LANGUAGE AND COMPUTERS:
STUDIES IN PRACTICAL LINGUISTICS

No 56

edited by
Christian Mair
Charles F. Meyer
Nelleke Oostdijk

Corpus linguistics around the world

Edited by
Andrew Wilson
Dawn Archer
Paul Rayson

Rodopi
Amsterdam - New York, NY 2006

Cover design: Pier Post

©Cover image: NASA & the Visible Earth (http://visibleearth.nasa.gov/)

Online access is included in print subscriptions.
see www.rodopi.nl

The paper on which this book is printed meets the requirements of "ISO 9706:1994, Information and documentation - Paper for documents - Requirements for permanence".

ISBN: 90-420-1836-4 (bound)

Printed in The Netherlands

Contents

Corpus linguistics around the world

Andrew Wilson, Dawn Archer and Paul Rayson

Lancaster University

Preface

The scope of corpus-based research is becoming ever wider.

Not so many years ago, the vast majority of corpus-linguistic research was concerned with the grammar and vocabulary of standard language varieties – the latter meaning, in many cases, British or American English. Whilst research on other topics, languages, and varieties was by no means completely absent from the scene, this was the general picture of the field which came across to the interested observer.

Today, things have changed dramatically. As this volume shows, the range of languages, research questions, and, indeed, methodologies which are addressed by corpus linguists has diversified. It is probably true to say that none of the papers published in this volume focuses primarily on standard English as a general variety. Here we find work not only on English dialects (Lee & Ziegeler) but also on learner language (Carson-Berndsen, Gut & Kelly; Lenko-Szymanska; Tagnin) and on a wide range of world languages - Basque (Aduriz et al.), Chinese (Xiao-jun), Danish (Asmussen), Dutch (Speelman et al.), German (Jones), Maltese (Bovingdon & Dalli), Russian (Sharoff), Slovene (Gorjanc), and Spanish (Guirao et al.). In terms of the research questions addressed, the more 'traditional' areas of corpus linguistics are still well represented, with papers on vocabulary (Jones), spoken language (Carson-Berndsen, Gut & Kelly; Guirao et al.), synchronic and diachronic variation (Asmussen, Beeching; Bovingdon & Dalli; Gorjanc; Lee & Ziegeler; Speelman et al.), Languages for Special Purposes (Noguchi, Orr & Tono), tagging, and corpus development (Aduriz et al.; Sharoff). However, exciting new departures are also present, with corpus-based work now extending into areas such as cross-cultural rhetoric and social psychology (Lenko-Szymanska; Wilson & Moudraia) and even economic forecasting (Ahmad et al.).

The papers published in this volume are but a small selection from the many which were presented at the Corpus Linguistics 2003 conference, held at Lancaster University in March 2003. This was the second Corpus Linguistics conference which we hosted at Lancaster (the first was in 2001), and, like its predecessor, it truly amazed us with the range of corpus-informed work being carried out world-wide. Computer corpus linguistics continues to thrive and to extend into so many areas of inquiry, many of which would probably have been unimaginable for its pioneers in the 1960s and 1970s. We are sure that it will

Manning C., and H. Schütze (1999) *Foundations of Statistical Natural Language Processing*. Cambridge MA: The MIT Press.

McEnery T., J. Langé, Oakes, M. and J. Véronis (1997), The exploration of multilingual annotated corpora for term extraction, in R. Garside, G. Leech, A. McEnery (eds.), *Corpus Annotation. Linguistic Information from Computer Text Corpora*. London, Longman.

Meyer, I., Mackintosh K., Barriere, C. and T. Morgan (1999), Conceptual sampling for terminological corpus analysis, in Sandrini (ed.), *Proceedings of TKE '99*. Vienna, TermNet, pp. 256-267.

Mladenić D. (2002), Automatic word lemmatisation, in: T. Erjavec and J. Gros (eds.), *Jezikovne tehnologije. Language Technologies*. Ljubljana, Institut Jozef Stefan, pp. 153-159.

