereas filled pauses, silent pauses, lengthenings and discourse markers play a different role as editing terms that the speaker inserts while planning the repair. Nevertheless, as previously observed for English (Lickley, 1994, pages 32–43), no acoustic and prosodic event was found to be a reliable cue by itself. Moreover, lengthenings and cutoff words are not currently available in a speech recognizer's output. Lengthenings might be made available by some computation based on duration with some normalization, but cutoff words are not easy to detect. As previous works for English suggest (Bear, Dowding, & Shriberg, 1992; Lickley, 1994; Nakatani & Hirschberg, 1994), the proposed cues should be augmented and combined with more information, especially at the acoustic-prosodic level, to reliably

detect disfluencies, for example, as done in (Shriberg, Bates, & Stolcke, 1997; Stolcke, Shriberg, Bates, Ostendorf, Hakkani, Plauche, Tür, & Lu, 1998).

The rest of the paper is organized as follows. We begin by enumerating the main features of the spontaneous speech databases. We proceed by describing the inventory of spontaneous speech events in detail, including examples. Then the absolute and relative counts of events and their dependence on utterance length and on the speaker are shown and discussed. The analysis is followed by a preliminary test, not previously done for Spanish, of the potential usefulness of various cues of speech repairs, based on their frequencies and their coverage of speech repairs.

2 The spontaneous speech databases

The database of human-computer dialogs, which henceforth we will call INFOTREN, was collected in the framework of a Spanish project (Bonafonte, Aibar, Castell, Lleida, Mariño, Sanchís, & Torres, 2000) that aimed to provide automatic access to information about train schedules through a spoken dialog interface. Since no dialog system was working at that time, the well known *Wizard of Oz* mechanism was applied to emulate human-machine communication: A human operator simulated the behavior of the machine side, including recognition and/or understanding errors, so that users could think they were interacting with a real system. Each *dialog* of INFOTREN consisted of a strict sequence of user requests and system answers, which means that interactive events common in *true* dialogs, such as speech overlaps and backchannel events, will not be found there. INFOTREN consists of 227 dialogs, recorded at 8 kHz across telephone lines in an office environment (see details in Table 1). Only the user turns are used for modeling purposes. Note that the function of the speech recognizer in this case is just to feed the upper levels of the dialog system with the recognized sequence/lattice of words corresponding to each user turn.

The spontaneous speech events found in a human-computer domain-restricted task may differ in both type distribution and intensity from those found in human-human conversational speech. To carry out a comparative study, we consider a second database here, consisting of 42 face-to-face interviews taken from radio and TV broadcasts, which henceforth we will call CORLEC-EHU (see details in Table 1).

CORLEC-EHU is a subset of a larger database of spoken contemporary mainland Spanish called *Corpus Oral de Referencia de la Lengua Española Contemporánea* (CORLEC), recorded by the *Universidad Autónoma de Madrid* for use in theoretical studies of spoken language (Ballester, Santamaría, & Marcos-Marín, 1993). CORLEC is a representative corpus of casual, spontaneous, completely unrestricted but, unfortunately, quite noisy and low-quality speech. CORLEC is composed of a heterogeneous and domain-unrestricted set of conversations, monologs and interviews, taken from radio and television broadcasts, daily conversations, academic lectures, round-table discussions/debates, etc. *Informants* (speakers) were drawn from various sociocultural backgrounds, and dialogs were held in different situations, either formal or familiar (and all intermediate types).

CORLEC contains 941386 words (around 100 hours of speech), with a vocabulary of 39785 words, and comprises 17 blocks, defined according to either the semantic

Table 1
Main features of the human-computer and human-human spontaneous speech databases

	INFOTREN	*CORLEC-EHU*
f_s*(kHz)*	8	16
Environment	Office	Radio/TV studios
Channel	Telephone line	Analog tape
Interaction	Human-Computer	Human-Human
Domain	Information, task	Free
# Dialogs	227	42
# Speakers	75	118
# Turns	1657	2856
Duration (hours)	2.50	6.41
Turns/Dialog	7.30	68.00
Duration/Turn (seconds)	5.43	8.08
# Words	18734	72461
Vocabulary size	788	8237
Samples/Word	23.77	8.80

domain or the speech modality. The most generic blocks (interviews and conversations) were preselected to define a smaller subcorpus. Conversations are open dialogs involving two or more speakers, with a large number of overlaps, since turns are not given but freely taken. Interviews are more formal dialogs involving, in most cases, only two participants: a journalist and a famous person (a writer, a politician, an actress, etc.). Typically the famous person is introduced by the journalist, and then he/she must deal with several questions, sometimes briefly (1s or 2s) and other times spending too long on the answers (2mins or more), depending on the subjects (which are completely free). The conversations were recorded at home, in family meetings, or while traveling, so they were very noisy, with echo, and so forth. The interviews were all taken from TV or radio broadcasts, so signals were quite clean. After discarding noisy dialogs, a preliminary set of 67 interviews and 65 conversations was obtained. A subset of 42 interviews was drawn from it to define CORLEC-EHU. Those interested in the process of recycling and adapting a subset of signals and transcriptions of CORLEC to build CORLEC-EHU can find a description in (Rodríguez & Torres, 2003).

3 The inventory of spontaneous speech events

Spontaneous speech shows a number of acoustic and syntactic features that make it difficult to recognize and understand. Disfluency is the most challenging issue from the point of view of natural language processing and psycholinguistics. The term *disfluency* refers to any break in fluency, any interruption, pause and/or reformulation of the discourse, due either to macroplanning delays or to problems detected at various levels: phonological, lexical, syntactic, and so forth (Levelt, 1989, ch. 12).

Most authors use the term *disfluency* to refer to filled pauses, word fragments, repeats or self corrections. Acoustic events such as silent pauses and lengthenings, and some specific words and phrases, often appear playing the role of editing signals in self corrections. So we define a new concept, called *spontaneous speech event* (SSE), which includes not only disfluencies but also some other (related) events.

The set of SSEs is organized into four broad categories: (1) events that would need an acoustic model or that should be taken into account in the estimation of acoustic models of a speech recognizer; (2) events distorting lexical content or the way lexical baseforms are defined; (3) nongrammatical structures that might lead to misunderstandings or ambiguities; and (4) words or phrases used as editing signals in self corrections.

3.1
Acoustic events

3.1.1
Filled pauses

These elements of spoken language, acoustically realized as either lengthened vowels or nasalizations, play an important role as communication resources, and are classified by some authors as conventional words with specific functions and meanings (Clark & Fox-Tree, 2002). Filled pauses require the definition of specific acoustic models, since they could be easily confused with short words. Various papers have shown that filled pauses may help in the recognition and understanding of spontaneous speech (Liu, Nguyen, Matsoukas, Davenport, Kubala, & Schwartz, 1998; Rose & Riccardi, 1999; Shriberg & Stolcke, 1996; Wu & Yan, 2001). From the point of view of speech production, filled pauses are used to hold the turn while the speaker is deciding what to say next. This may also happen in the context of a speech repair: The speaker realizes that he has committed an error, so he interrupts himself before giving the correction, and signals with a filled pause that something has gone wrong. Unlike English, the most common realization found in Spanish sounds like the vowel /e/; second comes a sort of nasalization sounding like /m/, which is close to the English *um*; and last, a sound like the vowel /a/, which is close to the English *uh*. A sizeable number of distorted or phonetically unidentified filled pauses, due to glottalization, laryngealization or misarticulation, can be also found.

3.1.2
Lengthenings

Unlike other SSEs, lengthenings, that is, the stretching out of speech segments inside words, also called *prolongations*, have received little attention in literature (Eklund, 2001). Moreover, of the most widely known spontaneous speech databases for English, only ATIS includes a explicit convention (the symbol ':') to mark lengthenings. However, lengthenings appear to play the same role as filled pauses, either to hold the turn or to mark a correction (Eklund, 2001). In fact, in Spanish the case of an /e/ lengthened at the end of a word may be easily confused with a nonlengthened /e/ followed by a filled pause.

Some authors (e.g., Heeman & Allen, 1999; Shriberg, 1994) accept that lengthenings express hesitation, just as silent pauses or filled pauses do, but they consider, from the point of view of a speech recognition system, that lengthenings do not alter the word stream or the dictionary pronunciation, so they are not included in the set of phenomena relevant to acoustic modeling. Here, lengthenings will be treated in the same way as filled pauses: as acoustic events which could mark the presence of a hesitant segment, a linguistic boundary, or nothing at all. Certainly, a lengthening just stays in the same phone for a longer time, situation that can be handled in a suitable way by current hidden Markov models. The key issue is that we want to know whether or not a phone was lengthened, since it may be used as a relevant cue to speech repairs. Though lengthenings do not strictly require specific acoustic models, since they could be modeled by repeated instances of the same phone, we found it more suitable the use of explicit models for them.

3.1.3
Silent pauses

In read and spontaneous speech, silent pauses are used to mark breaking points between two sentences or two semantic units, a feature that may help in identifying smaller units to recognize, thereby making the recognition process easier and more accurate. However, silent pauses also play a key role in the production and understanding of spontaneous speech, sometimes marking hesitant segments, preceded or followed by filled pauses or self-repairs, and sometimes fulfilling more complex communicative functions, for example, the speaker explores the understanding level of the listeners, stopping his/her discourse and giving them the opportunity either to interrupt or to confirm, as an instance of the turn-assignment Rule 2 in (Levelt, 1989, pages 31 – 32). Although silent pauses are taken into account by current acoustic models, they are usually deleted from the recognized sequence of words. However, recognizing and keeping them explicitly in the word string would greatly help the understanding of spontaneous speech.

3.2
Lexical distortions

Spontaneous speech is far more relaxed than read speech, so a high number of pronunciation variants and pronunciation errors due to speaker specific features, high speech rates, and so forth can be found. These events are included in the inventory primarily to avoid errors in training acoustic models, but also to allow the training of pronunciation models based on the variations observed with regard to *canonical* pronunciations. To that end, we define lexical distortions as *those events affecting the construction of word baseforms that result in noncanonical pronunciations or cut off words.*

3.2.1
Mispronunciations

Mispronunciations are defined as *nonproperly or noncanonically pronounced words.* They represent either alternative pronunciations or articulatory errors that, in the opinion of the speaker, do not pose a problem of understanding, so he/she leaves

them *uncorrected*. What we call *proper or canonical pronunciation* matches the standard Spanish used in broadcast news in Spain (close to central mainland Spanish), and is used to build the word baseforms in the speech recognizer. However, deviations from that pronunciation are allowed (e.g., contextual variations). We define mispronunciation as a pronunciation that involves a phone deletion or substitution with regard to the canonical pronunciation (e.g., 'Madrí' instead of 'Madrid', or 'pasiensia' instead of 'paciencia'). Certainly, mispronunciations are not specific to spontaneous speech, but a much smaller number are usually found in read speech. By annotating both the canonical and the observed word pronunciations, the performance of the recognizer can be evaluated to check whether it fails to recognize the target word due to a mispronunciation, or it is robust to these phenomena.

