# Language Resources Development at the Spanish Royal Academy

## Ángel Martín Municio, Guillermo Rojo, Fernando Sánchez León, Octavio Pinillos

Real Academia Española

Felipe IV 4, 28014 Madrid, Spain {amunicio, grojo, fsanchez, pinillos}@rae.es

#### Abstract

This paper explains some of the most relevant issues concerning the development of language resources at the Spanish Royal Academy. Two 125-M words corpus of Spanish language (synchronic and diachronic) and three specialized corpus has been developed. Around the corpus, RAE is also developing NLP tools and resources to morpho-syntactically annotate them. Some of the most relevant are: The Computational Lexicon, the Morphological analysis tools, the Disambiguation grammars and the Tokenizer generator. The last section describes the lexicographic use of corpus materials and includes a brief description of the Corpus-based lexicographical workbench and his related tools.

## 1. Introduction

The Spanish Royal Academy has traditionally exerted during the last three centuries the role of fixing the linguistic norms of use in the successive editions of Dictionaries, Grammars and Orthographies. These activities constitute the basic instrument to promote the unity of the language. The relation between the Spanish Royal Academy and the nineteen Academies from the Spanish-speaking countries, as well as those from the United States and Philippines (all of which integrating the "Association of Academies" of the Spanish Language has always produced beneficial results such as the incorporation of American words into Dictionaries and the general agreement for the recent edition of a new Orthography.

After a long time effort stemming in the 18th century driven by a traditional view of lexicography, the Royal Spanish Academy (RAE) has shifted in recent years towards a corpus-based lexicographic methodology. To this aim, RAE started two huge projects for corpus compilation aiming at the development of a reference corpus for present-day Spanish (CREA) and a representative historical corpus ranging from origins of Spanish to the start of the period covered by the former corpus (CORDE).

Each corpus will have grown by mid-2000 to the aimed size of 125M words. CREA is defined as a monitor corpus, so a follow-up for this project on the time axis is guaranteed. On the other hand, the size of both corpora, but specially of CORDE, is felt insufficient to provide full coverage of the lexical richness of certain periods of the evolution of the Spanish, thus more materials will be included in the near future.

## 2. Corpus development

### 2.1. Modern Spanish Reference Corpus (CREA)

CREA (Corpus de Referencia del Español Actual) is a project developed with the aim of providing researchers as well as people interested in Spanish with a balanced and representative sample of standard contemporary language. In order to achieve the largest flexibility in data retrieval, CREA is structured into several sections to make possible different kinds of corpus research, ranging from fullcorpus to specific subcorpus research, defined in terms of parameters such as geographic origin, subject, date, etc.

This complex structure is the result of combining different criteria:

Chronological criteria: recent texts (last twenty five years 1975-1999).

Geographical criteria: texts from Spain and America (and other Spanish-speaking countries, European and non-European), equally distributed.

Medium: books, newspaper, ephemera, spoken.

Superfields: Science, Arts, Leisure, Fiction...

The size of the corpus, 125 million words, is a balanced result of the need of getting the largest amount of different words without giving up encoding and grammatical annotation. Written texts represent 90% of the whole corpus, while spoken and ephemera texts make up the remaining (10%). Since this size is large enough to guarantee variety and representativeness, the corpus has been compiled with complete texts.

Texts are distributed according to the following general parameters:

Geographic distribution: Spain: 50 % - America: 50 %.

Chronological distribution: 1975-79(10%), 1980-84 (15%), 1985-89(20%), 1990-94(25%), 1995-1999 (30%).

«Superfield» taxonomy: 1. Science and Technology (Biology, Physics, Mathematics...), 2. Social Science, Belief & Thought (History, Religion, Education...), 3. Politics, Economy, Commerce & Finance (Industry, Government, Business...), 4. Arts (Music, Painting, Films...), 5. Leisure (Gastronomy, Tourism, Hobbies...), 6. Health (Medicine, Nutrition, Psychiatry...), 7. Fiction (Novel, Tales, Drama...).