Pearson J. (1998), *Terms in Context*. Amsterdam, John Benjamins.

Vintar Š, and V. Gorjanc (2000) *Identifying markers of semantic relations in Slovene*. http://www2.arnes.si/vinta/telri.rtf

Relating linguistic units to socio-contextual information in a spontaneous speech corpus of Spanish

José María Guirao

Universidad de Granada

Antonio Moreno Sandoval, Ana González Ledesma, Guillermo de la Madrid, Manuel Alcántara

Universidad Autónoma de Madrid

Abstract

This chapter shows the application of statistical tests to a corpus of spontaneous spoken Spanish. Our goal is to find representative differences between different parts of the corpus. To this end, we tagged n-grams in the corpus with features related to the speaker (age, gender, etc.), or the context (dialogue, monologue, media, etc.), and applied the log-likelihood test (Dunning, 1993) in order to find the most distinctive lexical or grammatical items for each specific socio-contextual feature.

This chapter is divided in three sections. In the first, the characteristics of the spoken corpus are shown. The second section is devoted to the explanation of the computational tool. In the third section, a first rough estimate of the results obtained is given, as well as possible applications of the model.

1. The Spanish corpus of the C-ORAL-ROM project.

C-ORAL-ROM is a multi-lingual corpus of spontaneous speech for the main four Romance languages, French, Italian, Portuguese and Spanish (Cresti et al. 2002). The project is funded by the EU under the V Framework Programme (IST-2000-26228) and the consortium consists of 9 partners, co-ordinated by the University of Florence. The remarkable feature of C-ORAL-ROM is its spontaneity: texts have been recorded in their actual context and without any script. Each sub-corpus is made up of 300,000 words with the same text distribution to assure comparability and sufficient register representation. The resource will be delivered in several formats: an orthographic transcription, an xml-tagged version, and the aligned audio source. Partial linguistic annotation will be provided, as well as some programs to handle the resources and quantitative studies. This paper shows preliminary results with respect to the Spanish corpus.

1.1 Differences between a speech database and a corpus of spontaneous speech

When discussing spoken resources, a preliminary distinction has to be made. Most linguistic resources currently available are speech databases: collections of high-quality recordings and detailed phonetic transcriptions of speech set up in controlled environments (typically telephone services). These speech databases are mostly used for training and testing speech systems and they are developed by and for the language engineering industry. They aim to serve as a basis for recognizing and producing speech in restricted, predictable domains. In most cases, those databases contain many samples of the same word (that is, many tokens of the same type). Usually, the utterances are prepared and pronounced by professional speakers. The acoustic quality of the recording is essential. Speech databases usually provide detailed phonetic descriptions, including disfluencies, noises and other sounds. In general, those databases reflect the standard register, and distant variants (dialects, jargons) are poorly represented. Instances of those are SpeechDat (LRE-63314, Infrastructure for Spoken Language Resources), SpeechDat II (LRE2-4001, Speech Databases for the Creation of Voice Driven Teleservices), which have set up a standard for this type of resource.

On the other hand, corpora of spontaneous speech are typically collections of a wide variety of spoken registers and non-scripted speech. Those corpora are collected mainly for linguistic analyses and applications (language teaching, grammars and dictionaries). In such corpora the acoustic quality is not essential. What is important is that the texts reflect as much variation as possible and the speaker behaves in a spontaneous manner. In some cases, those corpora are only concerned with a given register, for instance, a dialect or children's speech. An important difference with respect to speech databases is the transcription: spontaneous spoken corpora usually are less precise in the acoustic and phonetic parts. On the contrary, they include detailed information about the context and the speakers. These corpora are used mainly for sociolinguistic, text-typologic, or psycholinguistic analyses. Examples are CHILDES and London-Lund.

C-ORAL-ROM is a corpus of spontaneous speech, but it also shows some distinctive features:

Multilingual: the main goal is to compare the four languages, on the same grounds, and provide comparative studies at different linguistic levels.

Acoustic quality: in order to be re-usable by the speech industry, sufficient samples of digital recordings, media and phone conversations are included.

