Studies in Corpus Linguistics

Studies in Corpus Linguistics aims to provide insights into the way a corpus can be used, the type of findings that can be obtained, the possible applications of these findings as well as the theoretical changes that corpus work can bring into linguistics and language engineering. The main concern of SCL is to present findings based on, or related to, the cumulative effect of naturally occuring language and on the interpretation of frequency and distributional data.

Volume 15
C-ORAL-ROM: Integrated Reference Corpora
for Spoken Romance Languages
Edited by Emanuela Cresti and Massimo Moneglia

C-ORAL-ROM

Integrated Reference Corpora for Spoken Romance Languages

Edited by
Emanuela Cresti
Massimo Moneglia
University of Florence

John Benjamins Publishing Company
Amsterdam/Philadelphia

Studies in Corpus Linguistics

Studies in Corpus Linguistics aims to provide insights into the way a corpus can be used, the type of findings that can be obtained, the possible applications of these findings as well as the theoretical changes that corpus work can bring into linguistics and language engineering. The main concern of SCL is to present findings based on, or related to, the cumulative effect of naturally occuring language and on the interpretation of frequency and distributional data.

Volume 15
C-ORAL-ROM: Integrated Reference Corpora
for Spoken Romance Languages
Edited by Emanuela Cresti and Massimo Moneglia

C-ORAL-ROM

Integrated Reference Corpora for Spoken Romance Languages

Edited by
Emanuela Cresti
Massimo Moneglia
University of Florence

John Benjamins Publishing Company
Amsterdam / Philadelphia

Cover illustration from original painting *Random Order* by Lorenzo Pezzatini, Florence, 1996.

Library of Congress Cataloging-in-Publication Data

C-ORAL-ROM : integrated reference corpora for spoken Romance languages / edited by Emanuela Cresti and Massimo Moneglia.
p. cm. (Studies in Corpus Linguistics, ISSN 1388–0373 ; v. 15)
Includes bibliographical references and index.
1. Romance languages--Data processing. 2. Romance languages--Variation. 3. Romance languages--Spoken Romance languages. I. Cresti, E. (Emanuela) II. Moneglia, Massimo. III. Series.

PC44.5.C2 2005
440'.0285--dc22 2005041056
ISBN 90 272 2286 X (Eur.) / 1 58811 548 8 (US) (Hb; alk. paper)

John Benjamins Publishing Co. · P.O. Box 36224 · 1020 ME Amsterdam · The Netherlands
John Benjamins North America · P.O. Box 27519 · Philadelphia PA 19118-0519 · USA

Table of contents

Cover illustration from original painting *Random Order* by Lorenzo Pezzatini, Florence, 1996.

Library of Congress Cataloging-in-Publication Data

C-ORAL-ROM : integrated reference corpora for spoken Romance languages / edited by Emanuela Cresti and Massimo Moneglia.
p. cm. (Studies in Corpus Linguistics, ISSN 1388–0373 ; v. 15)
Includes bibliographical references and index.
1. Romance languages--Data processing. 2. Romance languages--Variation. 3. Romance languages--Spoken Romance languages. I. Cresti, E. (Emanuela) II. Moneglia, Massimo. III. Series.

PC44.5.C2 2005
440'0285--dc22 2005041056
ISBN 90 272 2286 X (Eur.) / 1 58811 548 8 (US) (Hb; alk. paper)

John Benjamins Publishing Co. · P.O. Box 36224 · 1020 ME Amsterdam · The Netherlands
John Benjamins North America · P.O. Box 27519 · Philadelphia PA 19118-0519 · USA

Table of contents

∞™ The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

Cover illustration from original painting *Random Order* by Lorenzo Pezzatini, Florence, 1996.

Library of Congress Cataloging-in-Publication Data

C-ORAL-ROM : integrated reference corpora for spoken Romance languages / edited by Emanuela Cresti and Massimo Moneglia.
p. cm. (Studies in Corpus Linguistics, ISSN 1388–0373 ; v. 15)
Includes bibliographical references and index.
1. Romance languages--Data processing. 2. Romance languages--Variation. 3. Romance languages--Spoken Romance languages. I. Cresti, E. (Emanuela) II. Moneglia, Massimo. III. Series.

PC44.5.C2 2005
440'.0285--dc22 2005041056
ISBN 90 272 2286 X (Eur.) / 1 58811 548 8 (US) (Hb; alk. paper)

John Benjamins Publishing Co. · P.O. Box 36224 · 1020 ME Amsterdam · The Netherlands
John Benjamins North America · P.O. Box 27519 · Philadelphia PA 19118-0519 · USA

Table of contents

CHAPTER 4

The Spanish corpus

Antonio Moreno, Gillermo de la Madrid, Manuel Alcántara,
Ana Gonzalez, José M. Guirao, and Raúl De la Torre

4.1 History of the corpus in the national framework

4.1.1 Historical overview

The Corpus Oral de Referencia de la Lengua Española Contemporánea (CORLEC)
first spontaneous speech corpus for Spanish, compiled in the LLI-UAM (Laborat
de Lingüística Informática – Universidad Autónoma de Madrid) under the supervi
of Francisco Marcos Marín between 1991 and 1992 (Marcos 1992), is the refer
for the Spanish corpus for C-ORAL-ROM. However, only one text from the ori
CORLEC is included in C-ORAL-ROM. The reasons for this virtually new compil
will be addressed in this section. The historical evolution, both in methodology
format, between CORLEC and C-ORAL-ROM will also be outlined. In addition,
similar corpora in Spanish will be compared with those compiled at LLI-UAM.

The temporal distance between the two corpora mentioned is a decade, fron
beginning of the 1990s to the beginning of the 21st century. This time gap al
us to present a historical overview of the progess in knowledge and of those as
that remain to be improved. These corpora are comparable in several aspects
same language, and the same laboratory (though with different transcribers).
all, both share the same goal: to accurately register contemporary spoken varia
produced in spontaneous situations. The methodology has stayed the same, in
ing monologue and conversation recordings in real scenes, without a pre-establ
script and with no restriction in expression for participants. Both corpora use a
thographic transcription following Spanish written conventions. Finally, both co
are reference corpora, that is to say, they are created to be used by the scientific
munity. Their public nature adds to the transcription accuracy requirement, since
accuracy may be contrasted with the original sound source.

Differences between these two corpora may be also noted, but only as far as
historical aspects are concerned. As we will see, these differences are basicall
to the budget and the technology available at the time each was compiled; imp
ments in the later corpus were possible as a result of experience and contact with

The LLI-UAM is a pioneering group in the creation of spoken corpora for the Spanish language in Spain. A comprehensive account of the state of the art in the 1990s may be found in the work of Llisterri (1997).

1. CORLEC (Corpus Oral de Referencia de la Lengua Española Contemporánea), funded by IBM, was the first spoken language corpus for the Spanish language. It was compiled under the supervision of Prof. Marcos Marín. The transcription and mark-up scheme was taken from TEI (Text Encoding Initiative). This corpus was encoded and revised in SGML in 1997 by the Real Academia Española group to be included in the Corpus de Referencia del Español Actual (CREA). The corpus is freely available through the LLI-UAM FTP server (ftp://ftp.lllf.uam.es/pub/corpus/oral/) and also through the CREA retrieval service (http://www.rae.es/).
   Among other spoken corpora compiled in Spain in the 1990s, we may find ALBAYZIN, the Corpus de Conversación Coloquial of the Universidad de Valencia (VALESCO) and the Corpus de Variedades Vernáculas Malagueñas of the Universidad de Málaga (VUM).
2. ALBAYZIN is a phonetic database of around 7,000 sentences from 300 speakers. It was developed in the beginning of the 1990s and its commitment was to create a database for speech recognition and for the testing of phonetic transcription systems. The main difference as regards corpora compiled within the LLI is that ALBAYZIN is not a corpus of spontaneous speech, but of 'phonetically balanced' sentences, that is to say, sentences designed to clearly distinguish phonemes.
3. VALESCO[1] (Corpus de Conversación Coloquial of the Universidad de Valencia) was developed with the aim of studying pragmatic aspects of the Spanish colloquial language. For this goal, the authors collected and transcribed a corpus of conversations. This corpus is clearly different to our corpora. Firstly, VALESCO is not designed as a reference corpus (that is to say, to be used by other research groups), but as a corpus for their own research. Furthermore, it does not include as much register variation as CORLEC or C-ORAL-ROM, where an important part is committed to formal registers (monologues, conferences, and sermons). Finally, there is no use of computer tools for linguistic mark-up in VALESCO.
4. VUM (Corpus de Variedades Vernáculas Malagueñas of the Universidad de Málaga) is one of the first dialectal spoken corpora. Its goal is clearly sociolinguistic and phonetic, attempting to register the phonetic characteristics of Southern Spanish speech. This corpus is based on similar criteria adopted by CORLEC.
5. CLUVI (Corpus Linguistico de la Universidad de Vigo), funded by the Plan Nacional de I+D+I, is a recent project being developed by the Seminario de Lingüística Informática of the Universidad de Vigo,[2] and is similar to that of LLI-UAM. The main difference lies in the fact that their corpus is based on 5 subcorpora, two of which are dedicated to oral language. One of them is a corpus of bilingual (Spanish-Galician) spontaneous dialogues and the other is a corpus of Galician in the media. In this sense, the LLI-UAM corpus and that of SLI-UVI are c
   plementary.

C-ORAL-ROM is a "second generation" spoken corpus (Moreno 2002), since it in
porates innovations such as the consent forms signed by speakers who were recor
internal and external validation of the transcription and mark-up, exceptional acc
tic quality thanks to digital recording, and the use of XML mark-up language. In or
to make a comparison with previous approaches, we will now review CORLEC,
antecedent of C-ORAL-ROM.

