Orality and Difficulties in the Transcription of Spoken Corpora

González Ledesma, Ana; De la Madrid Heitzmann, Guillermo; Alcántara Plá, Manuel; De la Torre Cuesta, Raúl & Moreno Sandoval, Antonio

Laboratorio de Lingüística Informática
Universidad Autónoma de Madrid
e-mail address{ana; guille; manuel; raul; sandoval}@lllf.maria.uam.es

Abstract

This paper analyses the effects of certain oral features on the process of transcription of spontaneous speech recordings. On the basis of the statistical analysis of the data obtained from the C-ORAL-ROM corpus, it will be shown empirically that transcription difficulties vary according to the communicative situation, the degree of formality and the number of participants.

1. Introduction

This paper is the result of an experiment carried out by a group of transcribers at the Laboratorio de Lingüística Informática (LLI) at the Universidad Autónoma de Madrid, once the recording and transcribing phases of the C-ORAL-ROM project were over.

The goal of the experiment was to confirm certain hypothesis which had arisen after the transcription process as the team attempted to determine which communicative interactions caused more difficulties in the transcription phase and why.

The original hypothesis consisted in relating these transcription problems to the frequency of occurrence of two kinds of linguistic phenomena typical of spoken interactions:

- **Production features,** such as *fragmented words, supports, retractings*, etc.
- **Interaction features**, such as the *number of turns* or the *overlapping* (Llisterri, 1997).

This would lead to the conclusion that the more frequent these phenomena were in a spoken interaction, the more time and effort needed by the linguist in the process of transcription.

However, it was the team's aim to rationalize these impressions and confirm the causes in an empirical way. Thus, the following objectives were stated:

- Definition of orality.
- Development of a computational tool which could help to establish a relation between conversational genres and orality features.
- Showing how these features vary inside the corpus depending of the register.
- Data analysis and verification of how orality is related to the difficulties present in the transcription process.
- Establishing a typology of transcription problems based on the results of the analysis.

However, before attempting further explanation of the experiment and analysis of the results, it is necessary to discuss some features of the corpus used, focusing on those which are related to its design and distribution.

2. Description of the corpus.

C-ORAL-ROM is a multilingual spontaneous speech corpus (Cresti et al., 2002) of the four main roman languages: French, Italian, Portuguese and Spanish. Each subcorpus consists of around 300,000 words. With the aim of enabling comparability between the different subcorpora, several sampling criteria concerning the distribution of the corpus were established: as long as each of the variation parameters is fully present in the corpus, the linguistic variation will be well represented (Moreno, 2002). In that sense, two elements are to be considered as basic elements: on one hand, the *characteristics of the speakers* and, on the other hand, the *context of use*. As far as the speakers are concerned, age, sex, education, occupation and geographical origin were taken into account. As for the contexts of use, a basic distinction was made between the dialogic structure (monologues and dialogues or conversations) and kind of situation (familiar or public).

A second important distinction was made between formal and informal speech. Each is represented in the corpus by 50% of the texts. Inside the informal part, a distinction was made between the familiar and the public domains: the first is represented by 75% of the texts, while the public domain accounts for the other 25%. As for the formal speech, the distribution of the texts was made following a thematic criterion: the *natural context formal speech* area (43% of the texts) is formed by recordings such as conferences, political debates, political speeches, sermons, professional explanations and texts dealing with business, law and teaching. In the same way, the texts which are part of the *formal speech in media* section (40% of the texts) are grouped in the following categories: interviews, meteo, news, reportages, scientific press, sports and talk-shows. Finally, inside the formal speech part, a section made up of phone recordings (17 % of the texts) is included.

Other relevant criteria concerning the corpus design are: acoustic quality of the samples (all are digital recordings), legal status (recording, transcription and publishing were done after the written authorization of all participants) and spontaneity of the recordings (no previous scripts were used and there were no restrictions in the use of the language and the expression of opinions).