Alignment of the transcription and the original sound: this is useful both to verify the accuracy of the transcription and for teaching and other applied investigation purposes.

The main limitation of C-ORAL-ROM is its size. 300,000 words per language is not a sufficient number for stating classifications and statistically significant analyses. We believe that this corpus will show the relevancy and usefulness of an approach that pays as much attention to the acoustic quality of the register as to the linguistic annotation.

1.2 Multi-lingual comparability

Cross-linguistic comparison can provide two complementary perspectives: comparing a given feature or features across languages, and comparing a given register or text type across languages (Biber, 1995). On the C-ORAL-ROM project different traditions and experiences have interacted. On the practical side, the teams came to an agreement around two basic points: a text distribution (or sampling design) and a unified format for the transcription.

1.2.1 Text distribution

In order to compare the linguistic features of the four languages, the same common sampling criteria and the same proportion of each type in the four sub-corpora are needed. There is a long tradition in sociolinguistics and in corpus linguistics (Labov, 1966; Biber, 1988; Biber et al., 1999; Miller & Weinert, 1999) in determining the relevant non-linguistic parameters. Basically, authors agree with a series of socio-situational parameters, such as register and genre variation, sociological features of the speakers (sex, age, education, occupation, origin), and dialogic structure (monologue, dialogue, conversation). The disagreement is in how to combine these parameters. C-ORAL-ROM has chosen the design of the Spoken Dutch Corpus (http://lands.let.kun.nl/cgn/ehome.htm). The sampling design is different in both sub-corpora:

> *Informal* register is organised according to social context (familiar-private vs. public) and dialogic structure (monologue vs. dialogue-conversation).
>
> *Formal* register is organised according to channel (media, telephone, natural context). In addition, media texts and formal in natural context are grouped by genre (see table below).

Sociological features have not been taken into account for text selection, but they are explicitly marked in the metadata section of the transcription. Male/Female distinction has been the only feature to be balanced.

With respect to the text length, some decisions have been made. Only three types of size are allowed: short (1,500 words), medium (3,000 words) and large (4,500 words). Texts shorter than 1,500 words have allowed in genre types like meteorological reports, but always compounding segments of 1,500 words.

Tables 1 and 2 show the distribution design of the informal and formal sub-corpora for each language.

Table 1: Informal sub-corpus

Private/Familiar Context 113,000 words		Public Context 37,000 words	
Monologue 33,000 words	Dialogue 80,000 words	Monologue 6,000 words	Dialogue 31,000 words

Table 2: Formal sub-corpus

Formal in Natural Context 65,000 words	Formal in Media Context 60,000 words	Telephone 25,000 words
Political Speech	News	Private Dialogues
Political Debate	Meteo	Phone to Call Services
Preaching	Interviews	
Teaching	Reportage	
Professional Explanation	Scientific Press	
Conferences	Sport	
Business	Talk Show Political	
Law	Thematic Explanation	
	Talk Show Culture	
	Talk Show Science	

1.2.2 Common format

To ensure a valid comparison, it is also necessary to use a consistent annotation framework. The consortium developed the C-ORAL-ROM format, which is based in the known CHAT format. A conversion to XML is provided. The xml-tagged version guarantees easy interpretation though the corresponding DTD. The combined use of XML and DTD ensures that every text in each corpus complies with the same requirements. In this way, textual uniformity are obtained throughout and between the four corpora.