American texts have been classified into a variety of major geographical and linguistic areas of non-peninsular Spanish. The percentage assigned to each area is proportional to its population and cultural weight:

Mexican area (40% of the whole American section of CREA): Mexico, south-western USA, Guatemala, Honduras and El Salvador.

Central area (3%): Nicaragua, Costa Rica, Caribbean area (17%): Cuba, Puerto Rico, Panama, Dominican Republic, Venezuelan and Colombian coasts.

North-eastern USA.

Andean area (20%): rest of Venezuela and Colombia, Ecuador, Peru, Bolivia.

Chilean area: (6%): Chile, River Plate area: (14%): Argentina, Paraguay, Uruguay.

CREA has been conceived as a monitor corpus. This means that the corpus includes only texts dated in the last 25 years. After 2001, new texts will be included in CREA while the earliest texts will become part of CORDE (Diachronic corpus of Spanish).

#### 2.1.1. Oral subcorpus

Spoken texts have their own design and encoding scheme, as it will be shown in the next paragraphs.

The spoken part of CREA (10% of the whole corpus) contains two groups of texts: (i) radio or television programmes, which have been transcribed and encoded according to the CREA encoding scheme (see below) and procedure; (ii) texts already transcribed by other groups outside the Academy and adapted to the CREA encoding scheme.

The taxonomy of spoken texts is also structured according to the opposition between radio-television, on the one hand, and texts collected from a pre-defined set of domains or contexts, on the other. The following text genres are distinguished:

Radio and television programmes: news, reports, interviews, talk shows, documentaries, sport broadcast, magazines, varieties programmes and quizzes.

Other recordings: these texts are classified into eight categories defined in terms of three parameters: formality (low, high), addressee (active addressee, passive addressee), channels (face to face, other channel).

Spoken texts are also classified according to other parameters: means (radio, television, direct recording, telephone, other), formality (basically formal, basically informal), control (basically free, basically controlled).

The geographic origin of CREA spoken texts is equally distributed between Spain, America and other Spanish-speaking countries. Thus texts coming from Spain will represent 50% of the whole spoken corpus by the end of the project. The spoken part of CREA will be very useful for voice recognition programs or Spanish training as a second language.

### 2.1.2. Encoding scheme

Three different schemes have been designed to encode CREA texts: a first level scheme for written texts, a second level scheme for written texts and a transcriptionencoding scheme for spoken texts.

The two-level distinction of encoding schemes for written texts is due to the need of achieving a large amount of encoded texts within a limited period of time. The first level scheme contains a minimal set of mainly structural and easy-introducing tags, while the second level scheme adds a variety of content and contextual information which requires always human supervision or control. This double-stage encoding procedure guarantees both speed while compiling the corpus and depth in the encoded result. However, only part of the whole corpus will be encoded according to the second level scheme by the end of 2000. At the present stage, only a one millionword subcorpus fulfils second level mark-up requirements.

The encoding scheme for spoken texts<sup>1</sup> is as complex as the second level scheme for the written part of CREA. Spoken texts represent only a small part of the whole corpus, which means that it is possible to devote a longer time to transcription and encoding processes in this section than in the written part.

CREA encoding principles follow the recommendations of the TEI<sup>2</sup> (Sperberg-McQueen -Burnard, eds., 1994) and the CES<sup>3</sup> (Ide, ed., 1996). The former is an all-purpose description of encoding principles for texts, while the latter is an application of the TEI for linguistic corpus encoding. However, none of these proposals have been applied exactly as they have been defined. Almost all the TEI/CES content tags have been removed from the first level encoding scheme, although the TEI header still remains in all the texts. The second level encoding scheme has expanded the limits of the TEI in some aspects not very suitably solved. Finally, in the transcription and encoding scheme for spoken texts, most elements of the TEI and CES proposals are retained, but with some added attributes.