3.2.2
Cut off words

Cut off words, also called *word fragments* or *partial words*, appear due to errors, either in planning or in pronunciation, which are immediately repaired, either by the same word repeated correctly or by another word replacing the first one. So cut off words usually lead to self-corrections, also known as *self-repairs*. Both the word fragment and the full orthographic transcription of the intended word must be annotated, preserving the quality of the acoustic models on the one hand and allowing analyses at higher levels on the other.

3.3
Speech repairs

Speech repairs are resources of spontaneous speech which allow speakers either to hold the turn — often in the case of repeats — or to correct their discourse on the fly (Levelt, 1989, ch. 12). In this study we apply the categorization made by Lickley (1994, 1998) with minor changes. In regard to the surface structure we will distinguish between two types of speech repair: (1) *self-repairs*, which include repeats on the one hand, and self corrections or reformulations with substitution, insertion or deletion of words on the other; and (2) *abandoned phrases*, which sometimes correspond to false starts and sometimes to sentence parts left unfinished (also known as *verbal deletions*).

3.3.1
Self-repairs

In discourse structure, repeats often act as turn holders, when the speaker hesitates and decides what to say next. On the other hand, reformulations are the only way to correct errors committed in spoken language. Following Shriberg (1994), we will apply a common structure for repeats and reformulations, consisting of four elements: (1) a segment to be repaired, called *reparandum*; (2) the *interruption point*; (3) the *interregnum*, also called *editing term* or *editing phase*: an optional segment which may include silent and filled pauses and some editing phrases such as *"sorry"* or *"I mean"*; and (4) a segment, called *repair*, giving the replacing material. The annotations will include at most three elements: a reparandum, an (optional) editing phase, and a repair. The interruption point is not annotated, because it is always implicitly located at the offset of the reparandum. The following example, an original transcription

taken from INFOTREN, shows a reformulation with substitution of words. The segment *"en Granada"* (*reparandum*) is replaced by *"en Málaga"* (*repair*), with the editing term *"perdón"*:

Example 1

RM stands for reparandum, IP for interruption point, ET for editing terms and RR for repair.

> hola resido **(RM en Granada) (IP) (ET perdón) (RR en Málaga)** y me gustaría saber si hay trenes para el siete de agosto del dos mil para Granada.
>
> hello I live **(RM in Granada) (IP) (ET sorry) (RR in Málaga)** and I would like to know whether there are trains going to Granada on August the seventh two thousand.

In annotating self-repairs we do not attempt to determine whether the speaker's intention is to hold the turn or to correct an error. We only pay attention to the surface structure. The most important feature in identifying a self-repair is that both the segment said first and its correction fulfill the same syntactic/semantic function — for instance, both are direct objects. Depending on the intonation and/or the syntactic context, the annotator decides whether a phrase is replacing previous materials or is just a syntactically well-formed element which is detailing or extending the meaning of those materials.

Some remarks must be made regarding the annotation of self-repairs:

1. Self-repairs could be nested inside other self-repairs, and nesting could be as deep as necessary to describe the phenomena.
2. Most of the existing annotation schemes for speech repairs mark how words in the repair are related to those in the reparandum (Bear, Dowding, & Shriberg, 1993; Heeman & Allen, 1995; Lickley, 1998). In this study, words are not tagged or indexed, but instead only the type of self-repair is marked. In the case of repeats, reparandum and repair are identical; in the case of insertions, the repair is identical to the reparandum except for one or more words inserted at some point; in the case of deletions, reparandum and repair differ in one or more words deleted in the latter; all other cases are labeled as substitutions.
3. Marks for reparandum, editing phase and repair are allowed to appear — strictly in that order — only within self-repairs, once each, with the editing phase being optional.
4. Self-repairs usually contain other events, particularly lengthenings and cut off words in the reparandum, and silent pauses, filled pauses, speaker noises and discourse markers in the editing phase.

3.3.2 Abandoned phrases

The choice of a specific subcategory for abandoned phrases deserves a short discussion. In spontaneous speech we might find some cases initially identified as self-repairs

where either there is no syntactic correspondence between the reparandum and the hypothesized repair, or they cannot be semantically interchanged even though they play the same syntactic role. For instance, consider the following example:

Example 2

AP stands for abandoned phrase and FP for filled pause.

> no. es que aún no he acabado. **(AP tenía más)** (FP) quería hacerle otra pregunta ¿ es posible?
>
> no. I have not finished yet. **(AP I had more)** (FP) I wanted to ask another question, is it possible?

The segment *"tenía más"* is interrupted and, after a hesitation, a new sentence is started: *"quería hacerle otra pregunta."* This is easily identified as a *false start*. A sentence is left unfinished and a new sentence is constructed in a completely different way. Similar cases can be found not at the beginning but inside sentences. Such cases are harder to identify as false starts, because it is only a sentence part that is erroneously started (see example below).

Example 3

> pues sí. quería saber **(AP a cuánto)** (FP) (RM el) (IP) (RR el) precio del billete de ida y vuelta en coche cama.
>
> so yes. I would like to know **(AP how much)** (FP) (RM the) (IP) (RR the) price of a return ticket by sleeping car.

This time a question acting as direct object *"a cuánto"* is abandoned, but not the introductory segment *"quería saber,"* which is continued after a brief hesitation — including a filled pause and the repeat of *"el"* — by the new object *"el precio del billete de ida y vuelta en coche cama"*. Cases such as this may be handled as words erroneously inserted. In fact, just by deleting them a meaningful sentence results.

It is remarkable that abandoned phrases are commonly followed by hesitations (a filled pause, a repeat, a discourse marker, etc.), used to hold the turn while the speaker composes either the continuation or the new sentence. In this study, these hesitations are not included in the abandoned phrase, except for lengthenings, which cannot be separated from the last word in the abandoned phrase. In previous studies the abandoned phrase was taken as the reparandum, the hesitations were grouped into the editing term and the repair was empty (Heeman, 1997; Shriberg, 1994). Here we annnotate only the abandoned phrase; the hesitations, though annotated as such, are not grouped into an editing term. We chose this approach because, by definition, we understand that self-repairs require two key elements (reparandum and repair), and there is no repair in abandoned phrases. So, cases with no repair are set aside and given a special category. Finally, note that abandoned phrases sometimes include hesitations and self-repairs.

3.4
Discourse markers

Discourse markers (hereafter 'DMs'), also known as *lexical fillers*, are very common words or phrases that speakers use primarily to structure discourse. Though DMs are equally useful in written and spoken language, they occur primarily in spontaneous speech. In fact, including DMs in conversation provides useful clues for the hearer to understand how the speaker's utterances fit with the conversation. However, DMs are optional parts of speech and neither affect the meaning of the utterance nor operate syntactically in it. They show relationships between utterances or between the speaker and the utterance, and can perform a variety of functions: taking or holding the turn, beginning a new topic, concluding or summarizing, contrasting points of view, emphasizing, establishing a sequence of reasoning, referring to previous contents, accepting or acknowledging, and so forth (see Schourup, 1999). So, we look at DMs as pragmatic elements of spoken language, that is, events at the pragmatic level. Some of them may be useful in detecting speech repairs, thus improving the recognition and understanding of spontaneous speech, as shown in previous studies (Heeman & Allen, 1999).

Some expressions act only as DMs, but in most cases words that act at one point as DMs can be found playing different roles elsewhere. Also, the same word can have different functions as a DM. In any event, words acting as DMs are easily detected because they are always the same and happen in very specific contexts. On the other hand, as noted in EARS-MDE (2004, page 8):

> it is nearly impossible to establish an exhaustive list of DMs, due to their wide range of functions and the difficulty of defining them precisely. Moreover, DMs are subject to much dialectal and individual variation, and novel expressions can serve as DMs, which means that any list quickly becomes out of date.

In this work we will consider only two types of DMs closely related to marking speech repairs:

explain/edit: *perdón* (sorry), *quiero decir* (I mean), etc.

fill/pause: *bueno* (well), *mire* (look), etc.

4 Spontaneous speech events: Analysis and discussion

4.1
The markup

A specific task-focused XML-based markup scheme is designed as a sort of convenient and simplified choice to annotate disfluencies and other events in human-machine and human-human dialogs in Spanish. The scheme is based on previously defined markup schemes for spontaneous speech databases in English (Heeman, 1997; Lickley, 1998; Shriberg, 1994). As an XML application, the markup scheme may grow and generalize, including more features to fit other databases or more complex tasks for the same database. In fact, the scheme includes more events than those reflected in this study. Here, only those events relevant to modeling and detecting disfluencies have been considered. A simplified format is also defined to speed up the annotation process. The simplified format is conceived as an intermediate and more readable version of

the XML counterpart, which is automatically generated. The markup task involves listening to the speech signals, revising the existing marks and adding completely new marks. Annotation guidelines are written to maintain consistency between annotators (Rodríguez, 2002; Rodríguez, Torres, & Varona, 2000). Also, to help the annotators to detect and correct markup errors, a very simple parser is implemented which accounts for the well-formedness of the annotations. Three annotators are used in the case of INFOTREN, and two in the case of CORLEC-EHU. To further increase the consistency of the annotations, a single expert (the first author of this paper) has reviewed them all.

4.2
Absolute and relative frequencies of events

Table 2 presents the statistics of spontaneous speech events for INFOTREN and CORLEC-EHU, including the absolute counts and the average number of events per word. The length of utterances is measured in terms of efficient words, as defined by Shriberg (1994), that is, excluding words in reparanda and editing terms of self-repairs. Nor are words in abandoned phrases, filled pauses and the remaining acoustic events counted as efficient words.

Table 2

Statistics of spontaneous speech events for INFOTREN and CORLEC-EHU: Absolute counts (#E) and average number of events per 100 efficient words (%E/eW)

			INFOTREN		CORLEC-EHU	
Category/Subcategory		Type	#E	%E/eW	#E	%E/eW
Acoustic events	Silent pauses	-	753	4.21	1863	2.67
	Filled pauses	a	93	0.52	33	0.05
		e	546	3.05	794	1.14
		m	179	1.00	335	0.48
		trash	210	1.17	642	0.92
	Lengthenings	-	1019	5.70	3593	5.15
Lexical distortions		cut off	95	0.53	222	0.32
		mispronounced	105	0.59	968	1.39
Speech repairs	Self repairs	repetition	292	1.63	1657	2.38
		substitution	141	0.79	337	0.48
		insertion	37	0.21	94	0.13
		deletion	5	0.03	16	0.02
	Abandoned phrases	-	70	0.39	203	0.29
Discourse markers		explain/edit	71	0.40	234	0.34
		fill/pause	225	1.26	1380	1.98

The number of orthographic words in INFOTREN is 18734, whereas the number of efficient words is 17884. The number of annotated events is 3841, which yields an average of 21.46 events per 100 efficient words and 0.4268 events/second. Note that some of these events are nested inside other events, for instance, filled pauses and cutoff words may appear inside speech repairs. CORLEC-EHU contains 72461 orthographic words; 69726 of them are efficient words. The number of events in CORLEC-EHU is 12371, which yields an average of 17.74 events per 100 efficient words and 0.5361 events/second.