The format is divided into the *header* (with the meta-data) and the *transcription*. Most features in the header are compulsory, therefore a rich information is provided for every text. The transcription is divided into turns, where applicable. Each turn is marked by a three-letter code identifying the speaker. An orthographic transcription is provided, along with some tags marking disfluencies, noises, overlapping, and prosodic units. Morpho-syntactic tagging will be supplied in a separate tier. Figure 1 shows a fragment of a text. A large selection of fragments from the four languages can be consulted on the official webpage of the project, along with the sound source.


```
@Title: Raquel
@File: efamdl04
@Participants: PAT, Patricia, (woman, B, 2, hairdresser, participant,
Madrid)
   ROS, Rosa, (woman, B, 3, English teacher, participant, Madrid)
@Date: 10/03/2001
@Place: Madrid
@Situation: chat between friends at home, not hidden, researcher
observer
@Topic: friends, movies and future Use's works
@Source: C-ORAL-ROM
@Class: informal, familiar/private, dialogue
@Length: 7' 58''
@Words: 1509
@Acoustic_quality: A
@Transcriber: Guillermo
@Revisor: Manuel; Guillermo, Jesús and Manuel (prosody)
@Comments:

*PAT: si ya han [/] han decidido ir con ellos / y la conocen ...
*ROS: ya / pero si yo no [/] si a mí me da igual / si yo no digo nada de
Use y Nuria / yo digo que la peña es un / poco egoista //
*PAT: <no> //
*ROS: [<] <y ya está> //
```

Figure 1: A fragment of C-ORAL-ROM text

1.3 Other relevant aspects

C-ORAL-ROM is compliant with the state of the art in spoken corpora. These aspects are briefly summarised in the following paragraphs.

1.3.1 The legal issue

During the 1990s legislation on Copyright and Privacy changed in many European countries. In spoken language corpora, the law is applied when recording individuals or using sound documents from the mass media. In the first case, speakers retain their right to preserve privacy, and have to give their express authorisation in order to their speech will be transcribed and published. In order to preserve spontaneity, which is essential for our purposes, the procedure is to ask each participant to sign an authorisation *after* the recording. If a speaker refuses his/her consent, then the recording is discarded. The right to privacy applies to every recording in a private context, but not to ones in a public situation (a lecture, a political speech, a sermon).

On the other hand, many texts in the corpora are copyrighted, not only the media recordings but also those in which the speaker creates knowledge in the form of ideas or structure of contents. Typically, this is true of lectures and professional talks. We obtained the written authorisation from the authors or the copyright holders for all the texts included in the Spanish corpus.

1.3.2 The acoustic quality

The Spanish corpus of C-ORAL-ROM has been collected from scratch, although other teams in the project have reused part of their previous texts. In our case, we preferred to make new recordings because, on one hand, we did not have the written consent for our previous texts and, on the other hand, the acoustic quality of the analogical tapes was poor.

Most texts have been recorded with a DAT Tascam (model DA-P1) and two unidirectional microphones. The source has been converted into a WAV file, mono, 16 bit, 22.050 Hz, through a SPDIF port in a Sound Blaster Live Platinum 5.1, using the software Creative Recorder. In public places, when possible, the DAT recorder has been connected to the sound system. The media recordings either have been provided directly by the broadcasting station or recorded by a computer connected to the receiver.

Acoustic quality is essential for the application in speech technologies and language engineering.

1.3.3 The linguistic annotation

Corpora increase in value depending on the annotation layers provided. Tagging a spontaneous speech corpus is a task slightly different to the same for written corpora (Uchimoto et al., 2002). The difference is not in the tagged information but in the lesser efficiency of the taggers when applied to spoken corpora. For instance, some POS taggers are usually trained on written texts, which show a quite stable and determined word order. On the contrary, corpora of spontaneous speech are highly flexible in word order. In addition, they show repetition, re-startings, overlapping, and other features of spoken syntax which have to be trained specifically.

The lexicon is also different. One can find many words that are not included in printed dictionaries, because they are innovations, or belong to an informal register, or simply because they are mispronunciations.

A complete lemmatization and POS tagging is provided. Moreno and Guirao (2003) report the development of a POS tagger and unknown words recogniser for morphosyntactic annotation of the Spanish corpus. The results provided by the lemmatizer are used in this paper (see section 3).

1.3.4 The validation

To verify the reliability of data has become a fashionable topic in the recent years. Users of linguistic resources want to know how the resources have been collected and their accuracy. C-ORAL-ROM passes two types of evaluation.