Informal	*Familiar/Private*	*Monologue*
		Dialogue
	Public	
		Conversation

Formal		
Formal in natural context	*Media*	*Telephone*
political speech	news	private conversation
political debate	sport	phone call services (man interaction)
preaching	interviews	phone call services (machine interaction)
teaching	meteo	
Professional explanation	scientific press	
conference	reportage	
business	talk shows political debate	

Figure 1: Distribution of the C-ORAL-ROM corpus

3. The notion of orality.

It is well known that **spoken language** is not always a synonym to **orality**, if we understand orality as the presence of linguistic, paralinguistic and interactive phenomena, such as *retracting* or *overlapping*, which are not present in the written register. The registers in spoken language vary depending on the communicative situation. For instance, a text being a transcription of a sermon will differ significantly from a private conversation between friends, as far as the subject, the communicative context, the goals and the relation between participants are concerned (Romaine, 1996).

These differences are present not only at a morpho-syntactic, lexical and discoursive levels, but also at a more basic level which has to do with discourse production and which we will refer to as **degree of orality**.

The goal of this paper is to study that degree of orality considering the different conversational genres established in C-ORAL-ROM, in such a way that the hypothesis stated at the beginning -the presence of certain spoken features makes the transcription process much more difficult- can be confirmed.

Those phenomena chosen as the object of this experiment are typical features of spontaneous speech: overlapping, retractings, dialogic turns, speaking speed, fragmented words ("psicolog" instead of "psicología", for example) or supports, coded in C-ORAL-ROM as *&ah* and *&eh*.

4. Orality and transcription problems: the original hypothesis.

In order to find out what kind of relation there is between orality and linguistic registers, two **scales of transcription difficulty** were stated taking into consideration the following two parameters:

4.1. Degree of formality (Scale 1).

Two ends can be considered when dealing with the texts in terms of transcription difficulty: on one end, the most complex, those texts distinguished as *private*; on the other end, the easiest, those texts classified as *formal*.

informal media formal
+ difficult ---------- - difficult

4.2. Number of speakers (Scale 2).

This parameter affects only those texts classified as *informal* (the most complex according to Scale 1) and considers as most complex those texts with a higher number of participants (three or more), while those with one or two participant imply a lower degree of difficulty.

conversation dialog monolog
+ difficult ---------- - difficult

5. The computational tool.

The C-ORAL-ROM corpus is tagged with XML. Using the information included in the tags, we developed a program which automatically calculate the frequency of occurrence of each of the following features: overlapping, retracting, number of dialogic turns, speaking speed, fragmented words and supports. These frequencies were calculated for each class of texts.

Thus, the results show the average number of words between two occurrences of a phenomenon, except in the case of *speaking speed*, where the figures correspond to the number of words per second. The higher the number of words, the less important is the phenomenon in the class of text in question. In order to facilitate the reading of the figures, only one decimal was used in the final results.

6. Textual typology and transcription problems: analysis of the data.

In this section, the results obtained by the program are analyzed. The analysis procedure has always been the same for each of the linguistic phenomena studied:

First, the relation between frequency of occurrence of the features and textual typology is stated.

Second, we evaluated whether this relation confirms the original hypothesis, which states that certain kinds of

texts are harder to transcribe, according to the scales of difficulty.

6.1. Number of dialogic turns.

The first feature to be analyzed is the **number of dialogic turns**, understood as the number of times a speaker replaces another in the conversation. According to the original hypothesis, in the analysis of the data it is assumed that there is a direct relation between the number of turns and the effort needed in the transcription process.

Below, it is shown how this feature is reflected in the mentioned classifications from a quantitative point of view.

In *Figure 1*, which analyses the **degree of formality (scale 1)**, it can be observed how the participants in informal texts produce shorter turns, while those turns belonging to formal texts are longer and those turns produced in media texts have an intermediate length, closer to formal texts than to informal ones.