To compare these results with previous studies in literature, next we consider only the counts of disfluencies. Following Shriberg (1994), this category consists of non-nested speech repairs and filled pauses. There are 1454 disfluencies in INFOTREN and 3674 disfluencies in CORLEC-EHU, which means 8.13 and 5.27 disfluencies per 100 efficient words (0.1616 and 0.1592 disfluencies/second), respectively. A similar measure was used by Shriberg (1994), called *per-word disfluency rate* (*d*), defined as the probability of a disfluency at any particular word. In the case of ATIS, the estimated value of *d* is less than 1%. In the case of AMEX (a corpus of spontaneous human-human dialogs in English over the telephone in the air travel planning domain) and Switchboard, *d* is around around 6%. This latter value is similar to the rate of disfluencies obtained for CORLEC-EHU, which is also composed of human-human dialogs but in a face-to-face configuration. However, the rate of disfluencies obtained for INFOTREN seems too high when compared with that of ATIS, especially taking into account that in both cases utterances come from human-computer dialogs held in specific (and very similar) domains.

This latter result can be explained in part by the fact that INFOTREN consists of telephone speech, whereas ATIS was recorded in a face-to-face configuration. Oviatt (1995) reports around 1.8 disfluencies per 100 words in a human-computer face-to-face spoken dialog task similar to ATIS; 5.50 disfluencies in human-human task-oriented face-to-face dialogs; and 8.83 in human-human dialogs over telephone lines. Oviatt (1995, pages 31–32) concludes first that "all samples of human-human speech had a substantially higher disfluency rate than the human-computer samples," and second that "average disfluency rates were significantly higher for (human-human) telephone speech than other categories of human-human speech." However, the disfluency rate obtained for INFOTREN does not tally with the first statement, and the disfluency rates for AMEX and Switchboard obtained by Shriberg do not tally with the second, if we compare them with the rate reported by Oviatt for human-human face-to-face dialogs and our own rate for CORLEC-EHU. So, the disfluency rate may depend not only on the dialog type (human-computer vs. human-human) and the channel (face-to-face vs. telephone). Setup factors must be also taken into account.

In particular, the high rate of disfluencies in INFOTREN can be explained by looking at the task more closely. In the case of ATIS, subjects had a push to talk device, so they could plan their speech beforehand. The machine was not waiting for them to respond. Subjects had unlimited time to look at a visual display of information. That display can usually hold much more information than can be conveyed over the phone and remembered at one time. In the case of INFOTREN, there was no push-to-talk device, the system was waiting for a response and the subject was pressed to answer before he/she had planned what to say. Also, subjects could only

obtain information a bit at a time, via the audio. Sometimes they tended to chat with the system instead of just asking for timetables and prices. Scenarios were designed to allow free interaction, but in practice subjects did not plan their questions, but made them up on the fly, asking for many things at the same time, which involved many hesitations and self-corrections. This behavior was especially common in the first turns, before the dialog system started guiding the user through a virtual menu of information items. Probably, these factors must be leading to the high rate of disfluencies in INFOTREN.

Summarizing, natural conversations might show more disfluencies than human-computer dialogs because of a higher degree of spontaneity. Also, telephone speech might show more disfluencies than face-to-face dialogs because of a lack of feedback that needs to be solved somehow. However, the interface design and other factors (closely related to the task) may conceal these general trends.

The statistics obtained for INFOTREN and CORLEC-EHU are studied in more detail in the following paragraphs, category by category.

4.2.1 Acoustic events

The difference in the density of events between INFOTREN and CORLEC-EHU (around 5 points) can be explained almost exclusively by the different densities of acoustic events (see Table 2). There are 15.65 acoustic events per 100 efficient words in INFOTREN, and only 10.41 in CORLEC-EHU. On the other hand, the internal distribution of acoustic events reveals that silent pauses, filled pauses and lengthenings are all important resources of spontaneous speech.

The high number of lengthenings is remarkable, especially in the case of CORLEC-EHU, where their density is twice that of silent and filled pauses (see Table 2). Lengthenings affected mainly vowels (80.27% of the cases), with a clear preference for the vowel /e/, but also some consonants, especially /l/, /n/ and /s/. As shown in Table 3, more lengthenings of the vowel /o/ and consonants /n/ and /s/ are found in CORLEC-EHU than in INFOTREN, whereas there are fewer lengthenings of vowels /a/ and /e/ and consonant /l/. In any case, the set of lengthened phonemes is the same in both corpora, except for the consonant /rr/, which only appears in CORLEC-EHU.

An analysis of the position (initial, intermediate or final) of the lengthenings inside words reveals that in the case of INFOTREN 8.54% are initial, 12.07% intermediate and 67.04% final; in the case of CORLEC-EHU 3.12% are initial, 6.21% intermediate and 80.77% final. So a clear preference for final lengthenings is observed. On the other hand, 12.37% and 9.91% of the cases, respectively, were lengthenings of specific monophonemic words, especially the conjunction '*y*' and the preposition '*a*' (translated as '*and*' and '*to*', respectively).

In fact, analysis reveals that the words most commonly lengthened are conjunctions, prepositions, articles and pronouns. In Spanish most of these words precede nouns, noun syntagms or full predicates, where the problems in planning commonly arise. This gives us important information for predicting lengthenings. Table 4 shows the 20 most common lengthenings found in INFOTREN and CORLEC-EHU, taking into account not only the lengthened sound but also the word and position in which

Table 3

Relative percentages of lengthened sounds for INFOTREN and CORLEC-EHU

Type of sound	*Sound*	*INFOTREN*	*CORLEC-EHU*
Vowel	/i/	11.58	10.10
	/e/	35.72	27.05
	/a/	23.06	19.90
	/o/	9.81	21.04
	/u/	0.10	1.22
Consonant	/l/	10.30	3.17
	/m/	0.29	0.17
	/n/	4.61	8.79
	/rr/	–	0.61
	/s/	4.51	7.96

Table 4

Ranking of the 20 most common lengthenings found in INFOTREN and CORLEC-EHU. Not only the lengthened sound — underlined in these examples — but also the word and position in which it appears are used to define one specific lengthening. Translations to English are provided, where F stands for *Feminine*, M for *Masculine*, S for *Singular* and P for *Plural*

INFOTREN			CORLEC-EHU		
Lengthening	Translation	Count	Lengthening	Translation	Count
de	from	113	que	that	304
el	the (MS)	77	y	and	233
y	and	72	de	of	211
para	for	45	la	the (FS)	106
a	to	38	en	in	81
el	the (MS)	36	no	no	77
que	that	31	a	to	68
qué	what	22	un	a (MS)	66
saber	to know	18	el	the (MS)	58
la	the (FS)	16	pero	but	57
día	day	16	porque	because	55
las	the (FP)	15	o	or	53
del	from the (MS)	15	para	for	52
sobre	over	14	una	a (FS)	51
quería	I wanted	13	es	is	34
o	or	13	pues	so	33
si	if	12	se	*Generic Pronoun*	29
este	this (MS)	12	como	like	29
el	the (MS)	12	yo	I	28
sí	yes	11	sí	yes	25

it appears. For instance, the most common lengthening in INFOTREN is that of the vowel /e/ at the end of the word '*de*' (the lengthened sound appears underlined). The words shown in Table 4 account for 60% the lengthenings in INFOTREN and 46% the lengthenings in CORLEC-EHU. In the ranking of INFOTREN there are only three verbs or nouns related to the task: *saber* (rank 9), *día* (rank 11) and *quería* (rank 15); the remaining words in that ranking (and all the words in the ranking of CORLEC-EHU) are conjunctions, prepositions, articles and pronouns.

These results enforce our intuition that lengthenings play an important role in spontaneous speech, which is supported by other, more detailed and cross-lingual studies (see Eklund, 2004).

4.2.2 Lexical distortions

The rate of lexical distortions in INFOTREN is quite low: 1.12 per 100 efficient words, half of them mispronounced and half cut off words. A higher figure is obtained in CORLEC-EHU: 1.71 per 100 efficient words. Note the high rate of mispronunciations in CORLEC-EHU (1.39) compared to that of INFOTREN (0.59). This may reflect the fact that the more competent the audience is perceived to be, the more relaxed the pronunciation is. In other words, speakers talk to the computer in a more formal way than they do to humans, because they perceive that computers are less competent than humans in recognizing speech. On the other hand, the rate of cut off words is higher in INFOTREN (0.53) than in CORLEC-EHU (0.32). CORLEC-EHU consists of interviews with writers, politicians, journalists, and so forth, who are highly skilled speakers and seldom have problems in lexical access. So in most cases words are cut off due to problems in articulation or to interruptions. By contrast, subjects of INFOTREN are not sure about their own requirements, because they are playing out virtual scenarios. So they often simply forget city names or train types. More than 35% of cut off words in INFOTREN are due to this kind of problem.

4.2.3 Speech repairs

An average of 3.05 speech repairs per 100 efficient words is found in INFOTREN, made up largely of repeats (53.58% of the cases) and substitutions (25.87%). In the case of CORLEC-EHU, 3.30 speech repairs per 100 efficient words are found, 71.82% of them repeats and 14.61% substitutions. The high rate of repeats supports the hypothesis that speakers use them as resources to hold the turn while thinking what to say next — the same as when inserting filled pauses. Substitutions, insertions, and so forth show much lower rates, since they only appear as a result of errors that speakers detect and must fix on the fly.

Sometimes a word is repeated several times. This case is annotated as a sequence of nested repeats, from left to right, following a binary branching analysis algorithm similar to that used in Shriberg (1994). Some authors consider such cases as one single repeat. To clarify this, we show in Table 5 the percentages of simple (non-nested/nonoverlapped) and complex speech repairs found in INFOTREN and CORLEC-EHU. Complex repeats are almost always multiple repeats (without substitutions, insertions nor deletions). Only 9.25% in INFOTREN and 12.07% in

CORLEC-EHU are complex repeats. In fact, complex repairs are not common: around 13% in both cases. If multiple repeats are counted as single repeats, the rate of speech repairs per 100 efficient words in INFOTREN and CORLEC-EHU drops to 2.90 and 3.02, respectively.