An internal validation is carried out by the team itself. Each text passes through five steps: transcription, first revision, prosodic tagging, second revision, and sound-text alignment. At least, three linguists transcribe/revise each text. A program verifies format errors, blanks, typos, badly formed tags, etc. Therefore, content and form have been validated exhaustively, guaranteeing that the transcription is accurate to the sound source. We want to stress that the alignment of sound and text is the best guarantee for validation of a spoken text: any discrepancy between the actual speech and its transcription will be easily detected.

An external validation will be done by experts at the end of the project.

2. The computational tool

We have developed computational tools for transforming the C-ORAL-ROM format into a more suitable tagging scheme in order to relate meta-data with lexical items, and compute the appropriate statistics. We will divide this section into three sections.

2.1 Using xml-tagged corpus for relating meta-data and linguistic features

The original C-ORAL-ROM annotation has been designed for registering a wide range of features, including acoustic ones (prosodic marks, noises, etc.) which will be used by the speech technology community. An example of an xml-tagged file is shown:

```
<Turn>
<Name>PAT</Name>
<Says>
<Utterance Type= "interrogation"> y cómo está </Utterance>
<Notes Type= "act"> cough </Notes>
</Says>
</Turn>
<Turn>
<Name>ROS</Name>
<Says>    <Utterance Type= "enunciation"> bueno </Utterance>
<Utterance Type= "enunciation"> no está mal </Utterance>
</Says>
</Turn>
```

Our goal in this experiment is to seek out lexical units peculiar to each sub-corpus. The first step was to remove the non-lexical information from the original xml tagging. In particular, we wanted to capture two types of information:

i) The words that every speaker says, and

ii) The split of every turn into utterances in order to prevent ill-formed word clusters. This task is similar to tokenisation in written corpora. This division into utterances is also needed for delimiting the context that the POS tagger uses for disambiguation.

A Perl script generates a new tagged corpus with only two tags: one for TURN, with attributes for `speaker` and `file`, and another for UTTERANCE:

<turn speaker="PAT" file="efamcv01"> <utt> y como está </utt> </turn>

<turn speaker="ROS" file="efamcv01"> <utt> bueno </utt> no está mal </utt> </turn>

By this means, every word in the corpus can be related with the speaker and the text. The file keeps in the header all the socio-contextual information. The corpus is partitioned in as many sub-corpora as different features appeared in the header. For instance, a male sub-corpus, an informal sub-corpus, a telephone sub-corpus, a meteo sub-corpus, etc. are all generated. After partition into sub-corpora, all occurrences (the tokens) for every lexical unit (the types) are counted in each sub-corpus. The next table shows the distribution by sex of speaker.

Table 3: Distribution by sex

Sex	Tokens for the category	Total number of tokens	Percentage
Man	182832	327044	55.9 %
Woman	134693	327044	41.2 %
X	9519	327044	2.9 %

The "X" value is assigned when the sex of the speaker is unknown (typically in a media recording).

The procedure can be applied to any type of information derived from the corpus. For instance, we tagged it with POS and lemma, using a POS tagger for spoken Spanish developed in our laboratory (Moreno & Guirao, 2003). We show the previous example after lemmatization and POS tagging. Lemmas are shown in uppercase.

Lemmatisation

<turn speaker="PAT" file="efamcv01"> <utt> Y CÓMO ESTAR </utt> </turn>

<turn speaker="ROS" file="efamcv01"> <utt> BUENO </utt> NO ESTAR MAL </utt> </turn>

POS tagging

<turn speaker="PAT" file="efamcv01"> <utt> C P AUX </utt> </turn>

<turn speaker="ROS" file="efamcv01"> <utt> MD </utt> ADV AUX ADV</utt> </turn>

C= conjunction; P = pronoun; AUX = auxiliary; MD: discursive marker; ADV= adverb

In summary, in this experiment we have considered three levels of linguistic data: words, lemmas and POS.