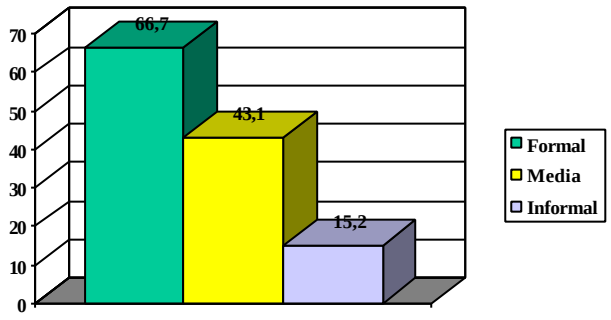


Figure 1: Words per turn in Scale 1.

In the groups dealing with **number of speakers (Scale 2)**, apart from the obvious conclusion about monologues, those turns belonging to dialogues are almost two and a half words longer than those belonging to conversations.

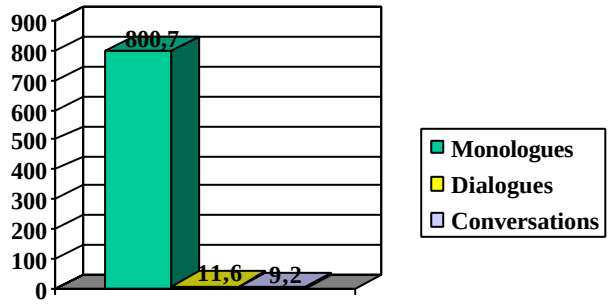


Figure 2: Words per turn in Scale 2

These results confirm the original hypothesis, that is to say: the higher the number of speakers, the shorter are the turns (considering the number of words per turn) and therefore the bigger is the effort necessary in the transcription. Furthermore, it has been proved that shorter turns are typical of informal texts and so it is in this area of the corpus where the transcriber will find more difficulties.

6.2. Overlapping.

This second feature is directly related to the previous one and represents, according to C-ORAL-ROM transcribers, one of the most important difficulties in the transcription task: **overlapping**. Again, the results are obtained dividing the number of words by the number of overlapping cases (except in monologues, where there is obviously no overlapping).

Figure 3 is the confirmation of *Figure 1*. As expected, overlapping is less frequent in the formal and media genres than in the informal one. In the informal genre, as shown in *Figure 4*, the difference between dialogues and conversations is an average of almost ten words. These data prove that overlapping is prototypical

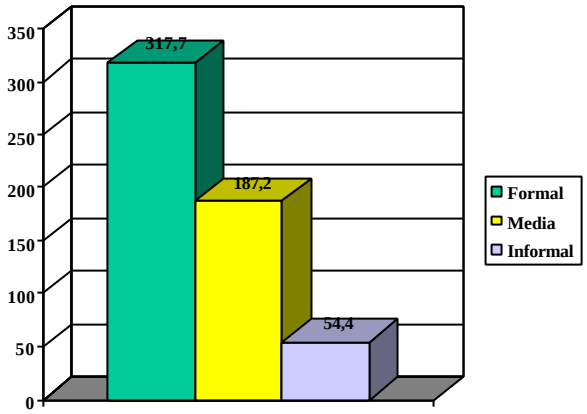


Figure 3: Wods per overlapping in Scale 1

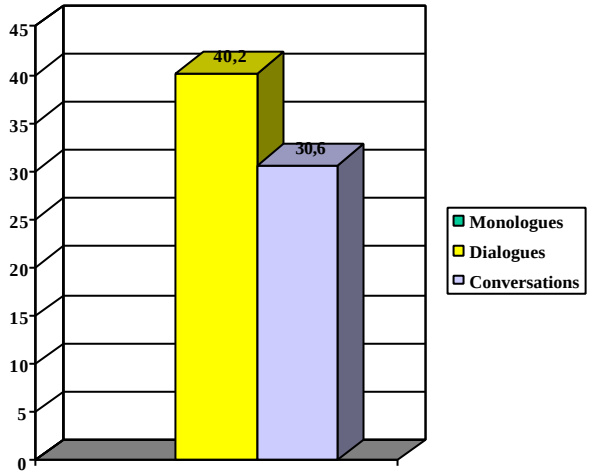


Figure 4: Words per overlapping in Scale 2.

of the informal genre and, furthermore, of the conversation subgenre. As far as the transcription task is concerned, this fact puts the conversation subgenre on the furthest end in terms of transcription difficulty.