4.1.2 CORLEC features

CORLEC is a database comprising around 1,100,000 transcribed words from spo
texts recorded on analogue audio tapes. The methodology consisted in carrying
recordings in their real contexts, usually without knowledge and permission of par
ipants. The transcription was made using an analogue recorder with headphones a
writing directly onto the word processor (WordPerfect). Digital technology was n
ther used in the recording nor in the later treatment of the data.[3] The limitations
this first generation methodology are noticeable: acoustic quality is usually deficie
and there is no alignment between the original sound and its transcription. Nevert
less, it must be kept in mind that the main goal of this corpus was to accurately regi
Spanish spoken varieties for the first time.[4]

As concerns transcription criteria, the most important feature is the accur
of what participants say: deleted phonetic segments, breaks, repeated occurren
self corrections, invented words or other languages are transcribed precisely as p
nounced by the speaker. For the retrieval of canonical forms, all these cases
marked-up with relevant tags.

Another transcription criterion is the use of punctuation marks (inverted c
mas, ellipsis, full stops, etc.) in order to mark discursive situations. Inverted com
are used to highlight words and mark titles in direct discourse. Ellipsis marks are u
to mark breaks, hesitations, sudden breaks. Commas and full stops are used as s
tactic unit markers. As a general rule, the transcriber was required to follow spel
rules for written texts: for instance, a pause must be marked even if the speaker d
not pause at the end of a sentence (Marcos Marín 1992). This decision is probably
most contradictory to the one just described before: on the one hand, there is
nunciation accuracy, but on the other, spelling rules are followed as regards wri
syntax. In contrast, in C-ORAL-ROM we have rejected the use of punctuation m
according to written language conventions.

The information provided in the transcription is enriched with a variety of
for phatic elements (sounds emitted by speakers that are interpreted as assertions
terrogations, etc.), noises (laughs, applauses, music, etc.), and, especially, discu
interactions: turn-taking and overlapping of participants are marked.

The LLI-UAM is a pioneering group in the creation of spoken corpora for the Spanish language in Spain. A comprehensive account of the state of the art in the 1990s may be found in the work of Llisterri (1997).

1. CORLEC (Corpus Oral de Referencia de la Lengua Española Contemporánea), funded by IBM, was the first spoken language corpus for the Spanish language. It was compiled under the supervision of Prof. Marcos Marín. The transcription and mark-up scheme was taken from TEI (Text Encoding Initiative). This corpus was encoded and revised in SGML in 1997 by the Real Academia Española group to be included in the Corpus de Referencia del Español Actual (CREA). The corpus is freely available through the LLI-UAM FTP server (ftp://ftp.lllf.uam.es/pub/corpus/oral/) and also through the CREA retrieval service (http://www.rae.es/).
   Among other spoken corpora compiled in Spain in the 1990s, we may find ALBAYZIN, the Corpus de Conversación Coloquial of the Universidad de Valencia (VALESCO) and the Corpus de Variedades Vernáculas Malagueñas of the Universidad de Málaga (VUM).
2. ALBAYZIN is a phonetic database of around 7,000 sentences from 300 speakers. It was developed in the beginning of the 1990s and its commitment was to create a database for speech recognition and for the testing of phonetic transcription systems. The main difference as regards corpora compiled within the LLI is that ALBAYZIN is not a corpus of spontaneous speech, but of 'phonetically balanced' sentences, that is to say, sentences designed to clearly distinguish phonemes.
3. VALESCO[1] (Corpus de Conversación Coloquial of the Universidad de Valencia) was developed with the aim of studying pragmatic aspects of the Spanish colloquial language. For this goal, the authors collected and transcribed a corpus of conversations. This corpus is clearly different to our corpora. Firstly, VALESCO is not designed as a reference corpus (that is to say, to be used by other research groups), but as a corpus for their own research. Furthermore, it does not include as much register variation as CORLEC or C-ORAL-ROM, where an important part is committed to formal registers (monologues, conferences, and sermons). Finally, there is no use of computer tools for linguistic mark-up in VALESCO.
4. VUM (Corpus de Variedades Vernáculas Malagueñas of the Universidad de Málaga) is one of the first dialectal spoken corpora. Its goal is clearly sociolinguistic and phonetic, attempting to register the phonetic characteristics of Southern Spanish speech. This corpus is based on similar criteria adopted by CORLEC.
5. CLUVI (Corpus Lingüístico de la Universidad de Vigo), funded by the Plan Nacional de I+D+I, is a recent project being developed by the Seminario de Lingüística Informática of the Universidad de Vigo,[2] and is similar to that of LLI-UAM. The main difference lies in the fact that their corpus is based on 5 subcorpora, two of which are dedicated to oral language. One of them is a corpus of bilingual (Spanish-Galician) spontaneous dialogues and the other is a corpus of Galician in the media. In this sense, the LLI-UAM corpus and that of SLI-UVI are complementary.

C-ORAL-ROM is a "second generation" spoken corpus (Moreno 2002), since it incorporates innovations such as the consent forms signed by speakers who were recorde internal and external validation of the transcription and mark-up, exceptional acou tic quality thanks to digital recording, and the use of XML mark-up language. In ord to make a comparison with previous approaches, we will now review CORLEC, t antecedent of C-ORAL-ROM.

4.1.2 CORLEC features

CORLEC is a database comprising around 1,100,000 transcribed words from spok texts recorded on analogue audio tapes. The methodology consisted in carrying recordings in their real contexts, usually without knowledge and permission of part ipants. The transcription was made using an analogue recorder with headphones a writing directly onto the word processor (WordPerfect). Digital technology was r ther used in the recording nor in the later treatment of the data.[3] The limitation this first generation methodology are noticeable: acoustic quality is usually deficie and there is no alignment between the original sound and its transcription. Nevert less, it must be kept in mind that the main goal of this corpus was to accurately regi Spanish spoken varieties for the first time.[4]

As concerns transcription criteria, the most important feature is the accur of what participants say: deleted phonetic segments, breaks, repeated occurren self corrections, invented words or other languages are transcribed precisely as p nounced by the speaker. For the retrieval of canonical forms, all these cases marked-up with relevant tags.

Another transcription criterion is the use of punctuation marks (inverted co mas, ellipsis, full stops, etc.) in order to mark discursive situations. Inverted com are used to highlight words and mark titles in direct discourse. Ellipsis marks are u to mark breaks, hesitations, sudden breaks. Commas and full stops are used as tactic unit markers. As a general rule, the transcriber was required to follow spel rules for written texts: for instance, a pause must be marked even if the speaker not pause at the end of a sentence (Marcos Marín 1992). This decision is probably most contradictory to the one just described before: on the one hand, there is nunciation accuracy, but on the other, spelling rules are followed as regards wr syntax. In contrast, in C-ORAL-ROM we have rejected the use of punctuation m according to written language conventions.

The information provided in the transcription is enriched with a variety of for phatic elements (sounds emitted by speakers that are interpreted as assertion terrogations, etc.), noises (laughs, applauses, music, etc.), and, especially, discu interactions: turn-taking and overlapping of participants are marked.

The LLI-UAM is a pioneering group in the creation of spoken corpora for the Spanish language in Spain. A comprehensive account of the state of the art in the 1990s may be found in the work of Llisterri (1997).

1. CORLEC (Corpus Oral de Referencia de la Lengua Española Contemporánea), funded by IBM, was the first spoken language corpus for the Spanish language. It was compiled under the supervision of Prof. Marcos Marín. The transcription and mark-up scheme was taken from TEI (Text Encoding Initiative). This corpus was encoded and revised in SGML in 1997 by the Real Academia Española group to be included in the Corpus de Referencia del Español Actual (CREA). The corpus is freely available through the LLI-UAM FTP server (ftp://ftp.lllf.uam.es/pub/corpus/oral/) and also through the CREA retrieval service (http://www.rae.es/).
   Among other spoken corpora compiled in Spain in the 1990s, we may find ALBAYZIN, the Corpus de Conversación Coloquial of the Universidad de Valencia (VALESCO) and the Corpus de Variedades Vernáculas Malagueñas of the Universidad de Málaga (VUM).
2. ALBAYZIN is a phonetic database of around 7,000 sentences from 300 speakers. It was developed in the beginning of the 1990s and its commitment was to create a database for speech recognition and for the testing of phonetic transcription systems. The main difference as regards corpora compiled within the LLI is that ALBAYZIN is not a corpus of spontaneous speech, but of 'phonetically balanced' sentences, that is to say, sentences designed to clearly distinguish phonemes.
3. VALESCO[1] (Corpus de Conversación Coloquial of the Universidad de Valencia) was developed with the aim of studying pragmatic aspects of the Spanish colloquial language. For this goal, the authors collected and transcribed a corpus of conversations. This corpus is clearly different to our corpora. Firstly, VALESCO is not designed as a reference corpus (that is to say, to be used by other research groups), but as a corpus for their own research. Furthermore, it does not include as much register variation as CORLEC or C-ORAL-ROM, where an important part is committed to formal registers (monologues, conferences, and sermons). Finally, there is no use of computer tools for linguistic mark-up in VALESCO.
4. VUM (Corpus de Variedades Vernáculas Malagueñas of the Universidad de Málaga) is one of the first dialectal spoken corpora. Its goal is clearly sociolinguistic and phonetic, attempting to register the phonetic characteristics of Southern Spanish speech. This corpus is based on similar criteria adopted by CORLEC.
5. CLUVI (Corpus Lingüístico de la Universidad de Vigo), funded by the Plan Nacional de I+D+I, is a recent project being developed by the Seminario de Lingüística Informática of the Universidad de Vigo,[2] and is similar to that of LLI-UAM, The main difference lies in the fact that their corpus is based on 5 subcorpora, two of which are dedicated to oral language. One of them is a corpus of bilingual (Spanish-Galician) spontaneous dialogues and the other is a corpus of Galician

in the media. In this sense, the LLI-UAM corpus and that of SLI-UVI are plementary.