## 2.2. Diachronic Corpus of Spanish (CORDE)

CORDE (Corpus Diacrónico del Español) is a very large corpus including 125 million words of texts ranging from origins of Spanish up to 1975, the year when the corpus CREA begins. CORDE has been designed to help all the researchers interested in language evolution and, besides, in order to create in the future a historical Dictionary.

<sup>&</sup>lt;sup>1</sup> Regarding the criteria for orthographic transcription, the spoken subcorpus of CREA is based on the following proposals: French (1992) –included in NERC-1 (1994)–, Marcos Marín (1992), TEI (Sperberg-McQueen – Burnard, eds., 1994), SpeechDat (1997) and EAGLES (Llisterri 1994, 1996). The main innovations are the result of applying those models –sometimes defined in a excessively theoretical way– to Spanish texts.

<sup>&</sup>lt;sup>2</sup> Text Encoding Initiative.

<sup>&</sup>lt;sup>3</sup> Corpus Encoding Standard.

CORDE is a written text corpus encoded with a minimal SGML mark-up scheme, just as CREA. In order to offer the users a wide versatility regarding exploitation, CORDE has been structured taking into account several parameters, such as the following ones:

Chronological: the Corpus is structured into three periods (Middle Ages, Golden Centuries and Contemporary Age), which can be grouped into three smaller periods according to both historical and linguistic criteria.

Geographical: CORDE includes Spanish texts from all around the world where this language is or has been spoken. Given the diachronic perspective, peninsular Spanish has a weight of 74% in the corpus and the rest represents a 26%.

Modality and genre: The corpus is structured into two groups: 1. Fiction, including verse and prose texts, subdivided into Lyric, Epic and Dramatic. 2. Non Fiction, including prose texts, which are structured into the following genres: Didactic, Scientific, Social, Press and Publicity, Religious, Historic-documentary and Law.

The generic distribution is the following:

Fiction 44% (verse, prose, drama).

Non Fiction 56% (Didactics 9%, Science and Technique 15%, Society, Press and Publicity 8%, Religion 6, History 13%, Law and Juridical Science 5%).

Regarding the chronological distribution, the structure is as follows:

Middle Ages

Origins of the language until 1250. Alfonso X Age. 1250-1492. Territorial unity, Nebrija grammar.

Golden Centuries

1493-1598. Felipe II's death. Decline of the Empire. 1599-1713. RAE foundation.

Contemporary Age

1714-1812. End of the Independence War. Cadiz Constitution.1813-1898. War of 1898.1899-1936. Civil War.1937-1974. Postwar period.

## 2.2.1. Encoding scheme

Given the specific features of the old and classic texts that make up the Diachronic Corpus of Spanish (CORDE), we have adapted and extended the SGML markup system designed for CREA.

Taking into account that the corpus is now at a first encoding stage, we have tried to collect all the information needed to accurately interpret the structure and the content of texts so that information extraction can be possible.

We have established several basic criteria for text interpretation, respecting exactly (as far as possible) the source edition: orthography and accentuation, punctuation, highlights, illustrations, pictures... translating eventually the editor's indications to SGML tags, noticing always any intervention in the comments section of each encoded work.

It is possible to distinguish the main author's words from the rest of authors who take part in preliminary compositions, licenses, approvals, censures, errata, etc., by means of SGML tags.

CORDE, unlike CREA, includes verse texts for which a basic markup system has been developed. Thus, each line of the text it is encoded in a presentational way, that is to say, without establishing where a line of the poem starts or ends. Besides, this minimal markup allow to represent meaningfully indentations made by the author, which is specially useful in verse dramatic works to keep the lines layout of the source edition, allowing a later use of this information.

All the words, expressions or sentences in a foreign language have their corresponding tag in order to be kept apart during the lemmatization process and to be retrieved as foreign words.