6.3. Speaking speed.

Another important feature for the transcriber is the speed at which the participants speak. These are the results obtained for C-ORAL-ROM.

This feature, expressed in words per second, confirms once more how, in terms of speaking speed and given that the faster a participant speaks the harder the is to transcribe, the informal genre and the conversational subgenre are the most laborious in the transcription task. As we can see in the figures, a prototypical participant in

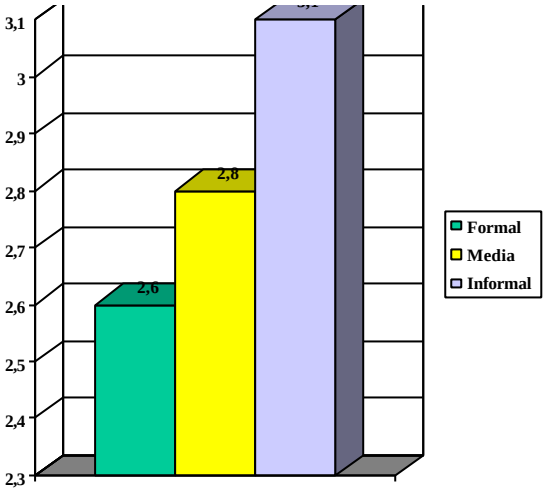


Figure 5: Words per second in Scale 1

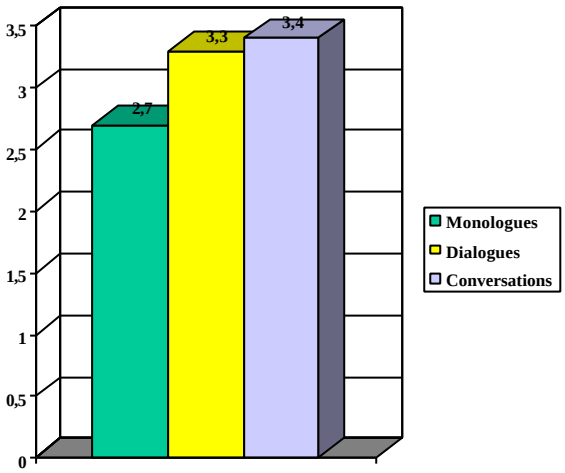


Figure 6: Words per second in Scale 2

an informal conversation utters approximately three and a half words per second, while for formal texts and informal monologues the average is 2.6-2.7 words per second.

6.4. Fragmented words.

So far, the data has confirmed the original hypothesis. However, in regards to the frequency of fragmented words, the hypothesis was not supported. A fragmented word occurs when a speaker does not complete the utterance of the word.

In *Figure 7*, it becomes obvious that most participants in the media genre are speaking *professionals*. Even though they speak at a higher speed than those appearing in formal texts (*Figure* 5), the frequency of occurrence of fragmented words in this kind of texts is much lower than it is in other genres, which share almost the same ratio. On the other hand, the formal genre is characterized by a high number of fragmented words.

Also, unexpectedly, *Figure 8* shows that the number of fragmented words is higher in dialogues than it is in monologues and conversations. The fact that dialogues are not in an intermediate position (as it happens in the rest of the results) leads to the conclusion that there is not a direct relation between number of participants and frequency of occurrence of fragmented words, an hypothesis that should be confirmed with further data.

All this would show how, in the transcription process, fragmented words are not perceived by the transcriber as an added difficulty, given that, in the difficulty scale (*Scale 1*), the formal genre is the easiest to transcribe.