Table 5

Percentages of simple (non-nested/nonoverlapped) and complex speech repairs in INFOTREN and CORLEC-EHU

	INFOTREN		*CORLEC-EHU*	
Type	*Simple*	*Complex*	*Simple*	*Complex*
Repeats	90.75	9.25	87.93	12.07
Substitutions	78.72	21.28	79.53	20.47
Insertions	81.08	18.92	86.17	13.83
Deletions	100.00	0.00	87.50	12.50
Abandoned phrases	88.57	11.43	91.13	8.87

4.2.4
Discourse markers

We found 1.66 and 2.32 discourse markers per 100 efficient words in INFOTREN and CORLEC-EHU, respectively. As shown in Table 2, filling words are more common than editing phrases. Though both types of discourse markers were included in the inventory as possible cues of speech repairs, many instances were not related to disfluencies. Also, a few expressions account for most of them. In the case of INFOTREN, the five most common expressions (see Table 6) account for 83.48% and 73.08% of the filling and editing markers, respectively. In the case of CORLEC-EHU, the figures are 73.33% and 81.69%, respectively.

4.3
Dependence on utterance length

We set out to learn how spontaneous speech events are distributed in the set of utterances, whether or not counts of events are only dependent on utterance length. We do not segment utterances into sentential units as done by Shriberg (1994), but rather consider the whole utterance as the reference unit. Utterances range from monosyllabic sentences lasting less than a second to a sequence of several sentences lasting two minutes or more. Five categories are considered: (1) acoustic events (silent pauses, filled pauses and lengthenings), (2) lexical distortions, (3) speech repairs, (4) discourse markers and (5) disfluencies, which include speech repairs and filled pauses. This latter category is defined to allow meaningful comparisons with previous works for English.

First, the events appearing at each utterance are counted, leading to the statistics shown in Table 7. It must be noted that utterances are on average much longer in CORLEC-EHU (24.41 efficient words per utterance) than in INFOTREN (10.73

Table 6

The five most common expressions used as filling and editing markers in INFOTREN and CORLEC-EHU

	INFOTREN		*CORLEC-EHU*	
	Expression (Spanish)	*Translation to English*	*Expression (Spanish)*	*Translation to English*
Fill	pues	so	pues	so
	mire	look	bueno	well
	a ver	let's see	no	no
	bueno	well	entonces	then
	entonces	then	hombre	man
Edit	o sea	that is	es decir	that is (to say)
	es que	it's just that	o sea	that is
	perdón	sorry	bueno	well
	bueno	well	vamos	come on
	es decir	that is (to say)	digamos	let's say

efficient words per utterance). That is why mean values are higher for CORLEC-EHU in most cases. As shown in Table 7, the average number of disfluencies per utterance is 0.95 in INFOTREN and 1.44 in CORLEC-EHU. A relatively high *SD* is found in both cases: 1.63 and 2.45, respectively. In fact, the maximum number of disfluencies found in a single utterance is 18 in INFOTREN and 23 in CORLEC-EHU. This variability in the counts of events per utterance is also remarkably high in the remaining categories, and may be explained to a great extent by the variable length of utterances. Correlation coefficients are computed between the counts of events and the length of utterances, measured in terms of efficient words, and are also shown in Table 7.

Table 7

Mean ($\bar{x}$), *SD* (σ) and maximum figures (x_{max}) of the count of spontaneous speech events per utterance in INFOTREN and CORLEC-EHU. The correlation coefficients (ρ) of the counts with regard to the length of utterances are also shown

	INFOTREN				*CORLEC-EHU*			
Category	$\bar{x}$	σ	x_{max}	ρ	$\bar{x}$	σ	x_{max}	ρ
Acoustic Events	1.69	2.52	32	0.6876	2.54	4.00	31	0.7930
Lexical Distortions	0.12	0.40	3	0.2433	0.52	1.22	16	0.3989
Speech Repairs	0.33	0.83	9	0.4609	0.81	1.53	21	0.5265
Discourse Markers	0.52	0.93	8	0.4228	1.03	1.34	10	0.4431
Disfluencies	0.95	1.63	18	0.5970	1.44	2.45	23	0.6667

Correlations are higher for CORLEC-EHU, maybe due to a wider distribution of utterance lengths and to the availability of more data. Some specific features related to the task may also explain lower correlations in the case of INFOTREN: For instance, speakers tend to insert filled pauses at the beginning of all their utterances, regardless their length. In any case, acoustic events show the highest correlations. We explain this through the fact that pauses and lengthenings are common resources of spoken language that most users tend to insert regularly.

With regard to speech repairs, two phenomena are combined, with opposite effects on the correlations. On the one hand repeats, which account for 50–70% of speech repairs, are common resources of spoken language, as are pauses and lengthenings. They are therefore likely to appear regularly, though this also depends on the speaker's habits. On the other hand, reformulations and abandoned phrases appear as a result of errors. So although the probability of errors increases with length, they may appear in both short and long utterances. Also, as we will show below, there are hesitant speakers who show an almost uniform distribution of reformulations, and fluent speakers who are rarely affected by errors. Disfluencies, which join together filled pauses and speech repairs, consistently show intermediate correlations.

Discourse markers appear in two main contexts, again with opposite effects on the correlations. On the one hand, filling expressions — here we include connectors such as *y* (*and*), *pero* (*but*) and *entonces* (*then*) — tend to appear regularly, especially in the case of low skilled speakers. On the other hand, explaining/editing expressions appear in both fluent and disfluent contexts, depending on the speaker habits, but are not inserted regularly. It is also observed that only a few speech repairs show explicit editing expressions. This leads to lower correlations than those obtained for acoustic events.

Finally, the lowest correlation coefficients are obtained for lexical distortions (cut off words and mispronunciations). We think this is mainly related to their low frequency. These phenomena, especially cut off words, arise only as errors in both short and long utterances, though they are more likely in long utterances. In the case of CORLEC-EHU, more data are available and a higher correlation coefficient is obtained.

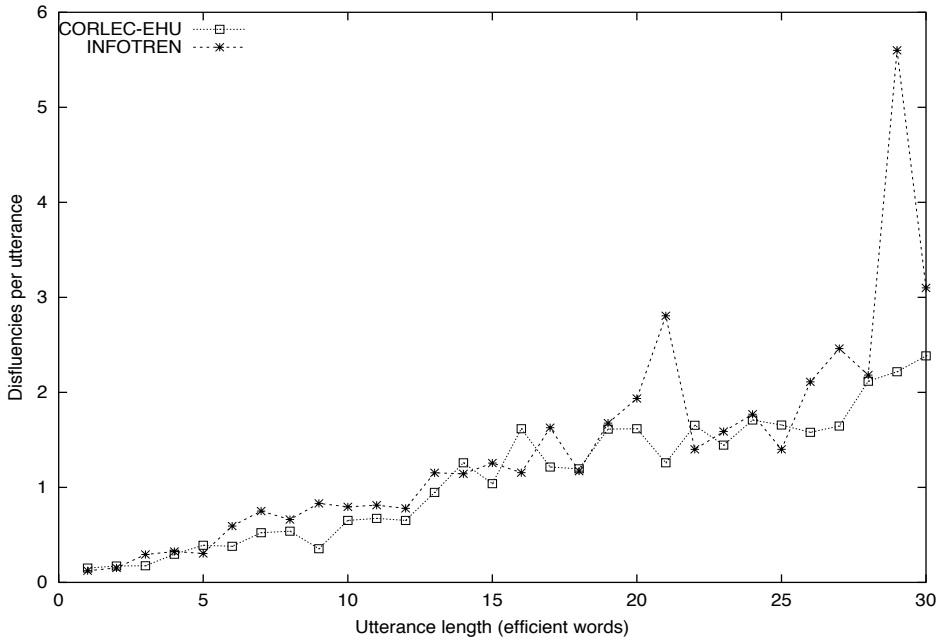
4.3.1 Comparisons with previous studies

In previous paragraphs we have shown that a positive, relatively high correlation exists between the number of disfluencies and the length of the utterance, and found that 35–45% of the variance in the number of disfluencies is accounted for by the variance in the length of the utterance. In fact, as shown in Figure 1, the number of disfluencies seems to increase linearly with utterance length. This does not mean that there is a causal relationship between the two quantities, since a third independent variable may be involved. However, previous studies for English have sought to define a parametric model describing the rate of disfluencies as a function of utterance length.

Oviatt (1995) suggested that "the disfluency rate would rise in longer utterances, since their production theoretically requires an increase in both macroplanning and microplanning." Note that it is the *rate of disfluencies*, not the *absolute number of disfluencies*, which Oviatt claims, would increase with the length of the utterance. In

Figure 1

The average number of disfluencies per utterance as a function of utterance length



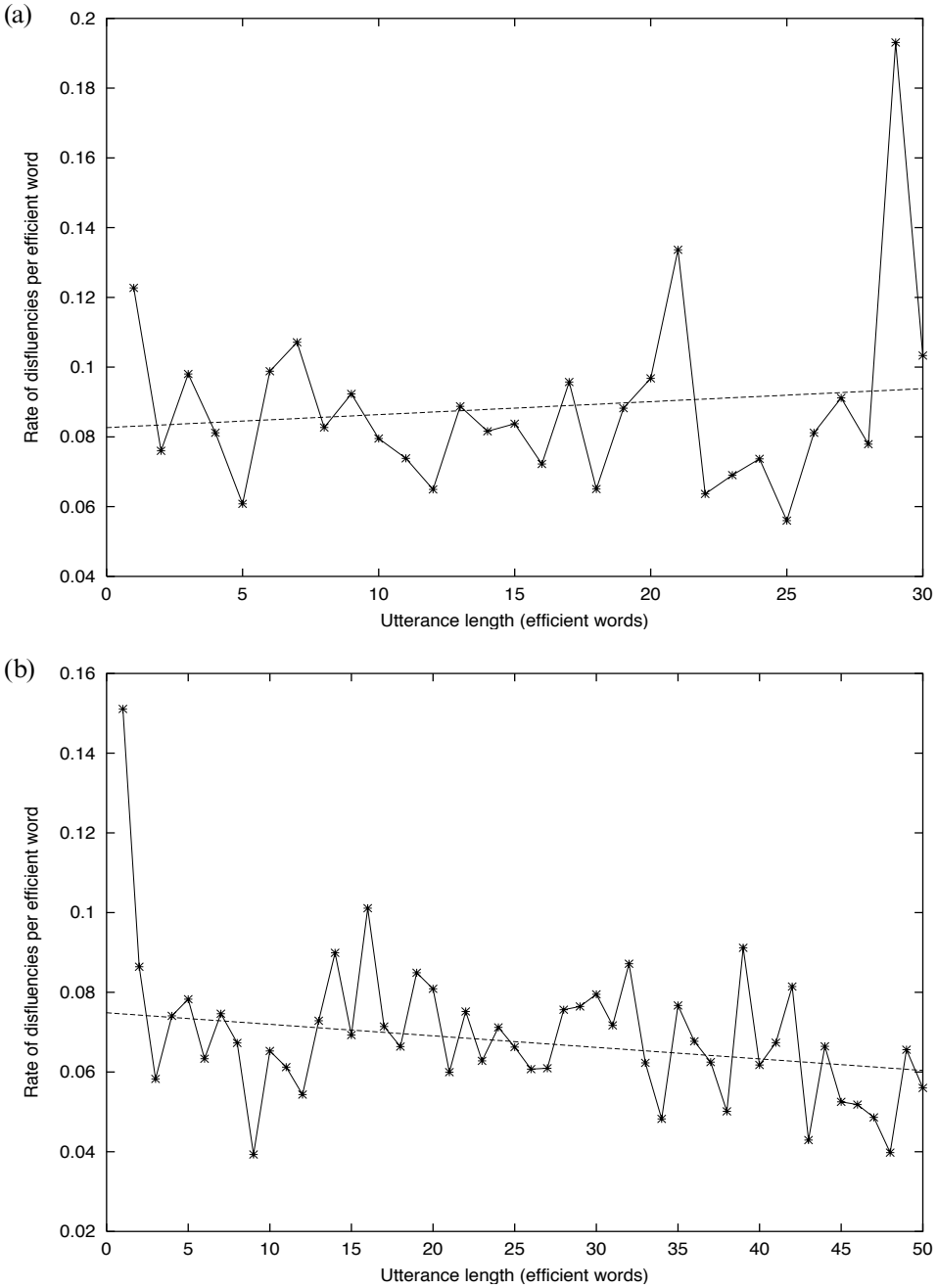
fact, she finds that the average rate of disfluencies per 100 words in unconstrained task-oriented human-computer spoken dialogs seems to depend linearly on utterance length, with a sizeable correlation of 0.8775, which means that 77% of the variance in the rate of disfluencies can be predicted simply by knowing the utterance length. The rate increases from around one disfluency per 100 words for utterances of length five to around four disfluencies per 100 words for utterances of length 15 (a 400% relative increase in a range of size 10). Oviatt reports 1.74 disfluencies per 100 words on average.