C-ORAL-ROM is a "second generation" spoken corpus (Moreno 2002), since it i porates innovations such as the consent forms signed by speakers who were reco internal and external validation of the transcription and mark-up, exceptional ac tic quality thanks to digital recording, and the use of XML mark-up language. In o to make a comparison with previous approaches, we will now review CORLEC antecedent of C-ORAL-ROM.

4.1.2 CORLEC features

CORLEC is a database comprising around 1,100,000 transcribed words from spo texts recorded on analogue audio tapes. The methodology consisted in carrying recordings in their real contexts, usually without knowledge and permission of pa ipants. The transcription was made using an analogue recorder with headphones writing directly onto the word processor (WordPerfect). Digital technology was ther used in the recording nor in the later treatment of the data.[3] The limitation this first generation methodology are noticeable: acoustic quality is usually defic and there is no alignment between the original sound and its transcription. Never less, it must be kept in mind that the main goal of this corpus was to accurately reg Spanish spoken varieties for the first time.[4]

As concerns transcription criteria, the most important feature is the accu of what participants say: deleted phonetic segments, breaks, repeated occurren self corrections, invented words or other languages are transcribed precisely as nounced by the speaker. For the retrieval of canonical forms, all these cases marked-up with relevant tags.

Another transcription criterion is the use of punctuation marks (inverted c mas, ellipsis, full stops, etc.) in order to mark discursive situations. Inverted com are used to highlight words and mark titles in direct discourse. Ellipsis marks are u to mark breaks, hesitations, sudden breaks. Commas and full stops are used as s tactic unit markers. As a general rule, the transcriber was required to follow spel rules for written texts: for instance, a pause must be marked even if the speaker d not pause at the end of a sentence (Marcos Marín 1992). This decision is probably most contradictory to the one just described before: on the one hand, there is p nunciation accuracy, but on the other, spelling rules are followed as regards wri syntax. In contrast, in C-ORAL-ROM we have rejected the use of punctuation m according to written language conventions.

The information provided in the transcription is enriched with a variety of for phatic elements (sounds emitted by speakers that are interpreted as assertions terrogations, etc.), noises (laughs, applauses, music, etc.), and, especially, discu interactions: turn-taking and overlapping of participants are marked.

The LLI-UAM is a pioneering group in the creation of spoken corpora for the Spanish language in Spain. A comprehensive account of the state of the art in the 1990s may be found in the work of Llisterri (1997).

1. CORLEC (Corpus Oral de Referencia de la Lengua Española Contemporánea), funded by IBM, was the first spoken language corpus for the Spanish language. It was compiled under the supervision of Prof. Marcos Marín. The transcription and mark-up scheme was taken from TEI (Text Encoding Initiative). This corpus was encoded and revised in SGML in 1997 by the Real Academia Española group to be included in the Corpus de Referencia del Español Actual (CREA). The corpus is freely available through the LLI-UAM FTP server (ftp://ftp.lllf.uam.es/pub/corpus/oral/) and also through the CREA retrieval service (http://www.rae.es/).
   Among other spoken corpora compiled in Spain in the 1990s, we may find ALBAYZIN, the Corpus de Conversación Coloquial of the Universidad de Valencia (VALESCO) and the Corpus de Variedades Vernáculas Malagueñas of the Universidad de Málaga (VUM).
2. ALBAYZIN is a phonetic database of around 7,000 sentences from 300 speakers. It was developed in the beginning of the 1990s and its commitment was to create a database for speech recognition and for the testing of phonetic transcription systems. The main difference as regards corpora compiled within the LLI is that ALBAYZIN is not a corpus of spontaneous speech, but of 'phonetically balanced' sentences, that is to say, sentences designed to clearly distinguish phonemes.
3. VALESCO[1] (Corpus de Conversación Coloquial of the Universidad de Valencia) was developed with the aim of studying pragmatic aspects of the Spanish colloquial language. For this goal, the authors collected and transcribed a corpus of conversations. This corpus is clearly different to our corpora. Firstly, VALESCO is not designed as a reference corpus (that is to say, to be used by other research groups), but as a corpus for their own research. Furthermore, it does not include as much register variation as CORLEC or C-ORAL-ROM, where an important part is committed to formal registers (monologues, conferences, and sermons). Finally, there is no use of computer tools for linguistic mark-up in VALESCO.
4. VUM (Corpus de Variedades Vernáculas Malagueñas of the Universidad de Málaga) is one of the first dialectal spoken corpora. Its goal is clearly sociolinguistic and phonetic, attempting to register the phonetic characteristics of Southern Spanish speech. This corpus is based on similar criteria adopted by CORLEC.
5. CLUVI (Corpus Linguístico de la Universidad de Vigo), funded by the Plan Nacional de I+D+I, is a recent project being developed by the Seminario de Lingüística Informática of the Universidad de Vigo,[2] and is similar to that of LLI-UAM. The main difference lies in the fact that their corpus is based on 5 subcorpora, two of which are dedicated to oral language. One of them is a corpus of bilingual (Spanish-Galician) spontaneous dialogues and the other is a corpus of Galician in the media. In this sense, the LLI-UAM corpus and that of SLI-UVI are complementary.

C-ORAL-ROM is a "second generation" spoken corpus (Moreno 2002), since it incorporates innovations such as the consent forms signed by speakers who were recorded, internal and external validation of the transcription and mark-up, exceptional acoustic quality thanks to digital recording, and the use of XML mark-up language. In order to make a comparison with previous approaches, we will now review CORLEC, the antecedent of C-ORAL-ROM.

4.1.2 CORLEC features

CORLEC is a database comprising around 1,100,000 transcribed words from spoken texts recorded on analogue audio tapes. The methodology consisted in carrying out recordings in their real contexts, usually without knowledge and permission of participants. The transcription was made using an analogue recorder with headphones and writing directly onto the word processor (WordPerfect). Digital technology was neither used in the recording nor in the later treatment of the data.[3] The limitations of this first generation methodology are noticeable: acoustic quality is usually deficient, and there is no alignment between the original sound and its transcription. Nevertheless, it must be kept in mind that the main goal of this corpus was to accurately register Spanish spoken varieties for the first time.[4]

As concerns transcription criteria, the most important feature is the accuracy of what participants say: deleted phonetic segments, breaks, repeated occurrences, self corrections, invented words or other languages are transcribed precisely as pronounced by the speaker. For the retrieval of canonical forms, all these cases are marked-up with relevant tags.

Another transcription criterion is the use of punctuation marks (inverted commas, ellipsis, full stops, etc.) in order to mark discursive situations. Inverted commas are used to highlight words and mark titles in direct discourse. Ellipsis marks are used to mark breaks, hesitations, sudden breaks. Commas and full stops are used as syntactic unit markers. As a general rule, the transcriber was required to follow spelling rules for written texts: for instance, a pause must be marked even if the speaker does not pause at the end of a sentence (Marcos Marín 1992). This decision is probably the most contradictory to the one just described before: on the one hand, there is pronunciation accuracy, but on the other, spelling rules are followed as regards written syntax. In contrast, in C-ORAL-ROM we have rejected the use of punctuation marks according to written language conventions.

The information provided in the transcription is enriched with a variety of tags for phatic elements (sounds emitted by speakers that are interpreted as assertions, interrogations, etc.), noises (laughs, applauses, music, etc.), and, especially, discursive interactions: turn-taking and overlapping of participants are marked.

Table 4.1 Distribution of CORLEC

Classes	Number of words	Percentage
Administrative and political	61,200	5.6%
Advertising	30,800	2.8%
Debates	93,500	8.5%
Documentary	28,600	2.6%
Educational	58,300	5.3%
Familiar	269,500	24.5%
Humanistic	61,200	5.6%
Instructions (megaphony)	6,600	0.6%
Interviews	171,200	15.6%
Journalistic		
Legal	35,200	3.2%
Ludic (competitions, etc.)	61,200	5.6%
News	72,600	6.6%
Religious	12,100	1.1%
Scientific	36,600	3.3%
Sports	58,300	5.3%
Technical	43,100	3.9%
Total estimated	1,100,000	100.0%

The encoding system used follows TEI norms but, since the corpus was compiled when the TEI norms were not yet published, the authors made particular encoding decisions which differ from those norms. It is also important to highlight that this corpus has not passed any format validation process (for example, validation by means of DTD).

The distribution of different kind of texts is one of the distinctive components of any corpus. CORLEC is organised by thematic criteria, as illustrated in Table 4.1. The most important texts are journalistic and familiar, comprising 38.6% and 24.5% of the corpus, respectively. However, the kind of communication event (monologue face to dialogue/conversation) and the channel (direct recording through the media, the telephone) are not representative of all possible discourse types. The most distinctive feature in the distribution of texts types in C-ORAL-ROM is the distinction between formal and informal registers, that was unmarked in CORLEC.

The complete transcription of CORLEC is freely available through the LLI-UAM FTP server (http://www.lllf.uam.es/).