It is worthy of mentioning the works obtained from agreements with several universities and cultural institutions (Madison, SECRIT, etc.). These texts, already encoded in a different markup system, have been encoded and translated into SGML format, in order to be included in the corpus with a deeper encoding in some cases (Madison).

To sum up, the CORDE markup system is now at a minimal stage, but all the tags, including those which are a trace of some element of the edition, can be used as a first step to re-encode all the texts in a deeper way.

## 2.3. Specialized Corpus

As time goes on, RAE considered the absolute necessity to add several specific subcorpus. Nowadays, RAE is working on next research lines on Corpus creation:

A technical and scientific textual Corpus, a very helpful tool to reach the unification and normalization of the scientific terminology in Spanish. For this purpose, the scientific and technological terminology has been divided into about 50 areas of knowledge according to a tree of relationships from Medicine an Biomedicine to the fields of engineering and technology, from texts published in Spanish during the last 10 years. So that, the percentages of each area change from 2 to 10 per cent according to the amount of Spanish texts, mainly in electronic form, to achieve a total of 20M words, names exclusively. Given this grammatical restriction, mark-up requirements will be minimal.

A juridical Corpus, that will be of use to create juridical and legal terminology.

A secondary school textual Corpus to elaborate the School Dictionary main body text.

## 3. Tools and other resources

Around the corpora described in the previous section, RAE is also developing NLP tools and resources (T&Rs) to morpho-syntactically annotate them. However, these T&Rs will serve not only the purpose of providing an automatic analysis of the text corpora available to the Institution or coming from other research groups, but also that of helping other RAE departments in their everyday lexicographic work. Some of the most salient T&Rs created are described below.

## 3.1. Lexicon

The computational lexicon used by the tools, called LEX-CREA, is a collection of lexical entries encoded to the morpho-syntactic level that has been structured according to different criteria. While diachronic and geographical issues are the most important axes to fetch entries during analysis of a given text (specially given the fact that both historical and regional varieties of Spanish are represented in the different corpora under construction) other aspects like sublanguage and frequency have been also gathered from the various corpora and included in the lexicon. This information will heuristically guide the disambiguation process along with linguistic information properly. As regards its extension, the lexicon contains the lemma set coming from the RAE dictionary and other major dictionaries. However, as it has been demonstrated in its use with the corpora described in this paper, none of the dictionaries consulted contain all the lexical entries observed in the corpora, so additional effort has been devoted to the encoding of these new entries

### 3.2. Morphological analysis

The Computational Linguistics Department at RAE has developed two morphological descriptions of Spanish, one is used by mmorph (Petitpierre & Russell, 1995), a two-level and unification based morphological generator, the other is a code library that can be plugged to different NL components. These inflectional modules are complemented by a set of specialized components ranging from the recognition and adequate treatment of clitic pronouns to verbs (morphological pasted to certain verbal forms in Spanish), some aspects of derivational morphology and guessing.

On top of these tools, a Web-based verbal generator specifically designed for pedagogical purposes has been implemented, using colors to represent different irregularities involved (phonological, morphological or simply orthographic, or a combination of them).

## 3.3. Disambiguation grammars

The main strategy adopted for disambiguation is a symbolic one, based on the linguistic information provided by the lexicon. However, this approach allows for the inclusion of other kinds of information (syntactic, semantic... but also frequency information as gathered from the corpora themselves) to the disambiguation process. The tool developed for this purpose, Latch, is vaguely inspired in Constraint Grammars (Karlsson et al., eds., 1995), although constraints, of a weighted nature, are applied sequentially to the ambiguous input stream.

Currently, several grammars exist although all of them are stored in a single repository that is handled with a *Literate Programming* approach. This unique repository of rules can be arranged in a set of configurations allowing its use for modern texts or historic ones. Besides, there are specialized rule sets dealing with certain linguistic phenomena, like agreement, function, and so on. Other criterium used for the dynamic extraction of a given configuration is its reliable vs. heuristic nature. The number of rules in the repository is near 800.