Figure 2 shows the average rate of disfluencies obtained for each utterance length for INFOTREN (17884 efficient words, 1454 disfluencies) and CORLEC-EHU (69726 efficient words, 3674 disfluencies).

In the case of INFOTREN a very low positive correlation is found. In the range of lengths [1,30] it would be $\rho = -0.1239$, which means that only 1.54% of the variance in the rate of disfluencies may be explained by the variance in the length of the utterance. The mean rate of disfluencies per word in that range is 0.0884 (± 0.0096), with a minimum rate of 0.0560 and a maximum rate of 0.1931. Regression analysis yields a slightly positive slope: $y = 0.0004\ x + 0.0827$, but not significant ($t = 0.6608$, $df = 28$, $p > .5$). If this relation holds, the rate would increase from 8.3 disfluencies per 100 efficient words for utterances of length 1 to 9.3 for utterances of length 30. This means a 12% relative increase in a range of size 30, far from the 400% relative increase reported by Oviatt. For lengths higher than 30, rates of disfluencies are not so reliably estimated, due to a lack of samples (the same can be stated for lengths

Figure 2

The rate of disfluencies per efficient word as a function of utterance length for INFOTREN (a) and CORLEC-EHU (b). A linear fit of the data is shown, which yields a slight positive slope for INFOTREN: $y = 0.0004\ x + 0.0827$ ($t = .6608$, $df = 28$, $p > .5$), and a slight negative slope for CORLEC-EHU: $y = -0.0005\ x + 0.0810$ ($t = 2.8772$, $df = 48$, $p < .006$)



higher than 50 in the case of CORLEC-EHU). Note that, in fact, the rate for length 29 appears to be an outlier. If that point was not considered here, a lower correlation and a flatter slope would be obtained, supporting the argument for a roughly constant rate of disfluencies.

In the case of CORLEC-EHU a negative correlation is found. In the range of lengths [1,50] it would be $\rho = -0.3835$, which means that 14.71% of the variance in the rate of disfluencies may be inversely explained by the variance in the length of the utterance. The mean rate of disfluencies per word in that range is 0.0690 (± 0.0050). Regression analysis yields a slightly negative slope: $y = -0.0005x + 0.0810$ ($t = 2.8772$, $df = 48$, $p < .006$). The high rate obtained for length 1 can be explained by the large number of simple repeats used as backchannels, such as *"Sí, sí"* (*Yes, yes*), or *"Claro, claro"* (*Sure, sure*), which are counted as self repairs. Once again, if that point was not considered here, a lower correlation and a flatter slope would be obtained.

In summary, we can conclude that the rate of disfluencies, though noisy, does not change significantly with utterance length, but instead appears to be fairly flat. Similar results are reported by Shriberg (1994) for two databases of human-human dialogs (AMEX and Switchboard). Also, though a statistically significant positive correlation is reported for ATIS, the slope of the linear fit is quite small and, as noted by Shriberg (1994, page 98), *"the average number of disfluencies in a sentence grows roughly linearly with sentence length also for ATIS."*

In Shriberg (1994) the probability of a sentence being fluent (having no disfluencies) is found to be well fit by an exponential decay function of utterance length, with a fractional base parameter (b) called *per-word fluency rate*, which may be different for each database and depends mainly on the type of interaction (human-human or human-computer). Figure 3 shows the distribution of fluent and disfluent utterances by utterance length in the range [0,40] for INFOTREN and CORLEC-EHU. In both cases, the number of fluent utterances seems to decrease exponentially. In the case of INFOTREN the number of disfluent utterances does not change substantially in the range [1,20] but then it decreases sharply, mainly due to a lack of samples. In the case of CORLEC-EHU the number of disfluent utterances decreases smoothly in the range [1,40].

Like Shriberg (1994), we compute the probability of a fluent utterance of length n as the number of fluent utterances of length n divided by the total number of utterances of length n. Figure 4 shows the log-probability of a fluent utterance as a function of utterance length. The data seem to fit the two-parameter model proposed by Shriberg: $P_{fluent}(W) = C\,b^W$, so that the log-probability of a fluent utterance would depend linearly on utterance length. In the case of INFOTREN the correlation coefficient in the range [1, 20] is $\rho = -0.9542$ and the linear fit yields $ln(P_{\text{fluent}}(W)) = -0.0628 \times W - 0.0330$. We get a fluency rate of $b = 0.9391 \pm 0.0091$ and $C = 0.9675 \pm 0.1000$. In the case of CORLEC-EHU the correlation coefficient in the range [1,20] is $\rho = -0.9578$ and the linear fit yields $ln(P_{\text{fluent}}(W)) = -0.0544 \times W - 0.0009$, which gives $b = 0.9471 \pm 0.0076$ and $C = 0.9991 \pm 0.085$.

Note that $C \approx 1$ in both cases, which, as noted by Shriberg (1994), suggests a lack of factors affecting all utterances to the same degree, regardless of their length. Note also that the trends for INFOTREN and CORLEC-EHU are quite close. In

Figure 3

Distribution of fluent and disfluent utterances by utterance length for INFOTREN (a) and CORLEC-EHU (b)

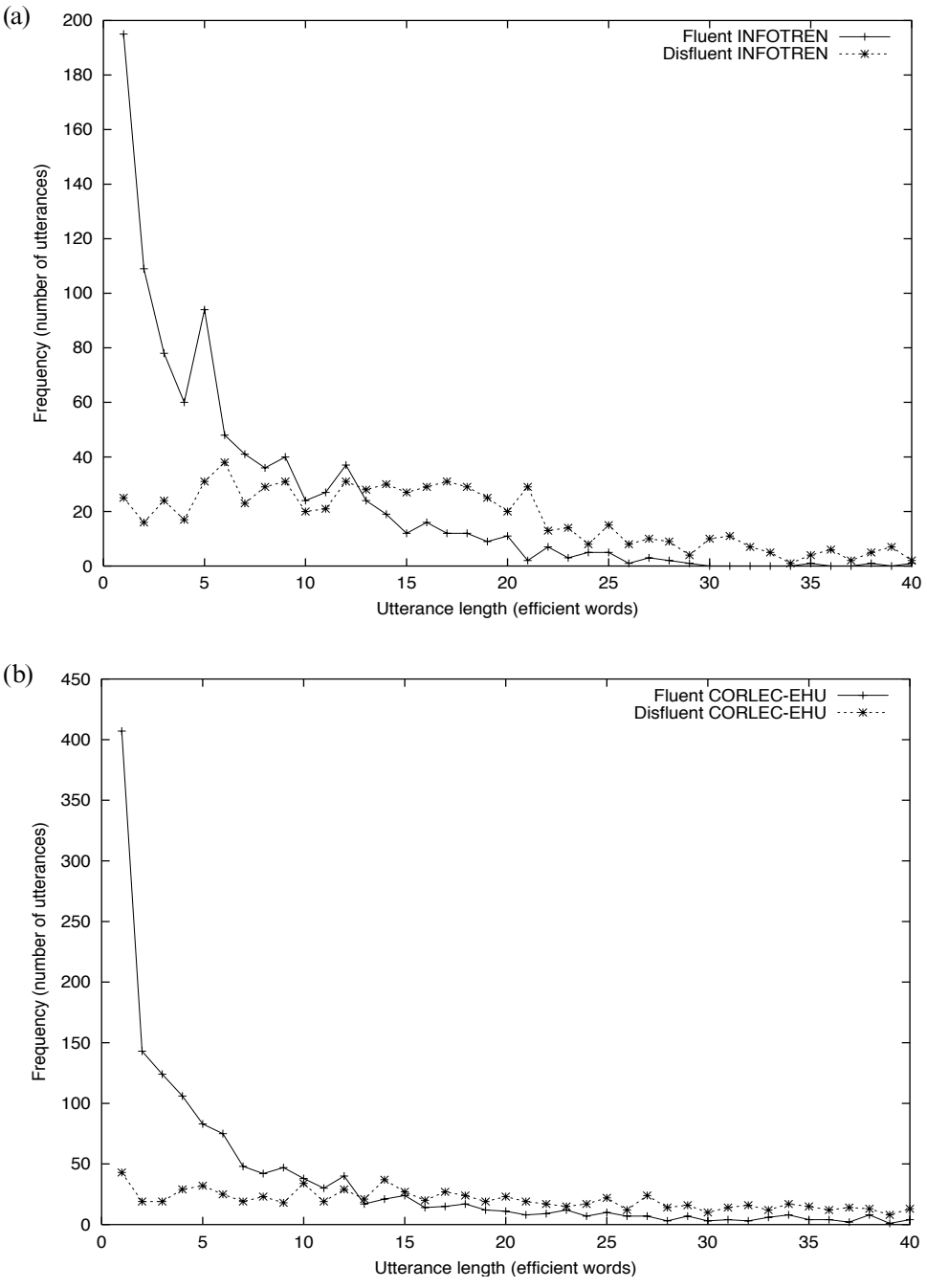
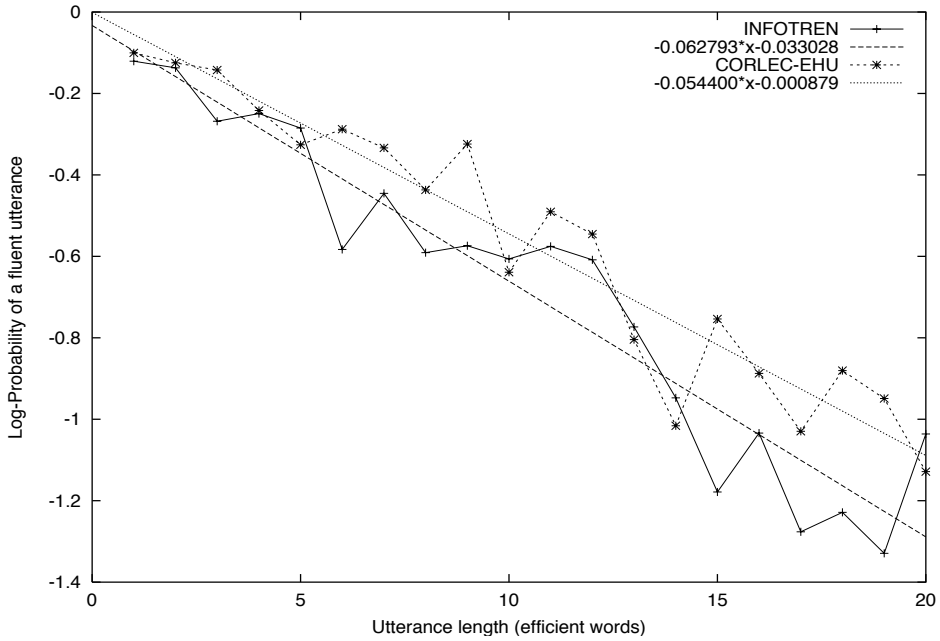