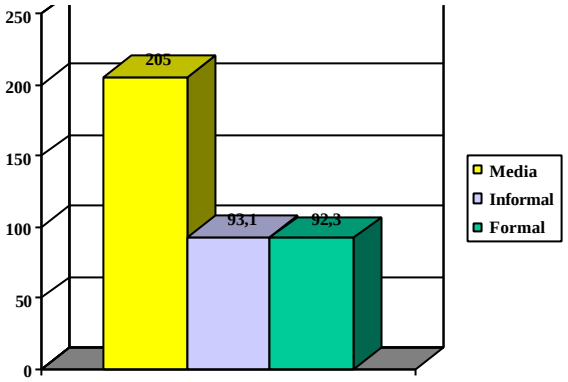


Figure 7: Words per fragment in Scale 1

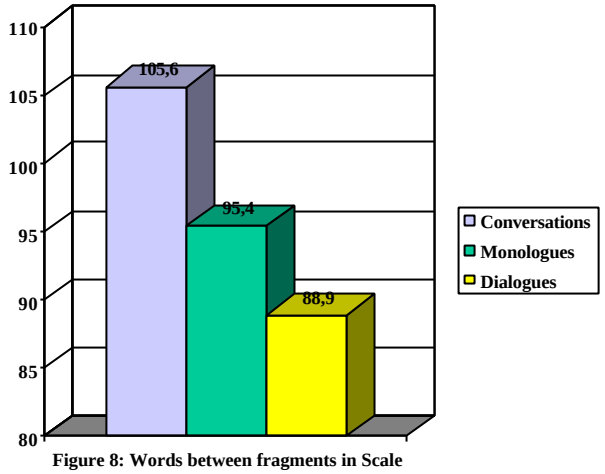


Figure 8: Words between fragments in Scale 2.

6.5. Supports.

The following analysis corresponds to the frequency of occurrence of **supports.**

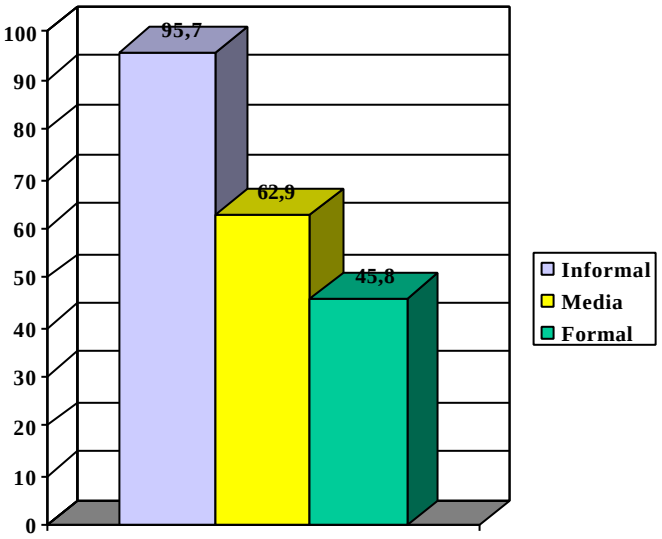


Figure 9: Words between supports in Scale 1

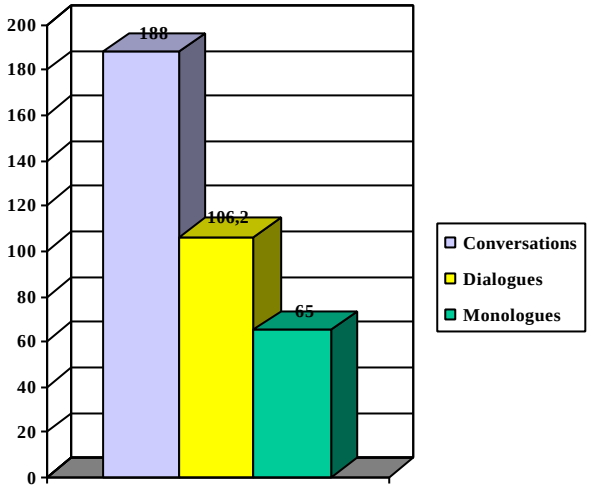


Figure 10: Words between support in Scale 2.