Being very conservative in rule discrimination, recall has been kept to its maximum (higher than 99.0%), while precision is in the range 84-88% with still much work to be done to exhaust this strategy.

#### 3.4. Manually disambiguated (sub)corpora

In order to evaluate automatic disambiguation, a 1M word subcorpus has been drawn from CREA. This corpus prototype, manually disambiguated by a group of human posteditors, has the same design criteria as the whole CREA, thus it provides a good basis to test Latch and the set of disambiguation rules developed within it. Manual disambiguation for every text has been double, so remaining errors have been minimized. Besides, this methodology has allowed post editors the development of an agreed annotation scheme that has benefited also the production of disambiguation rules.

Currently, a similar prototype corpus is being produced for CORDE, with more than 150,000 words already disambiguated.

## 3.5. Other tools

### 3.5.1. Tokenizer generator

A high level declarative formalism has been designed in order to write tokenizer specifications. Knowledge about tokens and their contexts is expressed using typed feature structures and regular expressions. A macro definition language (m4) has been plugged, thus allowing developers to create new constructs. Literate programming style has been for documenting and modularizing specifications. A generator to a lex specification is implemented as an intermediate step towards a C program that implements operationally fast deterministic tokenizers. Assuming that analysis begins at the very first character, linguistic analyses have been integrated in the process as well as guessing strategies. Discourse structure conventions and text structure are modeled using contextual orthotypographical string information. Finally, the ecology of tokenization classes is similar to that used in information extraction systems.

### 3.5.2. NP extractor

The extraction of a set of generic tools for lexicogrammatical chunks has been designed and implemented. A first version of this tool is capable of recursively extracting NPs from annotated corpora. It is being planned how to use this type of information, specially those NPs coming from the manually disambiguated prototype, during the disambiguation process, since it is felt that a cooperative model can be constructed where the disambiguation of certain difficult ambiguity classes (for instance, noun and adjective, a very fuzzy class in Romance languages) can benefit, like in DOP approaches (Bod, 1995) from a collection of previous analyses.

Moreover, this tool serves also the purpose of generating candidate term constructions coming from the special language corpus.

## 3.5.3. Spell checking

The Computational Linguistics Department is also working on search algorithms for spell checking on Spanish texts. Both the linguistic data collected from the corpora (specially systematic errors found in ephemera sections of the CREA) and the tuning of candidate search so as to reflect orthographic hesitations of common users besides the simple character distance have been taken into account in this new class of algorithms.

#### 4. In house exploitation

### 4.1. Corpus-based lexicographical workbench

One of the principal benefited projects carried out has been the computerization of the twenty-first edition of the *DRAE (the Academic Dictionary of Spanish language)*. To this aim, a set of applications has been developed making the basis for the new corpus-based lexicographic platform. The future editions of the academic dictionary will benefit from a system specifically designed to guarantee the coherence, lexical actuality and uniformity of the actual lexicographic approach.

The dictionary is included in a relational database running on UNIX platform. The system has been designed according to the Client-Server model, which conveys a number of technical advantages. All the applications use advanced graphic tools, which enables a real reproduction of the graphic characteristics of the dictionary.

In a traditional dictionary, each article constitutes a textual information differentiated block of by Starting typographical changes. out from these characteristics, a data structure has been generated. The lexicographical entry has been organized in its elementary units, which allow the recovery of the information by way of the characteristics. A complex set of coordinated tables through a referential constraint system breaks down each article in its basic units, giving the dictionary a coherence which will avoid in future editions typical lexicographic mistakes, i.e. remissions to deleted dictionary entries.

Each dictionary entry will contain until tree composite elements: *lemma, meaning* and *complex form*. At the

same time, each of these elements adopts the following structure:

Lemma: Etymology, gender, number, variations and other information.