4.1.3 C-ORAL-ROM features[5]

The main difference between C-ORAL-ROM and other corpora is its multilingual feature. On the one hand, this results in an enrichment due to the experience of other groups and the exchange of ideas. On the other hand, it required the commit-

Obtaining a similar distribution in the different corpora is essential in order to compare the final results in the different languages. Together with the difficulty of designing a significant distribution for a spoken corpus, due to the inherent variability of spoken language, there is a difficulty in joining together different approaches to the transcription. In C-ORAL-ROM we have reached the following commitment: the two factors that influence decisively the variation are the communication event and the register (but not the theme, which was the organisation axis of CORLEC).

Another aspect needed to obtain comparability is to use the same mark-up scheme. The consortium agreed to use a C-ORAL-ROM format based on the Italian model (whose origin is the CHAT format). To get full reuse of the transcription, the mark-up language is XML, assuring an easy interpretation (by means of the corresponding DTD). The LLI-UAM has developed for the project the format conversion software to convert C-ORAL-ROM to XML.[6]

Other innovative aspects of C-ORAL-ROM compared to CORLEC are elaborated on below:

1. *Legality of texts.* We have at our disposal the signed consent forms of all participants. This requirement is compulsory, because the law has changed and the lack of this permission affects royalties (in the case of conferences and media recordings) and privacy rights. This fact has been a new experience for all the teams, because no written permission was needed in previous corpora; scientists do not normally worry about legal questions, only about the diffusion of knowledge. We became aware of this regulation thanks to ELRA, the partner that will lead the distribution and commercialisation of the corpus.
2. *Validation.* Validation has come back into fashion in recent years for any linguistic resource in electronic format: final users of corpora demand reliability of the collected data. C-ORAL-ROM presents different levels of validation. The most important one is internal validation. Every text must pass through five stages: transcription, revision, prosodic annotation, revision of annotation, and text-sound alignment; with each stage performed by a different linguist. As a consequence, at least three researchers intervene in each text. Comparing this to CORLEC, where the same researcher both transcribed and revised texts, reliability has increased remarkably. Furthermore, software developed within the LLI for the project tests format errors (e.g. missing tags, printing mistakes, blank spaces not allowed, etc.). By means of this testing which is necessary for the conversion into XML format, it is possible to unify all texts in the four corpora. Finally, a group of international experts made an external validation in order to increase the reliability of the corpus.
3. *The use of XML.* We exploit the universality of this mark-up language that guarantees the reusability of the corpus. In fact, if the aim is to create a corpus to be used as a reference by the linguistic and voice recognition communities, it is necessary to work with a format valid in all the communities concerned. XML and its exten-

sions have become the format for linguistic technologies, since XML-encoded text can be converted to any other text formats.

4. *Digital quality of recordings.* The other teams of the project have reused in part previous analogue recordings. In our case, however, as we did not have written permission for previous texts, where the sound quality testing was in any case discouraging, we started recording afresh using a digital recorder that affords excellent quality. This has meant much more work for the Spanish team, but now two corpora are at the disposal of LLI.
5. *Linguistic information tagging.* Different linguistic levels are marked-up in C-ORAL-ROM. The basic one is the prosodic level, which has been completed for all texts. Additionally, a significant part of each corpus will be marked up morpho-syntactically, including verb, noun and adjective lemmatisation.
6. *Sound-transcription alignment.* This characteristic is not provided in CORLEC because at that time technology to fulfil such requirements, i.e. digitalised sound and software to develop the alignment, was not available at an affordable price for the project. For the alignment in C-ORAL-ROM, we have used a version of *Winpitch*, a tool specifically developed for corpus transcription. We would like to emphasise that this feature provides us with a considerable empirical added value: lack of correspondence between text and audio is revealed clearly by the aligned version.

4.1.4 Final remarks

This section has shown the clear evolution in the spoken corpora of the LLI, keeping in mind that they still maintain the same basis: the registry of spontaneous spoken language in real contexts. Other aspects that still remain are text features, containing not only a transcription but also a header with rich information regarding the text. All the relevant information must be tagged by using a mark-up language that allows its identification in a clear and unequivocal way. The choice of the encoding scheme must be based on standardised criteria, since it is the only effective way to make a reference corpus for use by other researchers.

Nevertheless, there are many changes as a product of experience and technology breakthroughs, as well as of budget and legislation. Second generation corpora must provide a validated quality, both in transcription/annotation and sound source reliability, that must necessarily be given in digital format. As a consequence, the alignment of text and sound must be provided now to allow the empirical verification of the transcription accuracy. As far as a spontaneous spoken language corpus is to be used freely by the scientific community, the authorisation by participants is required, so that both their privacy and copyright are protected.

4.2 Orthographic transcription

4.2.1 General criteria

As far as the transcription of the sound files is concerned, we have respected the rules set by the Real Academia Española. Here we show some of these rules which we thought we should state together with the decisions the group took in order to maintain coherence.

1. *Acronyms and symbols.* Acronyms are given in block capitals and without full stops, as in *IVA, ONG, NIF*; the plural suffix for acronyms is in lower case, as in *ONGs*. Other corpora and the internet have been useful in order to determine the most frequent orthography of specific acronyms. In the case of symbols related to science (chemistry, etc.) we have followed the conventions. The letter *x* was used in certain contexts to make reference to a non-specific quantity. No abbreviations were used in the transcription of the sound files.
2. *New words.* We included words which, although not included in the Real Academia Española dictionaries, have a high frequency of occurrence in spoken Spanish, such as *porfa, finde* or *pafeto*.
3. *Numbers.* Numbers were transcribed in letters, except in the cases of numbers which are part of proper nouns, as in *La 2*, and numbers included in mathematical formulae. Roman numbers were used when referring to popes or kings, as in *Juan XXIII*, as well as in names in which they are included, as *N-III*.
4. *Capital letters.* Initial capital letters were used when transcribing proper nouns which made reference to people (including nicknames), as *Inma* or *el Bibi*, as well as cities, countries, towns, regions, districts, squares, streets and so on, as in *Segovia, Carabanchel* or *Madrid*. The same was applied to names of institutions, entities, organisations, political parties, etc., as in the case of *Comunidad de Madrid, Ministerio de Hacienda, la Politécnica*. Initial capital letters were also used when naming scientific disciplines, as well as entities which are considered absolute concepts and religious concepts, such as *la Sociología, Internet, la Humanidad, tu Reino* y *el Universo*, while in the case of *Señor salvador* the second word, being an adjective, is not given an initial capital. Names of sports competitions were transcribed with initial capital letters for content words, as in *Copa del Rey* or *Champions League*. In the case of books and song titles, as well as all kinds of works of art, even television programmes, only the first word was given an initial capital letter, except in cases like *Las Meninas* (as is conventional). Both nouns and adjectives included in the names of newspapers, magazines and such were written with initial capital letters, as in *El Mundo* or *El País*. The names of stores and commercial brands as *El Corte Inglés* were transcribed following the registered name.
5. *Italics.* No italics were used in the transcription of the sound files.
6. *Foreign words.* Foreign words were transcribed as in the original language. Words of foreign origin were written with the original orthography when pronounced

in that language, whether they were included in the academic dictionaries or not. When adapted to Spanish, these words were transcribed following the rules set in these dictionaries.

4.2.2 Orthography for non-standard words

Non-standard productions have been labelled in C-ORAL-ROM in the %alt dependent lines. For example, the following non-standard productions have been transcribed following orthography:[7]

standard forms	non-standard forms
abris	abréis
básico	básimoco
a El	al
casa	ca
claro	cao
claro	caro
cassette	casé

4.2.3 Interjections

A list of interjections was created during the process of transcription. These words were transcribed with exclamation marks, which are present in the transcription only in these cases. The list is as follows:

¡ah!	¡ahí va!
¡anda!	¡bah!
¡brum!	¡buah!
¡bueno!	¡cachis en la mar!
¡chun!	¡coño!
¡Dios mío de mi alma!	¡eh!
¡ey!	¡hala!
¡hombre!	¡jo!
¡jobar!	¡joder!
¡jolines!	¡leche!
¡madre!	¡madre del amor hermoso!
¡madre mía de mi vida y mi corazón!	¡madre mía!
¡mua!	¡oh!
¡ojo!	¡ole!
¡ostras!	¡ouh!
¡oy!	¡por Dios!
¡pum!	¡uh!
¡ms!	¡yeah!

4.3 Morpho-syntactic tagging

4.3.1 Tools and strategy adopted for automatic PoS tagging and lemmatisation

With respect to the linguistic annotation, the main goal is to provide a complete morphological and PoS tagging, including lemmatisation. These tasks have been performed automatically and validated by expert annotators.

For the morphological analysis we use GRAMPAL (Moreno 1991; Moreno & Goñi 1995) which is based on a rich morpheme lexicon of over 40,000 lexical units, and morphological rules. This system has been successfully used in language engineering applications as ARIES (Goñi et al. 1997) and also in linguistic description (Moreno & Goñi 2002). Originally, GRAMPAL was developed for analysing written texts. The tagging has been the most useful test for showing the ability of GRAMPAL to deal with a wide-coverage corpus of Spanish. We use this application for enhancing GRAMPAL with new modules: a PoS tagger and an unknown words recogniser, both specifically developed for spoken Spanish.

GRAMPAL is theoretically based on feature unification grammars and originally implemented in Prolog. The system is reversible: the same set of rules and the same lexicon are used for both analysis and generation of inflected wordforms. It is designed to allow only grammatical forms. In other words, the most salient feature of this model is its linguistic rigour, which avoids both over-acceptance and over-generation.