Figure 9 shows an opposite arrangement to the difficulty order originally suggested, where formal texts were classified as the easiest ones. Similarly to the case of fragmented words, the prototypical participant in a formal recording resorts to supports every 46 words, in order to sustain his discourse. This contrasts with the ratio in informal texts, where the average is of almost 96 words.

The information presented in this figure is quite unexpected, especially when, in this sense, the formal genre is made up of communicative interactions such as conferences or lessons in an academic context, which are quite close to texts following some kind of script. This helps to understand this phenomenon not as a symptom of lack of planning, but as a support which participants in this kind of recordings find useful or necessary.

These data could somehow be connected to the number of participants, that is to say, the more the speakers in a conversation, the less supports are used, due to the dynamics of the interaction. In order to confirm this hypothesis, we can look at *Figure 10*, where the number of participants is one of the parameters.

Nevertheless, this table shows how conversations represent the texts with the lowest frequency of occurrence of supports. In fact, the data prove that, as the number of participants increases (and, if we look at *Figure 2*, the turn is longer), so do the frequencies of supports.

Therefore, this is again a revealing finding. First of all, for the description of the different linguistic registers of the spoken corpus and further studies on this field. Secondly, it is also important for the transcribers, as supports do not seem to constitute an added difficulty in the transcription process.

6.6. Retractings.

Finally, figures relating the frequency of occurrence of **retractings** were analyzed in the six groups, obtaining the following results.

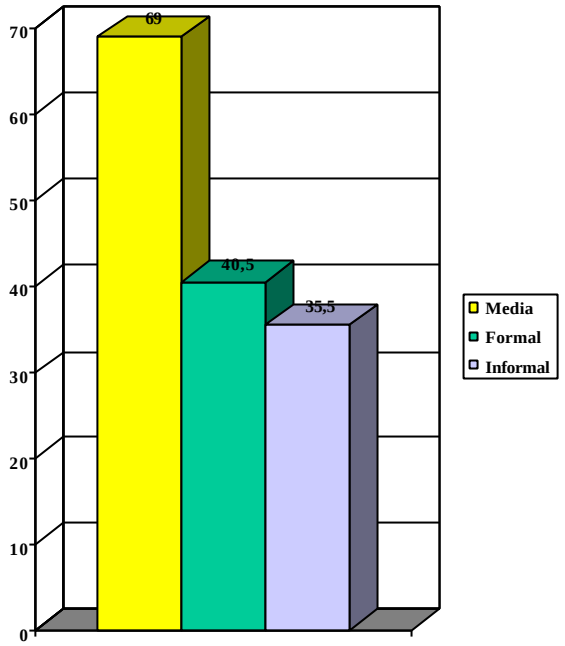


Figure 11: Words between restarts in Scale 1

Again, the scale of difficulty is inverted. Even though the frequency of occurrence of retractings increases in the informal texts, and this matches the predictions made, it is interesting to observe how the formal and media genres invert their positions with respect to the figures. This leads to important conclusions regarding the differences between these two genres, which are in principle quite similar, given that both are planned and are characterized by a register which is close to written language.

As for the results in *Figure 12*, retracting is a characteristic phenomenon of informal monologues, which again raises the question of the motive behind this phenomenon. It is interesting to remark that this feature presents the highest frequencies in a kind of text where there is no interaction at all between participants.