Figure 4

The log-probability of a fluent utterance (P_{fluent}) as a function of utterance length (W), fitted to a linear model: $ln(P_{fluent}(W)) = ln(b) * W + ln(C)$



fact, the differences may reflect random variabilities, the slopes being statistically identical. To solve this uncertainty, first the two-parameter model is approximated by assuming that $C = 1$, which results in a one-parameter model: $P_{fluent}(W) = b^w$. Then, by fitting the data to this model, we get that the difference between the slopes for INFOTREN and CORLEC-EHU exceeds the 95% confidence limits (see Table 8). So we conclude that the slope for INFOTREN is significantly higher than that obtained for CORLEC-EHU. Starting from the one-parameter model, the per-word disfluency rate is $1-b = 0.0631$ for INFOTREN, which is remarkably higher than the figure reported by Shriberg (1994) for ATIS (less than 0.01). In the case of CORLEC-EHU, $1-b = 0.0530$ which is close to the figure found by Shriberg (1994) for AMEX and Switchboard (around 0.055).

Table 8

Linear fit of the log-probability of a fluent utterance for the one-parameter exponential model

	Degrees of freedom	*Slope* (ln b)	*95% confidence limits for the slope*
INFOTREN	19	−0.0652	±0.0046
CORLEC-EHU	19	−0.0545	±0.0038

So far, we have implicitly or explicitly given three measures of the rate of disfluencies per efficient word: (1) the *overall rate*, computed as the number of disfluencies divided by the number of efficient words in the database; (2) the *average rate across utterance length*, computed as the average of the mean rate of disfluencies for each utterance length (Fig. 2); and (3) the *per-word rate*, $1-b$, based on an exponential model for the probability of a fluent utterance as a function of utterance length. The values obtained for INFOTREN and CORLEC-EHU are summarized in Table 9.

Table 9

Measures of the rate of disfluencies per efficient word

	Overall	*Average across utterance length*	*1-b, based on exponential model* $P_{fluent}(W) = b^W$
INFOTREN	0.0813	0.0884	0.0631
CORLEC-EHU	0.0527	0.0690	0.0530

The overall rate given in the first column must be taken as the reference figure in comparing the density of disfluencies in the two databases considered in this work. The two other figures are estimates of the rate of disfluencies as a function of utterance length, and may differ for two databases showing the same overall rate but different distributions of disfluencies across and within utterances. The average rate of disfluencies across utterance length is higher than the model-based per-word rate $1-b$ for both INFOTREN and CORLEC-EHU. As noted by Shriberg (1994), if disfluencies occurred independently, the average rate across utterance length could be directly determined from $1-b$.

What our data reveal is that disfluencies co-occur in the same utterance at a rate higher than that predicted by the binomial distribution. This finding is consistent with the results shown in (Shriberg, 1994). This implies that the exponential model oversimplifies the relation between the rate of disfluencies and the utterance length, because it assumes that disfluencies are independent each other. Nevertheless, the rate of disfluencies in INFOTREN seems to be consistently higher than that of CORLEC-EHU. Probably these differences are not attributable to the type of dialog (human-computer vs. human-human), at least taking into account other results in literature. Whereas CORLEC-EHU yields values near to those of similar databases in English, INFOTREN shows much more spontaneous speech events than other databases of human-computer dialogs (e.g., ATIS). As noted above, this may be due to the dialog interface design and other factors closely related to the process of speech data collection.

4.4
Dependence on the speaker

In this section we briefly explore how spontaneous speech events are distributed in the set of speakers. Situational and demographic factors (age, task role, difficulty of

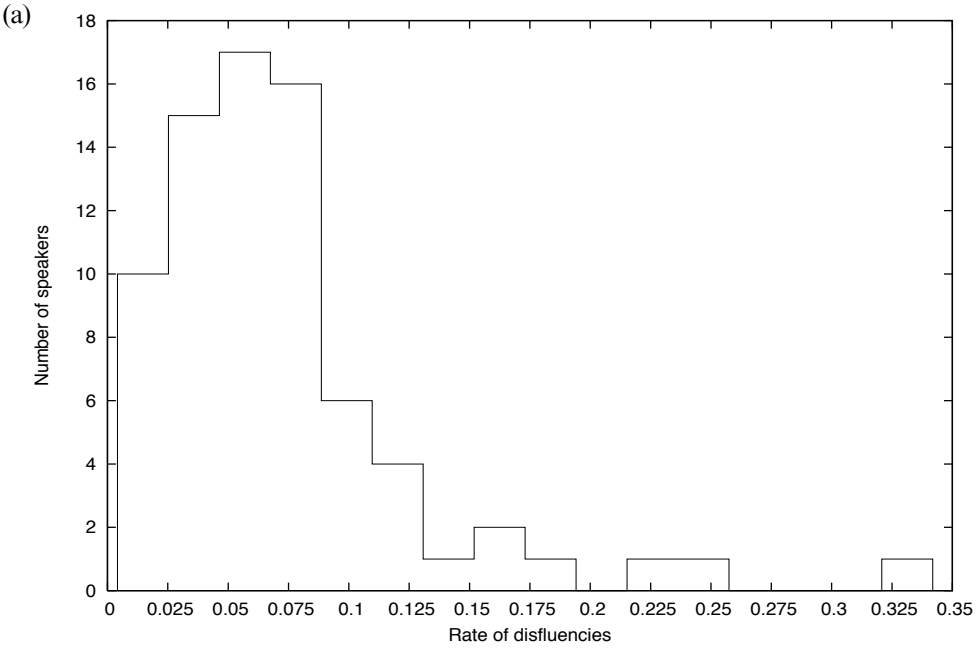
topic domain, gender, etc.) that could affect the rate of disfluencies of each particular speaker are not taken into account. Though those factors could explain to a great extent the different disfluency rates observed across speakers in dialog tasks (see Bortfeld, Leon, Bloom, Schober, & Brennan, 2001), we raise a different question: Is that variability high enough to support the need for speaker adaptation strategies in speech recognition?

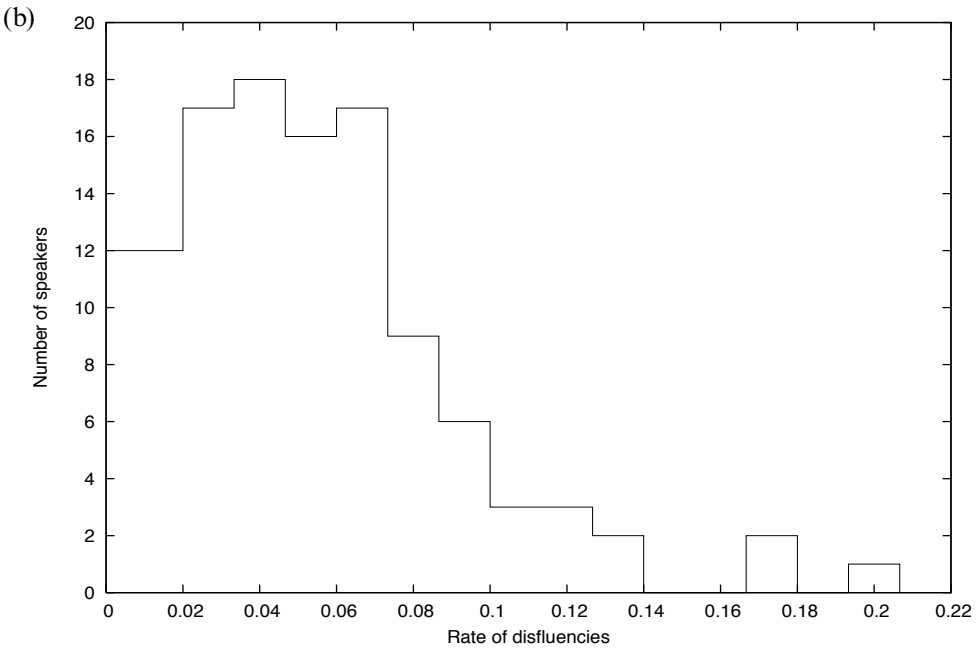
Data show a high variability in both databases, with some speakers showing very high counts and others almost zero counts. Such high variability depends primarily on the amount of data available for each speaker. This explains to a great extent why the average number of events per speaker in CORLEC-EHU (590.39) is higher than in INFOTREN (238.45). In the case of INFOTREN, more than 50% of the variance in the number of disfluencies per speaker can be explained by the variance in the number of efficient words. The percentage is even higher for CORLEC-EHU (70%).

However, regardless the amount of data available for each speaker, the rate of disfluencies still shows a sizeable variability. The distribution of speakers regarding the rate of disfluencies has the typical shape of a high population around the mean (roughly 0.08 for INFOTREN and 0.06 for CORLEC-EHU), a sizeable number of speakers being extremely fluent (those included in the first bin) and a long tail of extremely disfluent speakers (see Fig.5). This prevents us about the existence of different behaviors in the set of speakers, and supports the argument for the usefulness of adaptation strategies in a speech recognizer.