Meaning: Etymology, abbreviations, definition and examples.

Complex form: Complex form lemma and complex form meaning.

The relational model has multiple advantages in the lexicographical work, the most important are the integrity control, the structural coherence, the revision and search facilities and the edition flexibility and reliability.

The lexicographer's workbench includes the following applications:

Search system, update system, internal cross-reference system, pattern design system, hierarchical structuring system and edition system.

The *search system* is a user-friendly graphical interface that generates *SQL* queries, which provides a wide range of possibilities in the dictionary exploration.

Additionally, the system includes a dedicated corpussearching tool completely integrated with the lexicographer's workbench. The program offers the typical functionality of the concordance systems and adds some extra functions specially designed for the lexicographical work. A filter set allows the user dynamically select subcorpus and combine chronological, thematic, geographic and author criteria.

The query language includes Boolean and proximity operators, regular expressions, wildcards, etc.

Concordances and bibliographic references can be classified, summarized, printed or exported. Several statistical hit reduction functions complete the system facilities.

An experimental tool has been developed recently in order to explode the morpho-syntactically-annotated version of CREA. A new filter set performs "not form dependant" morpho-syntactic corpus queries.

The *update system*, that is the basis of the dictionary maintenance, allows the creation of "lexicographic units" that will later update the database. Using the *cross-reference system*, it is possible to establish internal relations between the lemmas: equivalence, opposition, etc. The *pattern design system* allows the creation of definition models in order to standardize the redaction of "word families". The *hierarchical structuring* application arranges and organizes the entries according to the lexicographer's criteria. The *edition system* displays and prints the dictionary in several ways, preserving the typographical characteristics (non-Latin fonts, old symbols, etc.).

#### 4.2. Lexicographic use of corpus materials

As a first approach to its new dictionary design, RAE has been working with word lists (both wordforms and lemmas) and concordances from the different corpora.

Thus, in the case of the secondary school corpus, the former information has been used to select the *lemmata* to be defined and the vocabulary to be used in the *definiens*. The use of frequency lists is a trivial way of approaching not only macrostructural issues in lexicography but also those affecting the microstructure of any dictionary.

However, the corpus is already being used to support other aspects of both lexicographic and grammar work. In this respect, a representative corpus is a very valuable source of information on paradigmatic hesitation by users on certain inflectional categories. For instance, certain irregular or defective verbs are actually used in a different way by native speakers than Spanish grammars pretend, and this information should, thus, be updated in all reference books produced from the corpus.

Besides, central to the new dictionary design is the use of naturally produced examples of language use. The corpus has already revealed itself as fundamental for the mining of recurring patterns in word concordances. This information, as already stated by (Sinclair, 1991), allows the frequency-based ranging of word meanings and also the extraction from the corpus of the most representative patterns of word usage.

Some of the tools described (NP extractor) and another set of tools currently being devised to help the terminologist in his/her work will be used in candidate term identification tasks for the sublanguage corpus and also in measuring the a priori relevance, for a given text, with respect to a given subdomain.

Finally, a long-time stemming effort that dates back to the 50s, which is the production of a historical dictionary for Spanish, will be the paramount exponent of the corpus-based methodology at RAE. Given that the Institution already drafted this dictionary, although only for words starting with a and b, making use, for this purpose, of its 13M lexicographic card file, the information contained in this draft version will be used to measure the representative of the lexicon contained in CORDE, in different periods. Those lexical items included in the draft dictionary and not in the corpus will serve as a seed to extend the corpus horizontally (as opposed to the vertical extension along the time axis). Once the corpus is considered to be fully representative when compared to existing lexicographic resources, a new dictionary will be developed based exclusively on the information drawn from the corpus.