2.2 Extracting word clusters

If we calculate statistics directly on every unit, the result will not be correct, since multi-words units will not be included in this count. Discourse markers as frequent as "por ejemplo" (for instance), "es decir" (in other words) or "o sea" (that is) will not appear if we work on single word units. To solve this, we developed an algorithm based on n-grams in order to extract multi-word candidates. We took out all n-grams with three or more occurrences, for n = 4, 3, and 2. Next, a filter is applied for discarding all n-grams that start or end with a determiner or auxiliary. Finally, multi-words are selected by hand. Every multi-word is regarded as a lexical unit, equivalent to the simple/single words.

2.3 Applying the statistics of surprise

In order to identify the distinctive words, lemmas or POS for a given sub-corpus, we have employed the log-likelihood ratio test proposed by Dunning (1993). This method does not assume normal statistical distributions of units in a corpus. Instead, the log-likelihood ratio λ assumes a binomial distribution more appropriate for rare but distinctive words. "Texts are composed largely of such rare events" (Dunning, 1993). In addition, this test does not need balanced sub-corpora for comparison.

This method has been successfully applied for finding collocations (Dunning, 1993) and terms (Daille, 1994). In order to test the method for finding distinctive units in specific domains, we can work on two hypotheses:

i) Two registers (or sub-corpora) show no difference in distinctive units (*Null hypothesis*)

ii) For a given sub-corpus, we can find out distinctive units (*Alternative hypothesis*).

We applied the test to two well-defined sub-corpora, meteorological reports and law, in order to discard one of the hypotheses. Results are shown in Table 4. The critical value for one degree of freedom is 7.88.

Table 4: Dunning Test applied to well-defined sub-corpora

Meteo			Law		
-2 log λ	Freq	Lemmas	-2 log λ	Freq	Lemmas
165	50	norte	200	58	policía
160	38	fuerza	200	289	persona
145	27	viento	116	85	derecho
128	6466	en	113	45	contrato
97	16	componente	101	22	judicial
91	19	temperatura	84	31	delito
80	19	noroeste	83	26	delincuente
69	12	oeste	78	50	ley
69	12	nube	77	69	determinar
64	92	zona	65	17	cometer

Results confirm the alternative hypothesis and the suitability of the Dunning test for the task. Most of the "top 10" lemmas in both domains have a low occurrence, but all are typical terms in its domain.

3. Preliminary results

Our goal is to show a range of possibilities for the application of this method. We will show here a very incomplete set of data. Currently, there is a disproportion of social and register features with respect to annotated linguistic features. The linguistic annotation is being carried out this year (2003). Comprehensive results will be delivered with the final version of the four corpora, including a cross-linguistic comparison. In this paper, the only linguistic features that will be taken into account are:

words and multi-words
lemmas
POS tags.

First, we show the 10 most frequent word types in our corpus of spontaneous speech for three professions: a consumers' association manager, a football coach and a computer system administrator. Notice that words and multi-words are regarded as equivalent units.

Table 5: Most characteristic words in three professions

Consumers' association manager		Football coach		Computer system administrator	
-2 log λ		-2 log λ		-2 log λ	
110	establecimiento	31	directiva	91	tío
107	establecimientos	30	club	48	grabando
106	encuesta	28	Real Madrid	41	web
103	cesta	25	rueda de prensa	33	yo qué sé
65	política de precios	18	director general	33	no sé
45	marcas	17	no	30	joder
44	precios	16	estabilidad	30	linux
44	índice	16	confianza	28	cabrón
42	consumidor	16	hombre yo creo que	28	detectan
41	insisto	16	Y que	26	barato

The procedure can be extended to any profession registered in the corpus, as a means of detecting sociolectal information. Now we will provide the results of the Dunning test on formal and informal registers, approximately 150,000 words each:

Table 6: More characteristic words in formal and informal registers

Formal		Informal	
-2 log λ		-2 log λ	
134	de	422	s
91	es decir	279	a
62	su	238	sabes
61	en	231	claro
60	gobierno	231	me
55	en este momento	222	tía
49	desarrollo	182	yo
49	general	179	no sé
47	nuestra	173	no
46	países	171	ya

Another interesting comparison is to find out which POS are more typical in male and female registers. This table shows the results.