Figure 5

Histogram of speakers regarding the rate of disfluencies for INFOTREN (a) and CORLEC-EHU (b)





4.5
Analysis of various potential cues of speech repairs

Different cues of speech repairs have been explored and tested in the past, in most cases using databases in English: glottalizations, cut off words (Nakatani & Hirschberg, 1993; Shriberg, Bear, & Dowding, 1992), lengthenings (Duez, 1993), silent pauses (Nakatani & Hirschberg, 1993; O'Shaughnessy, 1993; Shriberg et al., 1992), filled pauses (Shriberg & Lickley, 1993), discourse markers (Heeman & Allen, 1999; Shriberg et al., 1992), and so forth. However, most of the above mentioned cues are not reliable, either because they are not specific to speech repairs (as in the case of silent pauses) or because their frequency is not high enough for them to be seen as strong cues (as in the case of filled pauses and discourse markers). Lickley (1994, page 43) concludes that there are:

> several possible avenues in the search for acoustic and prosodic cues for disfluency, but few hard and fast rules and no universal marker, (…) so, rather than a single signal applying to all disfluencies, it is likely that several different signals may alternate or combine as cues.

In this section, we use the information sources currently available, that is, the *annotated events*, to detect speech repairs. Additionally, we aim to check whether or not the conclusions previously obtained for detecting speech repairs in English also apply in Spanish. The set of cues considered in this work consists of acoustic events (silent pauses, filled pauses and lengthenings), lexical distortions (mispronunciations and cut off words) and discourse markers. A study was carried out by counting the occurrences of the proposed cues in speech repairs and obtaining the percentage of speech repairs covered by them, which gives an indication of their potential predictive

power. As shown in Table 10, no single signal can be reliably used to detect speech repairs, since either the frequency or the coverage (or both) are too low. In the following paragraphs data are broken down and briefly analyzed.

Table 10

Percentage of events found inside speech repairs and percentage of speech repairs covered by events

	INFOTREN		*CORLEC-EHU*	
	% events inside speech repairs	*% speech repairs containing events*	*% events inside speech repairs*	*% speech repairs containing events*
Acoustic events	16.32	63.49	22.20	56.18
Lexical distortions	45.00	16.51	22.18	10.62
Discourse markers	10.81	5.87	7.19	4.51

4.5.1 Acoustic events and speech repairs

In the case of INFOTREN we find that 63.49% of speech repairs contain acoustic events; on the other hand, only 16.32% of acoustic events happen inside speech repairs. Similar figures (56.18% and 22.20%, respectively) are found for CORLEC-EHU. The detailed distribution of acoustic events inside speech repairs is shown in Table 11. Note that acoustic events may happen, for instance, in the reparandum of a self-repair which is nested in the repair component of a second self-repair. Such events would be counted twice, so the counts in Table 11 do not correspond to those given above. However, still they are useful as a relative measure of the kind of acoustic events and the position where those events appear inside speech repairs.

Table 11

Acoustic events found inside speech repairs, by repair region

	INFOTREN			*CORLEC-EHU*		
	Reparandum	*Editing phase*	*Repair*	*Reparandum*	*Editing phase*	*Repair*
Silent pauses	14	64	10	19	245	17
Filled pauses	7	91	4	17	226	20
Lengthenings	289	0	47	1032	5	187

When silent or filled pauses appear inside self-repairs, they are found almost exclusively at the editing phase. On the other hand, the frequency of lengthenings inside speech repairs is much higher than that shown in Table 2, suggesting that lengthenings are one of the most reliable cues of speech repairs. In fact, around 50% of the speech repairs we study contain lengthenings—mostly at the end of the last word in the reparandum. But around 70% of lengthenings happen outside speech

repairs. Therefore, though acoustic events (especially lengthenings, but also pauses) often mark the presence of speech repairs, they would also cause false alarms.

4.5.2 Lexical distortions and speech repairs

In the case of INFOTREN, 45% of lexical distortions are found inside speech repairs, and only 16.51% of speech repairs contain lexical distortions. In the case of CORLEC-EHU, even lower figures are obtained: 22.18% and 10.62%, respectively. However, a more detailed study reveals that most of the lexical distortions appearing inside speech repairs are cut off words, whereas most of those appearing outside are mispronounced words. In summary, around 80% of cut off words appear inside speech repairs, almost always at the end of the reparandum (see Table 12). So each time a cut off word is detected, a reformulation is highly likely. However, this finding is not useful in practice, since current speech recognizers cannot easily detect word fragments. Moreover, only between 10% and 15% of speech repairs would be covered by cut off words, at least taking into account the numbers obtained for INFOTREN and CORLEC-EHU.

Table 12

Lexical distortions found inside speech repairs, by repair region

	INFOTREN			***CORLEC-EHU***		
	Reparandum	*Editing phase*	*Repair*	*Reparandum*	*Editing phase*	*Repair*
Mispronunciations	13	0	2	53	2	32
Cut off words	81	0	6	185	0	17

4.5.3 Discourse markers and speech repairs

When describing the structure of self-repairs, we noted that discourse markers might often appear at the editing phase. However, few speech repairs are found to be marked by editing expressions: less than 6% in INFOTREN and less than 5% in CORLEC-EHU. Also, few discourse markers occur inside speech repairs (10.81% in INFOTREN and 7.19% in CORLEC-EHU). Moreover, some of them are repeats, where discourse markers do not act as editing terms, but rather as the words being repeated (see Table 13). So, just like for acoustic and lexical cues, we conclude that, though discourse markers may help to detect speech repairs, they need to be combined with other cues.

Table 13

Discourse markers found inside speech repairs, by repair region

	INFOTREN			***CORLEC-EHU***		
	Reparandum	*Editing phase*	*Repair*	*Reparandum*	*Editing phase*	*Repair*
Discourse markers	4	32	4	28	77	22

5 Concluding remarks

In this paper, we follow a data driven approach to explore the potential benefit of modeling disfluencies and other spontaneous speech events in a speech recognizer, rather than treating them as noise. Our approach is similar to Shriberg (1994), but deals with conversational Spanish instead of conversational English. To our knowledge, this is the first comprehensive study of disfluencies in Spanish. Two corpora of spontaneous speech in Spanish are studied: INFOTREN (human-computer dialogs across telephone lines) and CORLEC-EHU (human-human face-to-face conversations).

First, we introduce the concept of spontaneous speech event, in contrast to the more conventional of disfluency. An inventory of spontaneous speech events is defined, including acoustic events (silent pauses, filled pauses and lengthenings), lexical distortions (mispronunciations and cut off words), speech repairs and discourse markers. Each event is described and illustrated with examples, and the annotation format and procedure are briefly outlined. Finally, frequencies of events are presented and analyzed.

Frequencies are expressed in terms of number of events per 100 efficient words. In the case of INFOTREN, 4.21 silent pauses, 5.74 filled pauses, 5.70 lengthenings, 1.12 lexical distortions, 3.05 speech repairs and 1.66 discourse markers per 100 efficient words are found on average. Considering only acoustic events and speech repairs, INFOTREN shows a rate of 18.7 events per 100 efficient words. Following Shriberg (1994), speech repairs and filled pauses are both counted as disfluencies. The overall probability of disfluency at each word is 0.0813, which is more than eight times the figure reported by Shriberg for ATIS. This may be explained in part by the fact that the dialogs in INFOTREN were held on the telephone, since, as noted by Oviatt (1995), dialogs on the phone involve more disfluencies than face-to-face dialogs such as those of ATIS. But there is a stronger argument if we take into account the setups of ATIS and INFOTREN, which were extremely different from the point of view of subjects using the dialog system. In particular, INFOTREN had no push-to-talk device and no information display; instead, subjects could only obtain information a bit at a time, via the audio, and the system was waiting for a response, so that subjects were pressed to answer before they had planned what to say.

In the case of CORLEC-EHU, 2.67 silent pauses, 2.59 filled pauses, 5.15 lengthenings 1.71 lexical distortions, 3.30 speech repairs and 2.32 discourse markers per 100 efficient words are found on average. The overall probability of disfluency at each word is 0.0527, which is similar to the figure reported by Shriberg for Switchboard, though this latter consists of human-human dialogs on the telephone. A lower rate of disfluencies is found in CORLEC-EHU (human-human dialogs) than in INFOTREN (human-computer dialogs), which seems contradictory when compared to previous results by Shriberg (1994) and Oviatt (1995). As noted above, this can be explained to a great extent by setup factors.

In any case, the high frequencies of events and the distinct acoustic features of some of them suggest the use of specific acoustic models. Also, the regularities shown by some of these events should be taken into account in the language model of a speech recognizer to improve the performance of recognition and understanding.

The extent to which the number of events depends on utterance length is also explored. Definitions and measures used in previous works of Oviatt and Shriberg have been applied to allow meaningful comparisons between their results and ours. According to the results obtained by Shriberg for AMEX and Switchboard, we find that the rate of disfluencies, though noisy, is fairly constant and independent on the utterance length, for both INFOTREN and CORLEC-EHU. Also, confirming a previous result by Shriberg, in both cases the probability of a fluent utterance seems to decay exponentially with utterance length.

On the other hand, the counts of disfluencies show a high variability in the set of speakers, depending primarily on the amount of data available for each speaker. The empirical distribution of speakers with regard to the rate of disfluencies is centered around the mean rate, but also reveals that some of them are either extremely fluent or extremely disfluent. This supports the argument for the usefulness of adapting the speech recognizer to each particular speaker.

Finally, to reliably understand spontaneous speech, the issue of modeling, detecting and correcting speech repairs must be addressed. As a first approach, in this paper we explore the extent to which acoustic events, lexical distortions and discourse markers may be used as cues for detecting speech repairs in Spanish. As previously observed for English, no single cue can be reliably used to detect speech repairs, since either the frequency or the coverage (or both) are too low. Lengthenings and cut off words are found to be the strongest cues but are not definitive. Some of the proposed cues could be simultaneously applied and enriched with more information, at the prosodic, syntactic or even semantic levels, to reliably detect speech repairs.

Our current work concerns modeling acoustic events as a first step towards a more general scheme which will include modeling approaches for lexical distortions and speech repairs. In particular, following recent approaches in the relevant literature, we plan to combine acoustic, prosodic and syntactic information to better detect speech repairs.

manuscript received: 06.15.2004
manuscript accepted: 11.08.2005

References

BAHL, L. R., BALAKRISHNAN-AIYER, S., BELLEGARDA, J. R., FRANZ, M., GOPALAKRISHNAN, P. S., NAHAMOO, D., NOVAK, M., PADMANABHAN, M., PICHENY, M. A., & ROUKOS, S. (1995). Performance of the IBM large vocabulary continuous speech recognition system on the ARPA Wall Street Journal task. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 41–44.