### 5. Conclusions

Although both corpus development and linguistic annotation have already been used extensively to assist lexicographer's work in the English (especially British) lexicographic tradition, this is relatively new to the Hispanic world. Shy attempts have been performed by a couple of private editors, but it can be said that the work described and the methodology outlined in this paper is, taken together, new for Spanish. Moreover, while other projects aiming at the creation of language resources have been the result of a joint effort, it must be stressed that RAE, without giving up from taking account of other's achievements, has produced a bunch of corpora, with a high quality set of tools and other language resources around them, as an exclusive in-house effort (with funding from the Spanish government).

## 6. References

- BOD, R. (1995) Enriching Linguistics with Statistics: Performance Models of Natural Language, Ph. D. Thesis.
- BURNARD, L. (1995a). The Text Encoding Initiative: an overview. G. Leech, G. Myers, J. Thomas (eds.), Spoken English on Computer: Transcription, Markup and Applications, Harlow, Longman.
- BURNARD, L., SPERBERG-MCQUEEN, C. M. (1995). «*TEI* Lite: An introduction to Text Encoding for Interchange. electronic publication:

<<u>http://www-tei.uic.edu/orgs/tei/intros/teiu5.tei</u>>

<<u>ftp://info.ox.ac.uk/pub/ota/TEI/doc/teiu5.tei</u>>

- GOLDFARB, C. F. (1990). *The SGML Handbook*, Oxford, Clarendon Press.
- IDE, N. M., SPERBERG-MCQUEEN, C. M. (1995). The TEI: History, Goals and Future, *Computers and the Humanities*, 29/1, (pp. 5-15).
- IDE, N., (ed.) (1996). Corpus Encoding Standard, en *EAGLES Final Report*, electronic publication: <<u>http://www.lpl.univ-</u>
- aix.fr/projects/multext/CES/CES1.html>
- IDE, N., VÉRONIS, J. (eds.) (1995). The Text Encoding Initiative: Background and Contexts. *Computers and the Humanities*, 29, Dordrecht/Boston/London, Kluwer.
- JOHANSSON, S. (1995). The approach of the Text Encoding initiative to the encoding of spoken discourse, en G. Leech, G. Myers, J. Thomas (eds.), *Spoken* English on Computer: Transcription, Markup and Applications, Harlow, Longman.
- KARLSSON, F. et al. (eds.) (1995) Constraint Grammar: a Language-Independent System for Parsing Unrestricted Text, Berlin: Mouton de Gruyter.
- LLISTERRI, J. (1994). Spoken texts. Draft-Work in progress, *EAGLES document EAG-CSG/IR-T7.1*, en N. Calzolari, J.M. McNaught (eds.), EAGLES Interim Report. EAG-EB-IR-2.
- MARCOS MARÍN, F. et alii (1992). Corpus Oral de Referencia de la Lengua Española Contemporánea, electronic publication: <ftp://lola.lllf.uam.es>
- NERC (1994). Network of European Reference Corpora. NERC-1, Final Report, Pisa, ILC-CNR.
- PETITPIERRE, D. & Russell, G. (1995) MMORPH The MULTEXT Morphology Program, ISSCO, University of Geneva, *MULTEXT deliverable report for the task* 2.3.1, February.
- PINO, M. (1996). Encoding two large Spanish corpora with the TEI scheme: design and technical aspects of textual markup, *TEI Workshop at the ACM Digital Libraries '96*, Bethesda, Maryland, electronic publication:

<http://www.cs.vassar.edu/~ide/DL96>

- SINCLAIR, J. (1991). *Corpus, concordance, collocation.* Oxford, Oxford University Press.
- SPEECHDAT (1997). Working standards for speech databases directed towards short and medium term applications, *LRE-63314 Report D3.1.1.1*, electronic publication:

<<u>http://www.icp.grenet.fr/SpeechDat/deliv.html</u>>

SPERBERG-MCQUEEN, BURNARD, L. (eds.) (1994). Guidelines for Electronic Text Encoding and Interchange (TEI P3), Chicago/Oxford, Tei Encoding Iniative.