In our corpus, men prefer to use nouns and women prefer clearly pronouns.

Finally, after lemmatization, we can show the 10 most frequent verbs in general, male and female sub-corpora.

Table 7: POS in male and female registers

General		Male		Female	
Total occurrences		-2 log λ		-2 log λ	
47052	V	515	N	1327	P
42531	N	422	ADJ	524	ADV
38210	PREP	399	ART	243	C
32284	ART	382	PREP	149	MD
31404	ADV	80	DEM	126	INTJ
30737	C	34	Q	28	V
25044	P	34	REL	16	POSS
17418	AUX			16	AUX
12611	ADJ			4	NPR
10112	Q				

Table 8: Most frequent verb lemmas in male and female registers

General		Male		Female	
Total occurrences		-2 log λ		-2 log λ	
3398	ir	28	Escuchar	159	ir
2973	tener	28	Recordar	154	decir
2579	decir	26	Aparecer	112	saber
2067	hacer	23	Llegar	99	venir
1046	poder	23	contemplar	86	dar
1026	saber	23	caminar	47	mirar
995	ver	22	intentar	46	comprar
802	dar	22	Amar	42	gustar
779	querer	20	juntar	36	quedar
577	creer	20	superar	36	contar

4. Conclusions and future work

Here, we have shown the relevance of this procedure as an empirical method for the validation of sociolinguistic hypotheses in spoken language, as well as for determining register typology.

The method correlates linguistic with socio-contextual data applying Dunning's Statistics of Surprise. In order to achieve this, a rich linguistic-tagged corpus and the use of xml have been essential. The preliminary results are promising and have not been shown previously for Spanish. However, extracting conclusions and interpretations for these figures is premature, since the corpus is clearly not sufficient. For this reason, we will apply the method to CORLEC corpus, also developed by the LLI-UAM (see Moreno 2002 for an overview). The combination of C-ORAL-ROM and CORLEC corpora will contain over 1,500,000 words of spontaneous spoken Spanish.

We also wish to find out more about the correlation between linguistic and socio-contextual features, when complete morphosyntactic annotation will be finished. For instance, verb tenses, number, gender, persons in pronouns, etc, will be tagged. Biber (1988, 1995) provides a rich catalogue of linguistic features that can be traced.

Finally, a cross-linguistic comparison between the four Romance languages in C-ORAL-ROM will be made, based on the same text distribution.

References

Biber, D. (1988), *Variation across speech and writing*. Cambridge: CUP.

Biber, D. (1995), *Dimensions of register variation*. Cambridge: CUP.

Biber D., S. Johansson, G Leech, S. Conrad and E. Finegan (eds.) (1999), *The Longman grammar of spoken and written English*. London: Longman.

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.

Daille, B. (1994), *Combined approach for terminology extraction: lexical statistics and linguistic filtering*. Ph.D. Thesis, Paris 7.

Dunning T. (1993), 'Accurate methods for the statistics of surprise and coincidence'. *Computational Linguistics* 19(1): 61-74.

Labov W. (1966), *The social stratification of English in New York City*. Washington: Center for Applied Linguistics.

Miller J. and R. Weinert (1999) *Spontaneous spoken language*. Oxford: Clarendon.

Moreno A. (2002), 'La evolución de los corpus de habla espontánea: la experiencia del LLI-UAM', in: *Proceedings of II Jornadas de Tecnologías del Habla*, Granada, Spain.

Moreno A. and J. M. Guirao (2003), 'Tagging a spontaneous speech corpus of Spanish', in: *Proceedings of Recent Advances in NLP (RANLP-2003)* Borovets, Bulgaria.

Uchimoto K., C. Nobata, A. Yamada, S. Sekine and H. Isahara, (2002), 'Morphological Analysis of the Spontaneous Speech corpus'. In: *Proceedings of Conference of Computational Linguistics (COLING 2002)* Taipei, Taiwan