The analysis provides a full set of morpho-syntactic information in terms of features: lemma, PoS, gender, number, tense, mood, etc. In order to be suitable for tagging the C-ORAL-ROM corpus, a number of developments has been introduced in GRAMPAL, reported in Moreno and Guirao (2003):

1. A new tokenisation for the spoken corpus.
2. A set of rules for derivative morphology

Tokenisation in spoken corpora is slightly different to the same task in written corpora. Neither sentence nor paragraph boundaries make sense in spontaneous speech. Instead, dialogue turns and prosodic tags are used for identifying utterance boundaries.

For disambiguation, specific features of spoken corpora directly affect tagging: repetition and retracting produce agrammatical sequences; fragments that are not full sentences appear very often; and there is a more relaxed word order. Finally, there are no punctuation marks. All those characteristics force the PoS tagger, which is typically trained for written texts, to adapt.

Fortunately Proper Names recognition is not a problem for C-ORAL-ROM since only names are transcribed with an initial capital letter. As a consequence, analysing them is a trivial task.

On the lexical side, we detected two specific features as compared to written corpora: there is a low presence of new terms (i.e. most vocabulary used by speakers in spontaneous conversations is common and basic); and there is a high frequency of

derivative prefixes and suffixes that do not change the syntactic category, because most of them are appreciative morphemes (for instance, diminutives).

In order to handle the recognition of derivatives, GRAMPAL has been extended with derivation rules. The Prefix rule is: *take any Prefix and any (inflected) word and form another word with the same features.*

This rule is effective for PoS tagging since in Spanish prefixes never change the syntactic category of the base. The rule assigns the category feature to the unknown word. 239 prefixes have been added to the GRAMPAL lexicon.

PoS disambiguation has been solved using a rule-based model, in particular, an extension of a Constraint Grammar using features in a Context-Sensible PS. The output of the tagger is a feature structure written in XML.

The formalism allows several types of context-sensitive rules. First in application are the lexical ones: those for a particular ambiguous word, as follows:

"word" → <cat="X"> / _ <cat ="Y">
"word" → <cat="Z"> / <cat ="W"> _

where a given ambigous word is assigned a category X before a word with a category Y, or the category Z after a category W.

Any kind of feature can be taken into account, not only the category. For instance, we can face the problem of two ambiguous verbs belonging to different lemmas, as follows:

"word" → <lemma = "L"> / _ <cat="X">
"word" → <lemma = "M"> / _ <cat="Y">

In addition to features, strings and punctuation marks can be specified in the RHS of the context sensitive rule as follows:

"word" → <cat="X"> / *string* _
"word" → <cat="Y"> / # _

where *string* is any token, and # is the symbol for start or end of utterance.

If no lexical rule exists for a given ambiguity, then more general, syntactic rules are applied:

<cat="N">,<cat="V"> → <cat="N"> / <cat="V"> _

where if a given word is analysed with two different tags, one as a noun (N), and the other as a verb (V), then the one with category N is chosen if it appears after a word with category V. In short, those syntactic rules apply when there is no a specific rule for the case, either because it is a new ambiguity not covered by the grammar, or because the grammar writer did not find a proper way to describe the ambiguity.

This method benefits from the fact that most frequent ambiguities for a given language become well-known after a training. As a consequence, many context-sensitive lexical rules can be written by hand or extracted automatically from the data. Figures

Finally, those words which did not undergo disambiguation are treated with TNT (Brants 2000), which assigns PoS following a statistic model obtained from a 50,000-word training corpus. It has been shown that TNT is the most precise statistical tagger (Bigert et al. 2003).

4.3.1.1 *Electronic vocabulary*

The GRAMPAL lexicon is a collection of allomorphs for both stems and endings. New additions can be easily incorporated, since every possibility in Spanish inflection has been classified in a particular class.

In an experiment reported in Moreno and Guirao (2003), 8% of the whole corpus are unknown words for the system, comprising the following categories:

1. Foreign words: *walkman, parking*
2. Missing words in the lexicon, typically from the spoken language: *caramba, hijoputa.*
3. Errors in the transcription.
4. Neologisms, mostly derivatives.

Rules for handling derivative morphology have been shown in the previous paragraph. For the remaining three classes of unknown words, a simple approach is adopted:

a. Foreign words are included in a list, updated regularly.
b. Any word in the corpus but not in the lexicon is added, expanding the base resource.
c. Errors in the source texts are corrected, and then analysed by the tool.

To summarise, the tagger procedure, consisting of seven parts, involves the following:

1. *Unknown word detection*: Once the tokeniser has segmented the transcription into tokens, a quick look-up for unknown words is run. The new words detected are added to the lexicon.
2. *Lexical pre-processing*: The programme splits portmanteau words (*al, del* → *a el, de el*) and verbs with clitics (*damelo* → *da me lo*).
3. *Multiword recognition*: The text is scanned for candidates for multiwords. A lexicon, compiled from printed dictionaries and corpora, is used for this task.
4. *Single word recognition*: Every single token is scanned for every possible analysis according to morphological rules and lexicon entries. Approximately 30% of the tokens are given more than one analysis, and some of them are assigned up to 5 different analyses.
5. *Unknown word recognition*: The remaining tokens that are not considered new words pass through the derivative morphology rules. If some tokens still remain without being analysed (because they were neither included in the lexicon nor recognised by the derivative rules), they are held until the stage of statistical processing, when the most probable tag, according the surrounding context, is given.

6. *Disambiguation phase 1*: A feature-based Constraint Grammar resolves some of the ambiguities.
7. *Disambiguation phase 2*: A statistical tagger (the TnT tagger) resolves the remaining ambiguous analyses.

After such automatic tagging, human annotators can revise and correct the tagged corpus; Guirao and Moreno-Sandoval (2004) describe a tool developed for aiding human annotators in this task.

4.3.2 Tagset

During the process of disambiguation, C-ORAL-ROM Madrid developed a document explaining the different matters which concern the construction of the morpho-syntactic tagging system for the Spanish spoken corpus. This section will introduce this system, outlining the tagset used, how the tags were defined and, finally, what decisions were taken to solve the problems derived from ambiguity and from the limitations traditional grammar has when approaching categorisation. General theoretical decisions involving the definition of tags and their morpho-syntactic features will also be addressed.

4.3.2.1 *Tagset adopted*

A tag is defined as a descriptive symbol which is either manually or automatically assigned to a word or multiword inside a text (van Halteren 1999).

In Spanish grammar studies, the PoS problem is still far from being solved. In present-day literature on the subject, it is widely admitted that the list of PoSs we work with is based on a strange mix of criteria: semantic criteria for nouns and verbs; local, at times, for adjectives and prepositions; and of different nature for the adverb. With the aim of avoiding possible incoherence when assigning the different PoSs, in C-ORAL-ROM Madrid these are defined from three different points of view: semantic, morphological, and syntactic.

As for the way the different parts of speech are assigned to the different lexical units, there are two main theoretical models.

In the first place, there is the functionalist model, according to which, words are assigned one PoS or another depending on their syntactic behaviour inside the sentence. Second, there is the generativist model where words are assigned a PoS at source and, as a consequence, have a concrete behaviour in the syntax of that language. Let us examine an example:

(1) el hijo del presidente se educó en un colegio privado
[the son of the President was educated in a private school]

If we explain this example from a functionalist point of view, the word *hijo* would be assigned the PoS "noun" for the following reasons:

1. It is in a syntactic position which is typical of nouns.
2. It can be replaced by another noun.
3. *Hijo* has a syntactic function inside the sentence, which is also typical of nouns.

However, from a generativist point of view, the word *hijo* is a noun in itself and, as such, it has the syntactic behaviour just described. That is, it is not a noun because it is the subject of the sentence; rather, being a noun, it can be the subject of the sentence.

From the point of view of C-ORAL-ROM Madrid, the 'syntactic position' premise is not enough to justify the change of PoS in a lexical unit; this change must be based on semantic and morphological criteria as well. Above all, the semantic criterion has been favoured, and so the rest of parameters will be described from that perspective.

According to this perspective, the semantics of the PoS of a concrete word has morphological and syntactic consequences in the language being dealt with. For example, we will see how nouns are defined as a group of properties which tell one group of individuals from another. Those words which, in Spanish, designate classes of individuals, have gender and number information, occupy the central position in a phrase, allow meaning modifiers (such as articles, determiners or quantifiers) and always have a syntactic function inside the sentence (mainly: subject, direct object, etc.).

Table 4.2 shows the different PoSs used in the morpho-syntactic tagging of the corpus; afterwards, we will undertake the definition of each of these PoSs.

4.3.3 The notion of "multiword"

In C-ORAL-ROM Madrid, each sound chain with a unique meaning has been considered a word unit, regardless of its orthographic representation.

Table 4.2 The Spanish C-ORAL-ROM PoS tagset

PoS	Subcategory	Tag	Example
Noun		N	mesa
	Proper noun	NP	María
Adjective		ADJ	azul
Determiner	Article	ART	
	Possessive	POSS	mi, tu, su
	Demonstrative	DEM	ese, este
Quantifier		Q	uno, dos, tres primer, segundo muchos, pocos
Pronoun		P	yo, tú, él, lo_que
		PR	que
Verb		V	cantar
Auxiliary		AUX	habrá cantado
Preposition		PREP	ante, bajo, con
Adverb		ADV	así, aquí, allí
Conjunction		C	y, pero, ni
Discourse marker		MD	oye, o sea, es decir

In this sense, two kinds of word units can be found in the Spanish corpus: simple words and multiwords. Simple words are those graphically represented between two blank spaces. Complex words, on the contrary, are made up of two or more graphic units. We see an example of each kind in (2) and (3) respectively:

(2) mesa [table]

(3) fin de semana [weekend]

For C-ORAL-ROM Madrid, the following qualities are required for two or more lexical units to become a multiword:

a. Absence of compositional meaning. The new compound, as a whole, can only denote one meaning. For example, *al revés* is not the result of the sum of the meanings of *al* and *revés*, as it is the phrase *al cine* in the sentence *Vamos al cine*.
b. No insertion of other words inside the expression is allowed. For example, in *de todas formas*, when we add the article *la*, the result is *de todas las formas*. Therefore, we now have an expression which denotes the different ways in which an event can happen and is not anymore the discourse marker which *de todas formas was*.