BALLESTER, A., SANTAMARÍA, C., & MARCOS-MARÍN, F. A. (1993). Transcription conventions used for the Corpus of Spoken Contemporary Spanish. *Literary and Linguistic Computing*, **8**(4), 283–292.

BEAR, J., DOWDING, J., & SHRIBERG, E. (1992). Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 56–63. University of Delaware, USA.

BEAR, J., DOWDING, J., SHRIBERG, E., & PRICE, P. (1993). *A system for labeling self-repairs in speech*. Technical Note 522, SRI

BONAFONTE, A., AIBAR, P., CASTELL, N., LLEIDA, E., MARIÑO, J. B., SANCHÍS, E., & TORRES, I. (2000). Desarrollo de un sistema de diálogo oral en dominios restringidos (in Spanish). In *Actas de las I Jornadas en Tecnología del Habla*, University of Sevilla, Spain.

BORTFELD, H., LEON, S. D., BLOOM, J. E., SCHOBER, M. F., & BRENNAN, S. E. (2001). Disfluency rates in conversation: Effects of age, relationship, topic, role and gender. *Language and Speech*, **44**(2), 123–147.

CLARK, H. H., & FOX-TREE, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition*, **84**(1), 73–111.

COLTHURST, T., KIMBALL, O., RICHARDSON, F., SHU, H., WOOTERS, C., IYER, R., & GISH, H. (2000). The 2000 BBN Byblos LVCSR System. In *Proceedings of the NIST Speech Transcription Workshop*, University of Maryland. <*http://www.nist.gov/speech/publications/tw00/*>.

DUEZ, D. (1993). Acoustic correlates of subjective pauses. *Journal of Psycholinguistic Research*, **22**(1), 21–39.

EARS-MDE. (2004). *Simple metadata annotation specification*. EARS Metadata Extraction (MDE) Project, Technical Report, Version 6.2, Linguistic Data Consortium. <*http://www.ldc.upenn.edu/Projects/MDE*>.

EKLUND, E. (2001). Prolongations: A dark horse in the disfluency stable. In *Proceedings of the ISCA Workshop on Disfluency in Spontaneous Speech*, 5–8, University of Edinburgh, Scotland.

EKLUND, E. (2004). Disfluency in Swedish human-human and human-machine travel booking dialogues. Ph.D. thesis, University of Linköping.

HAIN, T., WOODLAND, P. C., EVERMANN, G., & POVEY, D. (2000). The CU-HTK March 2000 Hub5E Transcription System. *In Proceedings of the NIST Speech Transcription Workshop*, University of Maryland. <*http://www.nist.gov/speech/publications/tw00/*>.

HEEMAN, P. A. (1997). *Speech repairs, intonational boundaries and discourse markers: Modeling speakers' utterances in spoken dialog*. Ph.D. thesis, University of Rochester.

HEEMAN, P. A. (1999). Modeling speech repairs and intonational phrasing to improve speech recognition. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop*, Keystone, Colorado.

HEEMAN, P. A., & ALLEN, J. F. (1995). *The TRAINS 93 dialogues*. TRAINS Technical Note 94–2, The University of Rochester, Computer Science Department, Rochester, New York.

HEEMAN, P. A., & ALLEN, J. F. (1997). Intonational boundaries, speech repairs and discourse markers: Modeling spoken dialog. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 254–261, Madrid, Spain.

HEEMAN, P. A., & ALLEN, J. F. (1999). Speech repairs, intonational phrases and discourse markers: Modeling speakers' utterances in spoken dialogue. *Computational Linguistics*, **25**(4), 527–571.

HEEMAN, P. A., & LOKEN-KIM, K. H. (1999). Detecting and correcting speech repairs in Japanese. In *Proceedings of the ICPhS Satellite Meeting on Disfluency in Spontaneous Speech*, 43–46, Berkeley, CA.

LDC94S19 (1994). ATIS SR Output (".sro") transcription conventions. Documentation for ATIS3 Training Data. Distributed by Linguistic Data Consortium, Corpus LDC94S19, <*http://wave.ldc.upenn.edu/Catalog/docs/LDC94S19/*>.

LEVELT, W. J. M. (1989). Speaking: From intention to articulation. The MIT Press. *ACL-MIT Press series in natural-language processing*, Cambridge, Massachusetts.

LICKLEY, R. J. (1994). *Detecting disfluency in spontaneous speech.* Ph.D. thesis, University of Edinburgh.

LICKLEY, R. J. (1998). *HCRC Disfluency Coding Manual*. HCRC Technical Report 100, Human Communication Research Centre, University of Edinburgh.

LIU, D., NGUYEN, L., MATSOUKAS, S., DAVENPORT, J., KUBALA, F., & SCHWARTZ, R. (1998). Improvements in spontaneous speech recognition. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, Virginia.

LJOLJE, A., HINDLE, D. M., RILEY, M. D., & SPROAT, R. W. (2000). The AT&T LVCSR-2000 System. In *Proceedings of the NIST Speech Transcription Workshop*, University of Maryland. <*http://www.nist.gov/speech/publications/tw00/*>.

METEER, M., TAYLOR, A., MACINTYRE, R., & IYER, R. (1995) *Disfluency annotation stylebook for the Switchboard Corpus*. Technical Report (revised by Ann Taylor, June, 1995). Published by the Linguistic Data Consortium, University of Pennsylvania, Department of Computer and Infomation Science.

NAKATANI, C. H., & HIRSCHBERG, J. (1993). A speech-first model for repair detection and correction. In *Proceedings of the ARPA Workshop on Human Language Technology*.

NAKATANI, C. H., & HIRSCHBERG, J. (1994). A corpus-based study of repair cues in spontaneous speech. *Journal of the Acoustical Society of America*, **95**(3), 1603–1616.

NEY, H., WELLING, L., ORTMANNS, S., BEULEN, K., & WESSEL, F. (1998). The RWTH large vocabulary continuous speech recognition system. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 853–856.

O'SHAUGHNESSY, D. (1993). Locating disfluencies in spontaneous speech. In *Proceedings of the European Conference on Speech Communications and Technology* (EUROSPEECH), 2187–2190, Berlin, Germany.

OVIATT, S. (1995). Predicting spoken disfluencies during human-computer interaction. *Computer, Speech and Language*, **9**(1), 19–35.

RILEY, M., LJOLJE, A., HINDLE, D., & PEREIRA, F. (1995). The AT&T 60,000 word speech-to-text system. In *Proceedings of the European Conference on Speech Communications and Technology* (EUROSPEECH), 207–210.

RODRÍGUEZ. L. J. (2002). *Cortado de señales y anotación de disfluencias en el Corpus Oral UAM* (in Spanish). Technical Report, Pattern Recognition and Speech Technology Group, Departamento de Electricidad y Electrónica, Facultad de Ciencia y Tecnología, Universidad del País Vasco.

RODRÍGUEZ, L. J., & TORRES, I. (2003). Annotation and analysis of acoustic and lexical events in a generic corpus of spontaneous speech in Spanish. In *Proceedings of the ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition* (SSPR),187–190, Tokyo Institute of Technology, Tokyo, Japan.

RODRÍGUEZ, L. J., TORRES, I., & VARONA, A. (2000). *Manual para el etiquetado de disfluencias* (in Spanish). Technical Report BS12BV30, Proyecto TIC98-0423-C06: Sistema de diálogo para habla espontánea en un dominio semántico restringido, UPV/EHU.

RODRÍGUEZ, L. J., TORRES, I., & VARONA, A. (2001a). Annotation and analysis of disfluencies in a spontaneous speech corpus in Spanish. In *Proceedings of the ISCA Workshop on Disfluency in Spontaneous Speech*, 1–4, University of Edinburgh, Scotland.

RODRÍGUEZ, L. J., TORRES, I., & VARONA, A. (2001b). Evaluation of sublexical and lexical models of acoustic disfluencies for spontaneous speech recognition in Spanish. In *Proceedings of the European Conference on Speech Communications and Technology* (EUROSPEECH), Aalborg, Denmark.

ROSE, R. C., & RICCARDI, G. (1999). Modeling disfluency and background events in ASR for a natural language understanding task. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 1709–1712.

SCHOURUP, L. (1999). Discourse markers (Tutorial overview). *Lingua*, **107**(3–4), 227–265.

SCHULTZ, T., & ROGINA, I. (1995). Acoustic and language modeling of human and nonhuman noises for human-to-human spontaneous speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), **1**, 293–296.

SHRIBERG, E. E. (1994). *Preliminaries to a theory of speech disfluencies.* Ph.D. thesis, University of California at Berkeley.

SHRIBERG, E. E. (1996). Disfluencies in Switchboard. In *Proceedings of the International Conference on Speech and Language Processing* (ICSLP)—Addendum,11–14.

SHRIBERG, E. E., BATES, R. A., & STOLCKE, A. (1997). A prosody-only decision-tree model for disfluency detection. In *Proceedings of the European Conference on Speech Communications and Technology* (EUROSPEECH), **5**, 2383–2386.

SHRIBERG, E. E., BEAR, J., & DOWDING, J. (1992). Automatic detection and correction of repairs in human-computer dialog. In *Proceedings of the DARPA Speech and Natural Language Workshop*, 419–424.

SHRIBERG, E. E., & LICKLEY, R. J. (1993). Intonation of clause-internal filled pauses. *Phonetica*, **50**(3), 172–179.

SHRIBERG, E. E., & STOLCKE, A. (1996). Word predictability after hesitations: A corpus-based study. In *Proceedings of the International Conference on Speech and Language Processing* (ICSLP),1868–1871.

SIU, M., & OSTENDORF, M. (1996). Modeling disfluencies in conversational speech. In *Proceedings of the International Conference on Speech and Language Processing* (ICSLP), 386–389.

STOLCKE, A., SHRIBERG, E., BATES, R., OSTENDORF, M., HAKKANI, D., PLAUCHE, M., TÜR, G., & LU, Y. (1998). Automatic detection of sentence boundaries and disfluencies based on recognized words. In *Proceedings of the International Conference on Speech and Language Processing* (ICSLP), **5**, 2247–2250.

STOLCKE, A., & SHRIBERG, E. E. (1996). Statistical language modeling for speech disfluencies. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), **1**, 405–408.

SUNDARAM, R., GANAPATHIRAJU, A., HAMAKER, J., & PICONE, J. (2000). ISIP 2000 Conversational speech evaluation system. In *Proceedings of the NIST Speech Transcription Workshop*, University of Maryland. <*http://www.nist.gov/speech/publications/tw00/*>.

WU, C. H., & YAN, G. L. (2001). Discriminative disfluency modeling for spontaneous speech recognition. In *Proceedings of the European Conference on Speech Communications and Technology* (EUROSPEECH), 1955–1958, Aalborg, Denmark.