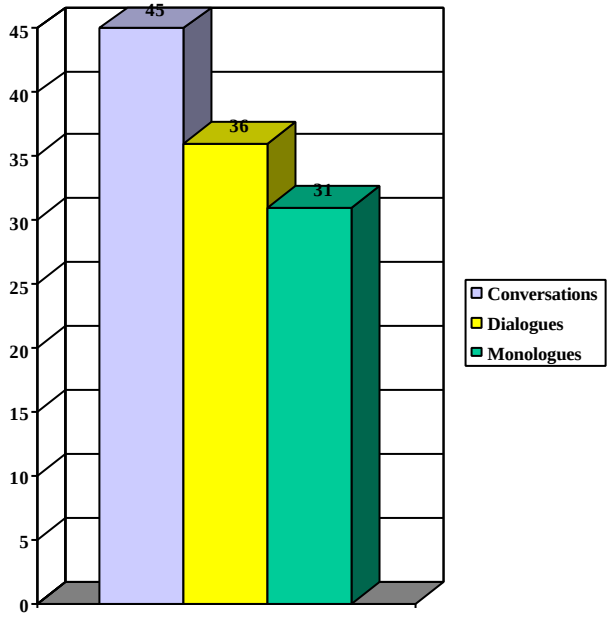


Figure 12: Words between restarts in Scale 2

Regarding the transcription problems, again the original hypothesis is inverted. Contrary to what was expected, there is not a direct relation between the presence of retractings and the degree of difficulty in the transcription process, as the *informal monologue* is the last in the scale of difficulty (Scale 2) introduced in section 3.

7. Conclusions.

Sometimes, the obvious facts has to be proven in order to question its value of truth, and this is exactly what has been accomplished in this paper. Starting from a apparently natural hypothesis, which consists in relating the presence of certain spoken features to a special difficulty in the transcription process, it has been deduced from the analysis of the results that this hypothesis is not always true because there are some spoken features such as supports, whose frequency of occurrence is higher in those texts which, as it is the case with formal texts and informal monologues, are not an added obstacle for the transcriber.

All this would lead to the classification of the features of the corpus into two groups:

- **Interactional features:** number of words per turn, frequency of overlapping and speaking speed (*Figures 1-6*).
- **Production features:** frequency of occurrence of fragmented words, supports and retractings (*Figures 7-12*).

The distribution of the types of text in the case of group 1 matches exactly the intuitive difficulty scales presented as *scales 1 and 2*.

However, the cases in group 2 (production features) show a distribution which is even opposite to scales 1 and 2 in some of its aspects: *media* is the genre with less influence coming from fragmented words and retractings, and the *formal* genre is the one with a highest ratio of fragmented words and supports (this last feature shows its lowest ratio in *informal* texts).

As for the second scale (informal texts), observing the second group of features (production features), *conversations* appear as the less affected subgenre, while *monologues* stand out as the richest in supports and retractings. The scales, if only the production features were taken into account, would be as the following, from less to most difficult:

- difficult ---------- + difficult
media informal formal

Figure 2: Scale 1 and production features.

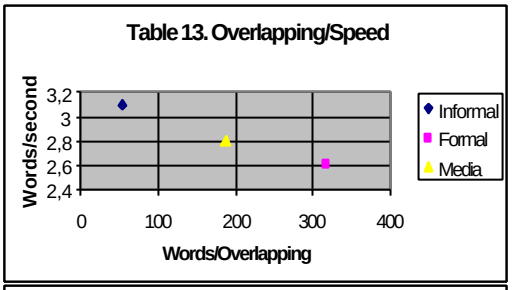
- difficult ---------- + difficult
conversations dialogues monologues

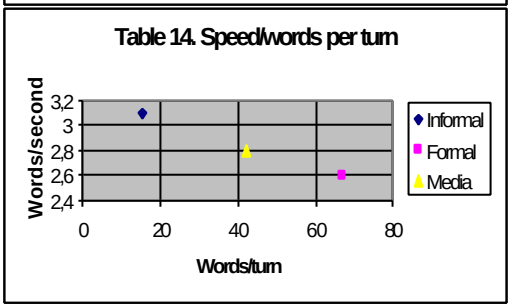
Figure 3: Scale 2 and production features.