The different lexical units which form multiwords are tagged together, joined by dashes, as in *fin_de_semana*.

4.3.4 Ambiguous clustering

In this group we have included those words which can have two kinds of syntactic analysis, that is, those words which can form either a phrase or a multiword, depending on the context, as illustrated in the following examples:

(4) es que\ES QUE\MD mañana no puedo ir a trabajar
[because tomorrow I cannot go to work]

(5) Lo que quiero decir es\SER\Vindp3s que\QUE\C no he dicho la verdad //
[what I mean is that I haven't told the truth]

(6) De eso nada\DE ESO NADA\MD guapa eso no lo haces tú ni por asomo.
[by no means darling you won't do that]

(7) de\DE\PREP eso\ÉSE\PPER3s nada\NADA\Q pero de esto que está aquí me da usted un kilo por favor.
[from that nothing but from this other piece here please give me a kilo]

Information given by prosodic tagging will help in the process of disambiguation through the use of rules. In the examples above then, the rules would tag (4) and (6) as MD, that is, they would analyse them as a multiword, while in (5) and (7) the analysis would be that of a phrase, where each word has its own PoS.

In those cases where rules cannot help, GRAMPAL will favour the syntactic anal-

4.3.5 Level of morpho-syntactic encoding of forms

In Table 4.3, the semantic, morpho-syntactic and syntactic features of each PoS are presented.

While Table 4.3 sums up the general features of each PoS, further explanation is needed concerning the decisions taken for some specific cases.

1. *Quantifiers* (*Q*). Those words tagged as quantifiers by C-ORAL-ROM Madrid have been classified by grammatical tradition either according to their syntactic position or to their semantic content. This way, in a traditional analysis of the same word, we can obtain two different categories, as illustrated in examples (8) and (9).

(8) algunos alumnos no han venido hoy a clase
[some students haven't attended the class today]

(9) algunos no han venido hoy a clase
[some haven't attended the class today]

In (8), *algunos* would be classified a determiner, while in (9) it would be a pronoun.

2. *Article or quantifier?* Words that in the grammatical tradition are tagged as articles have been classified as quantifiers in C-ORAL-ROM Madrid as well, as shown in (10) and (11):

(10) un niño
[a/one boy]

(11) unos niños
[several boys]

Choosing the article tag would mean sacrificing the enumerative interpretation in some contexts. It was therefore decided to consider these particles as quantifiers which, depending on different aspects which include their own semantic features, will denote a defined or undefined quantification.

3. *Discourse markers.* Discourse markers are linguistic units which, at the discourse level, "guide, according to their morpho-syntactic, semantic and pragmatic features, the inferences which take place in communication" (Portolés 1999); see, for instance, *vamos* and *mira* in (12) and (13):

(12) pero / vamos / a mí me gusta mucho //
[but / come on / I like it very much //]

(13) pues mira / fue / horrible //
[so look / it was / horrible //]

4.3 Semantic, morpho-syntactic, and syntactic features of each PoS

Semantic features	Morphological features				Syntactic features	Examples
	Per	Num	Gen	Tense-Mood		
Denote an element or object		+	+		Subject, direct object, complement	actitudes\ACTITUD\NCfp
classification					Absence of determiners	Luisa\LUISA\NP
Mono-referential	Fixed inflection				Noun complement, predicative complement	extranjeros\EXTRANJERO\ADJmp
Denote qualities or properties		+	+		Prenominal position, no syntactic function	los\EL\DETdmp
Restrict/define the referent of a noun phrase		+	+		Prenominal (1st series) and postnominal (2nd series) positions	mío\MIO\DETposs
Express relation of possesion or ownership		+	+		Pre and postnominal positions	este\ESTE\DETdem
Express location of the referent in space and time		+	+			una\UN\O
Express number of individuals or objects	Will vary depending on the kind of entity they are quantifying				Different syntactic functions inside the clause	que\QUE\PR
Retrieve the referent of the noun they modify inside the clause they introduce	Inherit their morphological features from the noun working as the referent				Maximum expansion of the noun phrase	yo\YO\PPER1s
Refer to a noun phrase	+	+	+		Central element of the sentence, determine the different syntactic functions	es\SER\Vindp3s
Express events	+	+		+		
Express a mental state	Invariability				Not been assigned	ah\AH\INT
Guide the inferences which take place in conversation	Invariability				Not been assigned	es decir\ES DECIR\MD
Establish logical or discourse bounds	Invariability				Relate sentences or elements in a sentence	pero\PERO\C
Establish semantic relationships associated to spatial concepts	Invariability				Establish relationships between two elements	de\DE\PREP
Set the meaning of the verb	Invariability				Do not introduce a second term	también\TAMBIEN\ADV

4. *The article lo.* C-ORAL-ROM Madrid has defined a pronoun as being able to recover its referent by itself. This *lo* cannot accomplish this function and this has been the main reason why it has been tagged as an article, as seen in examples (14) to (16).

(14) lo azul
[the blue]

(15) lo alto que eres
[* the tall you are]

(16) lo graciosa que es esta chica
[*the charming this girl is]

5. *Pronouns and adverbs.* This last group of words, traditionally classified as pronominal adverbs (Kovacci 1999), have been labelled as pronouns by C-ORAL-ROM Madrid, bearing in mind that, as other pronouns, these words are open terms whose referent is not fixed beforehand and does not keep itself constant, but is established every time there is a change of speaker, listener or space and time coordinates. We assume a semantic point of view in the definition of those parts of speech, understanding that they behave as entities and therefore they are a subclass of nouns. As a consequence, the following words, traditionally classified as adverbs, are classified as pronouns in the tagset:

abajo; acá; actualmente; ahí; ahora; allí; alrededor; anche; antaño; antes; aquí; arriba; así; atrás; ayer; debajo; delante; después; detrás; encima; enfrente; entonces; hoy; luego; mañana; mientras

4.3.6 Evaluation

The total number of units (where unique words and multiwords each count as one unit) in the test corpus (hand-annotated)[8] is 44,144.

The test corpus has been developed using a combined procedure of automatic and human tagging:

1. A fragment of approximately 50,000 words (15% of the corpus) was selected, taken from the different sections and intended to be a representative sampling of the whole. Each word in the 27 texts was tagged with all possible analyses. This means that some words (the unambiguous ones) have one tag, while other are given one tag per morpho-syntactic analysis.
2. Each file was revised by a linguist who selected the correct tag for every case, discarding the wrong ones.
3. From the revised corpus, a set of disambiguation rules were written for handling the most frequent cases.
4. A new run of the tagger, augmented with the disambiguation grammar, provided an automatically tagged corpus, with only one tag per unit.

5. The automatic and human tagged corpora were compared, and the differences were noted one by one, assuming that agreement on the same tag implied a correct analysis. While, in most cases the wrong tag was assigned by the tagger, in several cases it was the linguist who provided an incorrect tag: mistakes were probably due to a lapse in attention due to the repetitiveness of the task. After assigning the proper tag in all the disagreements, a final version of the test corpus was delivered.

Both the disambiguation grammar and the statistical tagger were trained against the test corpus. Finally, the rest of the corpus of over 250,000 words was tagged as described in Section 3.1.

In order to evaluate the tagger performance (including disambiguation), a new run of GRAMPAL against the test corpus was conducted. The mismatches between the GRAMPAL tagged corpus and the test corpus, working as a golden standard for evaluation, were counted (the figures are shown in Table 4.4). The precision rate was calculated by the number of correct tags assigned by the tagger divided by the total number of tagged units in the test corpus. In other words, 42,206 tags out of 44,144 were assigned correctly. No evaluation of the precision has been performed for the rest of the corpus, but a similar rate (95.61 %) as that obtained against the test corpus can be assumed.

With respect to the recall, understood as the ratio between the number of tagged units by the programme and the total number of units, the figure for the whole corpus is 99.96. Only 117 tokens were not given a tag by the programme.

There is a discrepancy between the number of words in the transcription corpus and the number of words in the tagged corpus. This can be explained by the fact that a different concept of 'word' has been used in transcription and in PoS tagging. In transcription, a word is simply a string between blank spaces, while a word in tagging is a lexical or grammatical unit. That is, it can be a single word (*hola*) or a multiword (*es decir*). Since there are many multiwords in the corpus, the actual number of tagged words is less than the number of transcribed words.

With respect to the evaluation of PoS tagging, it is important to stress that only a subset of approximately 50,000 transcribed words were revised by hand, resulting in over 44,000 tagged words. The rest of the tagged corpus has not been revised by human annotators. This fact has some consequences in the list of forms and lemmas. The tagger, when two or more tags are available, always assigns the tag with the shared information between the candidates. For instance, many verb forms are ambiguous with respect to first and third person singular: (*yo*) *cante* / (*él, ella*) *cante* 'I sing' vs. 'he/she sings'. The tags for each are Vp1s and Vp3s, respectively. When the context cannot solve

Table 4.4 Recall and precision rates of test corpus

Number of units	Number of tagged units	Recall	Precision
.....	44,144	100%	95.61%

the ambiguity (by means of the pronoun), however, the tag assigned is V, compatible with both. Human annotators, however, can normally resolve the ambiguity when they are revising the tagging, in which case, the appropriate full tag is provided. As a result, different tags for the same word can be found in the list of lemmas and forms.

4.3.7 Specific tagging problems with the Spanish spoken corpus

4.3.7.1 *Retracting and interruption phenomena*

Phenomena such as retracting and interruption will not be tagged as the rest of elements during the process of automatic tagging, that is, they will not be tagged as PoSs. Nonetheless, in C-ORAL-ROM Madrid we have considered it important not to erase this information because it will play an important part in the process of manual disambiguation and spontaneous speech corpora training by means of disambiguation rules.

1. *Retracting.* The retracting phenomenon presents itself as problematic when dealing with automatic morpho-syntactic tagging. In the transcription of spoken language we have included cues to signal various retracting phenomena and for this reason the rules for automatic disambiguation thus cannot work properly. Therefore, different rules should be written to avoid problems as those which would arise in the following example:

(17) entonces la [/] la llamé
[then I called her [/] her]

In this case, as the fragment lacks the proper context (*la* is a pronoun before a verb), the tagger will give the first *la* the "article" PoS tag, when it is really a pronoun. The retracting phenomenon will be labelled as <re> in the final morphologically tagged text. In C-ORAL-ROM this phenomenon can be divided into two kinds:

a. In the first model of retracting, a repetition of the same word on both sides of the label [/] will take place, for example:

(18) la [/] la [/] la moto de mi abuelo Pepe.
[the the the motorbike of my grandfather Pepe]

In these cases, after applying the contextual rules for disambiguation, GRAMPAL gives the likeliest analysis, which in the example would be the following:

(19) la\LO\PPER3s <re> la\LO\PPER3s <re> la\EL\DETdfs moto\MOTO\NCfs
de\DE\PREP mi\MI\DETposs abuelo\ABUELO\NCms Pepe\PEPE\Npi

It can be seen how the analysis of the first two instances of the word *la* results in the tag P (pronoun), because the contextual rule *la* → *ART/_N* could not be applied as it could be in the case of the third case of *la*. In order to solve these cases, C-ORAL-ROM

left side of the label [/] with the same PoS tag as the one on the right side. For example, in the first case we saw in example (18), the word *la* in the noun phrase *la moto* is an article; therefore, all the words *la* on the left side of the retracting will be tagged as ART, as illustrated in (20):

(20) la\EL\DETdfs <re> la\EL\DETdfs <re> la\EL\DETdfs moto\MOTO\NCfs
de\DE\PREP mi\MI\DETposs abuelo\ABUELO\NCms Pepe\PEPE\Npi

b. In the second model of retracting, there is no similarity between the words occurring on both sides of the label [///], as seen in (21):

(21) la [///] tiras de ahí / lo haces //
[the [///] pull out / you do it //]

In such cases, due to the break in discourse, it is not possible to predict the PoS of the word, so the tag for such words will be the first one GRAMPAL assigns automatically, as shown in (22):

(22) la\EL\DETdfs <re> tiras\TIRAR\Vindp2s de\DE\PREP ahí\AHÍ\P
lo\LO\PPER3s haces\HACER\Vindp2s

2. *Interruption.* In C-ORAL-ROM, both the interruption and the change of topic phenomena are represented with the symbol +, as seen in example (23):

(23) No quiero ir a ese + ayer me dijo que yo no lo había comprado //
[I don't want to go to that + yesterday he told me that I hadn't bought it //]

These cases will be dealt with in the same way as in the second case of retracting, i.e. maintaining the choice that the morphological tagger makes after applying GRAMPAL and the contextual disambiguation rules, as in (24).

(24) no\NO\ADV quiero\QUERER\Vindp2s ir\IR\V a\A\PREP
[no\ADV I-want\V I-go\V to\PREP
ese\ESE\DETdem +
that\DET]

4.3.7.2 *Linguistic forms whose distribution is not consistent with the distributional characters of written language*

In the Spanish corpus in C-ORAL-ROM, those linguistic occurences which are typical of spoken language are labelled according to the following conventions:

1. *Support:* &ah and &eh. These elements are tagged as <sup>. These are illustrated in examples (25) and (26):

(25) mañana &ah no voy

(26) mañana\MAÑANA\P <sup> no\NO\ADV voy\IR\Vindp1s
[tomorrow\P no\ADV I-go\V]

2. *Non linguistic forms,* transcribed as *hhh*, are labelled with the tag <nl>, as in (27) and (28):

(27) mañana hhh no voy
[tomorrow hhh I won't go]

(28) mañana\MAÑANA\P <nl> no\NO\ADV voy\IR\Vindp1s
[tomorrow\P no\ADV I-go\V]

3. *Onomatopoeia,* for example, the imitation of the sound of cars, birds, etc., will be dealt with depending on whether the sound has a conventional transcription or not, as elaborated on below:

a. Those onomatopoeia which correspond in the written form with a conventional linguistic expression will be tagged as INTJ (interjection), even though in some contexts, these words have a syntactic function, as in the following example, where the interjection *¡ras!* plays the role of a direct object:

(29) y le hizo ¡ras! en toda la cara
[and he did splash ! in his face]

(30) y\Y\C le\LO\PPER3s hizo\HACER\Vindp3s ¡ras!\RAS\INT en\EN\PREP
toda\TODO\Q la\EL\DETdfs cara\CARA\NCfs
[and\C to-him\PRER3s he-did\V splash\INT ...]

b. Those onomatopoeia not conventionalised with a linguistic expression and which were transcribed in C-ORAL-ROM Madrid as *hhh* have, as with the non-linguistic forms in point 2. above, the tag <nl>.

4. *Interjections:* The rest of the expressions expressing the speaker's states of mind were tagged as INTJ. Those interjections not included in the Diccionario de la Real Academia Española (DRAE) were added to a list called "C-ORAL-ROM interjections", which can be consulted in the document where the construction of the tagging is explained. An example from the corpus of interjections not included in the DRAE is shown in (31).

(31) uju!\YUJU\INT mañana\MAÑANA\P <nl> no\NO\ADV voy\IR\Vindp1s
[yuju\INT tomorrow\P no\ADV I-go\V]

5. *Meaningless linguistic forms:* This group includes those expressions which are not alphabetically transcribed and which do not have a referent. An example is the case of a speaker humming a song, shown in (32):

(32) *ALF: tachín tachín tachín para pa pa para para pa pa
[singing]

These cases, which GRAMPAL leaves untagged, are labelled afterwards as <nc> (non categorial).

4.3.7.3 *Linguistic forms and non-standard forms used as discourse markers*
The label Discourse Marker (MD) is a PoS which was created for the C-ORAL-ROM tagging; as such, this documentation includes its definition, the list of words belonging to the class, as well as an account of how C-ORAL-ROM Madrid solved the problems emerging from the choice of this PoS; this is summarised below. The list of discourse markers is made up of different groups of words, as elaborated on in the following.

1. *Words which are always MD.* In this group are words such as *o sea, sin duda, por tanto, y tal,* etc. These words are tagged automatically and do not present any problem of ambiguity.

2. *Words with PoS ambiguity.* There is a group of words which, in some syntactic contexts, behave as MD, while in other contexts behave as conjunctions, adverbs, nouns or adjectives. We can see in examples (33) to (35) the different interpretations the word *vamos* can have:

(33) hoy vamos\IR\Vindp1p al cine
[today we go\V to the movies]

(34) vamos\VAMOS\INT hombre eso no te lo crees ni tú
[come on\INT man you don't believe that]

(35) era muy competitivo vamos\VAMOS\MD un trepa
[he was very competitive say\MD a go-getter]

In these cases, during automatic tagging, the disambiguation rules will give priority to one PoS over another. In the case of MDs, the contextual disambiguation rules use the information given by the prosodic tags. If a word is transcribed between single or double slashes, or, for example, if it is a dialogical turn in itself, given its prosodic and syntactic independence, we will probably be dealing with a MD.

4.3.8 Main data from lemmatisation

In Tables 4.9 and 4.10 we compare the data obtained in C-ORAL-ROM from morphological tagging with those from other corpora: the Spanish UAM Treebank, a 21,420-word written corpus taken from newspapers (Moreno et al. 2003); the 10 million-word spoken part of the 41 million-word Longman Spoken and Written English (LSWE) Corpus, from which we have obtained the occurrence percentages for verbs, nouns, ad-

Table 4.5 High frequency verbs, excluding auxiliaries and modal verbs

Rank	Lemma	Rank	Lemma	Rank	Lemma	Rank	Lemma
1	ser	26	comer	51	subir	76	utilizar
2	decir	27	trabajar	52	perder	77	tirar
3	estar	28	contar	53	esperar	78	considerar
4	tener	29	coger	54	echar	79	jugar
5	hacer	30	unir	55	acordar	80	dormir
6	haber	31	valer	56	ganar	81	mantener
7	ir	32	encontrar	57	traer	82	levantar
8	ver	33	meter	58	abrir	83	morir
9	dar	34	empezar	59	mandar	84	terminar
10	saber	35	conocer	60	recordar	85	caer
11	pasar	36	poder	61	suponer	86	permitir
12	poner	37	mirar	62	quitar	87	mover
13	creer	38	pedir	63	sobrar	88	necesitar
14	venir	39	entender	64	acabar	89	salar
15	llamar	40	vivir	65	leer	90	conseguir
16	llevar	41	entrar	66	imaginar	91	fijar
17	hablar	42	seguir	67	tratar	92	servir
18	quedar	43	buscar	68	estudiar	93	aparecer
19	querer	44	sacar	69	intentar	94	bajar
20	llegar	45	volver	70	ocurrir	95	realizar
21	dejar	46	comprar	71	escuchar	96	costar
22	salir	47	pagar	72	sentar	97	referir
23	parecer	48	preguntar	73	tocar	98	interesar
24	gustar	49	tomar	74	casar	99	aprender
25	pensar	50	cambiar	75	explicar	100	andar