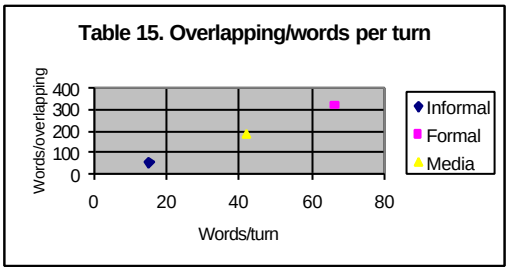
The fact that *interactional features* match exactly the intuitions made at the beginning and that production features are almost the opposite (the highest difficulty end corresponds to formal texts and informal monologues) gives way to the conclusion that the difficulty perceived by the transcriber comes from the features in the first group (interactional). This is shown below in the correlation of the data for interactional features in *Scale 1* (that is, the data on figures 1 to 6 for informal, formal and media). It is clearly shown on the figures how the order informal-media-formal is kept at all the times. The three figures correspond to relations between **speed and overlapping** (*Figure 13*), **speed and words per turn** (*Figure 14*) and **overlapping and words per turn** (*Figure 15*).

Those features belonging to the second group (*production features*) are problematic in the *establishment of the text* (Benveniste, 1998) phase, as it has to be decided what is going to be the written representation for that kind of recording; however, the transcriber does not consider these features as obstacles in the transcription process.

These empirical conclusions should be confirmed by applying the same analysis on new texts, as it will be







the case with a previously created corpus in LLI-UAM: CORLEC (Moreno, 2002). CORLEC does not follow exactly the same transcription criteria as C-ORAL-

ROM (mainly because it was recorded and transcribed 10 years before), but it has the advantage of being three times larger in terms of the number of words.

Nonetheless, the main conclusion (the complexity of a transcription derives from the interaction features and not from the production features) is fully justified by the representative character of the data used as a basis. More specifically, there were 429 different speakers, some of them participating in different recordings, which means a total of 554 participants. The number of texts is not very high (169 recordings) but its great variety should be highlighted (as shown in the distribution chapter). Finally, below is a summary of the data used:

Total data in C-ORAL-ROM (general data)				
Texts	Speakers	Participants	Turns	Words
169	429	554	15595	312597
Total data in C-ORAL-ROM (features)				
Overlapping	Retractings	Fragmented words	Supports	
4307	7860	3084	4807	

Table 16: Absolute data for C-ORAL-ROM.

Further investigation applying this methodology, extended to other criteria, might include characterizing the different spoken registers included in C-ORAL-ROM at all the linguistic levels. Besides, an optimum result of applying this methodology would lead to a prediction of the typology of a given spoken text, based on quantitative data not in qualitative ones.

8. References.

Blanche-Benveniste, C. (1998). Estudios lingüísticos sobre la relación entre oralidad y escritura. Barcelona, Gedisa.

Briz, A. (1996). El español coloquial: situación y uso. Madrid, Arco Libros.

Cresti, E. et al. (2002). The C-ORAL-ROM project. New methods for spoken language archives in a multilingual romance corpus. In Proceedings of LREC 2002. Las Palmas de Gran Canaria.

Gallardo, B (1993). La transición entre turnos conversacionales: silencios, solapamientos e interrupciones. In Contextos, XI/21-22, (pp. 189--220).

Halliday, M.A.K. (1985). Spoken and Written Language. Oxford University Press.

Listerri, J. (1997). Transcripción, etiquetado y codificación de corpus orales. In Fundación Duques de Soria. Seminario de Industrias de la Lengua. http://liceu.uab.es/~joaquim/publicacions/FDS97.html

Martí, M.A (Ed.) (2003). Tecnologías del lenguaje. Barcelona, UOC.

Moreno, A. (2002). La evolución de los corpus de habla espontánea: la experiencia del LLI-UAM. In Actas de las II Jornadas en Tecnologías del Habla, diciembre de 2002.Granada.

Tusón, A. (1997). Análisis de la conversación. Barcelona, Ariel.

Rodríguez L.J., I. Torres & A. Varona (2001). Annotation and analysis of disfluencies in a spontaneous speech corpus in Spanish. In Proceedings of the workshop on Disfluency on Spontaneous Speech. Scotland, University of Edimburgh.

Romaine, S. (1996). El lenguaje en la sociedad. Una introducción a la sociolingüística. Barcelona, Ariel.