Table 4.6 High frequency nouns

Rank	Lemma	Rank	Lemma	Rank	Lemma	Rank	Lemma
1	día	26	padre	51	estudio	76	peseta
2	año	27	hermano	52	país	77	teléfono
3	cosa	28	forma	53	hombre	78	sábado
4	tío	29	hijo	54	punto	79	kilómetro
5	vez	30	caso	55	partido	80	derecho
6	casa	31	amigo	56	nivel	81	sistema
7	gente	32	madre	57	cuenta	82	piso
8	persona	33	tarde	58	precio	83	cara
9	tiempo	34	gobierno	59	mujer	84	información
10	momento	35	grupo	60	producto	85	montón
11	problema	36	señor	61	palabra	86	sol
12	niño	37	sitio	62	programa	87	novio
13	hora	38	mes	63	historia	88	ordenador

Table 4.6 (*continued*)

Rank	Lemma	Rank	Lemma	Rank	Lemma	Rank	Lemma
15	mundo	40	semana	65	falta	90	señora
16	idea	41	ciudad	66	zona	91	compañero
17	vida	42	manera	67	libro	92	domingo
18	trabajo	43	clase	68	servicio	93	equipo
19	tipo	44	dinero	69	empresa	94	relación
20	parte	45	sentido	70	centro	95	cliente
21	tema	46	situación	71	minuto	96	película
22	uno	47	coche	72	familia	97	lunes
23	noche	48	gracia	73	virus	98	ciencia
24	mañana	49	agua	74	calle	99	lengua
25	pueblo	50	medio	75	puerta	100	sociedad

Table 4.7 High frequency adverbs

Rank	Lemma	Rank	Lemma	Rank	Lemma
1	no	35	tarde	68	máximo
2	sí	36	prácticamente	69	especialmente
3	ya	37	absolutamente	70	claramente
4	también	38	efectivamente	71	últimamente
5	bien	39	totalmente	72	básicamente
6	claro	40	evidentemente	73	inmediatamente
7	así	41	precisamente	74	del todo
8	siempre	42	en general	75	recién
9	además	43	probablemente	76	sinceramente
10	tampoco	44	quizá	77	quizás
11	tal	45	perfectamente	78	generalmente
12	todavía	46	justo	79	obviamente
13	a lo mejor	47	directamente	80	poco a poco
14	sólo	48	aún	81	concretamente
15	igual	49	de repente	82	principalmente
16	casi	50	pronto	83	franco
17	al final	51	seguramente	84	en teoría
18	nunca	52	seguro	85	exclusivamente
19	sobre todo	53	de vez en cuando	86	menos mal
20	más o menos	54	al principio	87	curiosamente
21	mal	55	o así	88	posteriormente
22	fuera	56	fundamentalmente	89	indudablemente
23	realmente	57	aparte	90	a la vez
24	por lo menos	58	completamente	91	habitualmente
25	mejor	59	justamente	92	a su vez
26	a veces	60	de pie	93	cierto
27	solamente	61	lógicamente	94	alrededor
[illegible]	[illegible]	[illegible]	[illegible]	95	inicialmente

Table 4.7 (*continued*)

Rank	Lemma	Rank	Lemma	Rank	Lemma
30	primero	64	de nuevo	97	en coche
31	normalmente	65	al revés	98	afortunado
32	dentro	66	en serio	99	finalmente
33	cerca	67	en absoluto	100	previamente
34	exactamente				

Table 4.8 High frequency adjectives

Rank	Lemma	Rank	Lemma	Rank	Lemma	Rank	Lemma
1	mismo	26	buen	51	fundamental	76	correcto
2	bueno	27	cierto	52	especial	77	vecino
3	mejor	28	bajo	53	humano	78	pesado
4	pequeño	29	general	54	rápido	79	caro
5	grande	30	diferente	55	próximo	80	histórico
6	importante	31	junto	56	barato	81	rico
7	último	32	alto	57	blanco	82	joven
8	solo	33	inglés	58	natural	83	imposible
9	nuevo	34	pobre	59	fatal	84	terreno
10	normal	35	peor	60	técnico	85	necesario
11	mayor	36	interesante	61	urbano	86	público
12	social	37	raro	62	mental	87	verde
13	distinto	38	claro	63	mínimo	88	primate
14	español	39	subordinado	64	suficiente	89	internacional
15	bonito	40	nacional	65	material	90	delincuente
16	propio	41	pasado	66	europeo	91	moderno
17	único	42	fácil	67	guapo	92	frío
18	siguiente	43	anterior	68	tonto	93	típico
19	político	44	principal	69	radical	94	gracioso
20	malo	45	extranjero	70	informático	95	serio
21	posible	46	libre	71	contento	96	clásico
22	gran	47	capaz	72	electoral	97	habitual
23	mal	48	científico	73	estupendo	98	central
24	fuerte	49	económico	74	viejo	99	castellano
25	difícil	50	determinado	75	menor	100	grave

jectives and adverbs in conversation; and finally, the frequency word lists from Juilland and Chang-Rodríguez (1964) (Spanish) and P. M. Alexejew et al. (1968) (English).

In the comparison of a spoken Spanish corpus (C-ORAL-ROM) and a written one (UAM Treebank), two conclusions can be reached:

1. There is an inverse distribution for lexical (nouns, verbs, adjectives etc.) and non-lexical (conjunctions, prepositions etc.) categories: the latter are much more

Table 4.9 Frequency word lists: comparison between C-ORAL-ROM and other corpora

	C-ORAL-ROM	Spanish UAM Treebank	Juilland & Chang-Rodríguez*	Alexejew et al. (English)*
Verbs (lemmas)	1 ser	1 ser		
	2 decir	2 tener		
	3 estar	3 estar		
	4 tener	4 pedir		
	5 hacer	5 anunciar		
	6 haber	6 haber		
	7 ir	7 hacer		
	8 ver	8 querer		
	9 dar	9 dar		
	10 saber	10 morir		
Nouns (lemmas)	1 día	1 año		
	2 año	2 millón		
	3 cosa	3 gobierno		
	4 tío	4 día		
	5 vez	5 país		
	6 casa	6 vida		
	7 gente	7 mundo		
	8 persona	8 niño		
	9 tiempo	9 presidente		
	10 momento	10 juez		
Adjectives (lemmas)	1 mismo	1 español		
	2 bueno	2 grande		
	3 grande	3 nuevo		
	4 mejor	4 bueno		
	5 pequeño	5 último		
	6 importante	6 pequeño		
	7 último	7 político		
	8 solo	8 europeo		
	9 nuevo	9 italiano		
	10 normal	10 vasco		
Most frequent words	1 de	1 el	1 de	1 the
	2 el	2 de	2 el	2 of
	3 y	3 en	3 la	3 to
	4 a	4 a	4 y	4 in
	5 que (conjunction)	5 un	5 a	5 and
	6 la	6 y	6 en	6 a
	7 no	7 suyo	7 él	7 for
	8 en	8 ser	8 que (pronoun)	8 was
	9 es	9 con	9 ser (lemma)	9 is
	10 se	10 que (conjunction)	10 que (conjunction)	10 that

* These corpora do not contain specific information on lemmatisation. Reflected in the table are

Table 4.10 Distribution of lexical classes: comparison between C-ORAL-ROM and other corpora

	C-ORAL-ROM %	Spanish UAM Treebank %	Longman (Conversation) %
Verbs	17.12	10.36	13
Nouns	13.39	31.25	14
Adjectives	3.97	6.99	3
Adverbs	6.50	2.88	6
Total	40.98	51.48	36

2. It is also interesting to note the difference in the distribution of verbs, nouns, adjectives and verbs in the two corpora. In speech, the presence of verbs and adverbs is much greater than in writing, while nouns and adjectives are much more frequent in written than in spoken texts. The correlation between the frequencies of, first, nouns and adjectives, and, second, verbs and adverbs seems obvious, as the members of these pairs of words are closely related through syntactic features, with adjectives and adverbs being modifiers of nouns and verbs, respectively.

As for the comparison between spoken Spanish and spoken English, the results are very similar in general terms: the distribution for nouns, adjectives and adverbs is almost equal, while in the case of verbs, the frequency is higher in Spanish; it is also noticeable how verbs are the most frequent category in Spanish, while in English this position corresponds to nouns.

Notes

1. The website of VALESCO is http://www.uv.es/~valesco/.

2. Its website is http://www.uvigo.es/webs/sli/index.html

3. We have recently digitalised the analogue audio tapes. For more information, please contact francisco.marcos.marin@uam.es

4. In this sense, CORLEC is more extensive, spontaneous and natural than C-ORAL-ROM, since it is three times larger and does not have the constraint of needing to obtain the permission of all the participants.

5. Universidad Autónoma de Madrid acknowledges the source of the sound files in the media recordings as being kindly provided by RTVE (Radio Televisión Española), Radio Televisión Madrid, COPE (Cadena de Ondas Populares Españolas/Radio Popular) and Onda Cero Radio.

6. Not available in this edition.

7. See the complete list of non-orthographic productions in the table of correspondence in the DVD, in the Spanish section of the corpus Metadata Menu.

8. The files from the C-ORAL-ROM corpus annotated by hand are the following: efamcv03; efamcv06; efamcv07; efamdl03; efamdl04; efamdl05; efamdl10; efammn05; emedin01; emdmt01; [illegible]; emedts01; emedts05; emedts07; emedts09;