Written Linguistic Resources in Catalan: the DCC project

Joan Soler i Bou
Institut d'Estudis Catalans
Carme, 47
08001 Barcelona, SPAIN
[jsoler@iec.es]

Abstract

This paper is a general presentation of the most important lexicographical project that is being carried out in Catalonia: the DCC (Dictionary of Contemporary Catalan) project. After making a brief description of the different stages of the project, the specific goal of my contribution is to present the resources built for lexicographic purposes and, more specially, to stress the steps taken for the standardisation and harmonisation of the resources with other projects like PAROLE, or with EAGLES recommendations and TEI guidelines.

General outline of the DCC

The Dictionary of Contemporary Catalan (DCC) is a large scope lexicographic project that started at the beginning of 1985 in the Institut d'Estudis Catalans. The general aim of the DCC is the creation of a descriptive dictionary of the contemporary Catalan language. Up to this moment, the Catalan lexicography has remarkable resources as regards its different aspects: historical and dialectal (the DCVB), etymological (the DECat) and normative, but there is a lack in a specific work concerned with the synchronic description of the lexis without a normative perspective.

The DCC project corresponds to the activities related to the study of language, which are specific to the Institute, and accomplishes thus its institutional mandate. The IEC has the responsibility, acknowledged by the Catalan and Spanish national authorities, of updating and establishing a standard and normative Catalan. The carrying out of this institutional mandate means that the creation of linguistic resources as the basis for such activities becomes a crucial issue, which is behind most programmes and research projects fostered by the Institute.

According to this plan, the development of the DCC is structured in two stages which correspond, on the one hand, to the creation of linguistic resources that will enable the realisation of the project and, on the other, to the exploitation of these resources for the purposes mentioned above.

The first stage of the DCC project took place from 1985 to 1997, and its total cost came to more than 900 million pesetas (6 M ECUS). The second stage has an estimated duration of 8 years, and an approximate cost of 950 million pesetas. The DCC project has been financed with specific funds by the Science and Education Ministry of the Spanish Government and by the Generalitat (autonomous government) of Catalonia. Both administrations have signed, through the "Secretaria de Estado para Universitades e Investigación" and the "Comissió Interdepartamental per a la Recerca i el Desenvolupament Tecnològic (CIRIT)", specific agreements to carry out the several stages of the project. The participation of various administrations in a work of this kind has enabled the execution of the project as it was conceived, and represents an unprecedented experience in the field of Catalan linguistic resources.

First Stage: creation of the CTILC

Objectives

The *Corpus Textual Informatitzat de la Llengua Catalana* (*Catalan language machine-based textual corpus*) (CTILC) was created at the beginning of 1985 in the Institut d'Estudis Catalans. The elaboration of this corpus was conceived as the first stage of the DCC project, according to the methodological principles of current lexicography. Among these principles and as a main resource, there stands out the use of a sufficiently large and representative textual corpus of the language of the time one intends to study.

Characteristics

The corpus has a total extent of 52,3 million words, divided, approximately, into a 56% of non-literary language and a 44% of literary language. The criteria that have been followed to classify the texts are based on the subject and the genre of the texts. Thus, the literary language part of the corpus has been structured into four genres (corresponding to the four traditional literary genres: narrative, poetry, theatre and essay), whereas the non-literary corpus has been structured into ten subject groups (philosophy, religion and theology, social sciences, etc.). Each one of these ten groups can still be subdivided into ten subspecifications. In the table 1 the quantitative data of the specifically lexical information (that is, without proper names) included in each group is shown.

Typological group	Tokens	%
1 Philosophy	1.730.749	6,05
2 Theology and Religion	2.922.415	10,22
3 Social Sciences	5.490.046	19,20
4 Newspapers and Periodicals	3.447.517	12,06
5 Mathematics and Natural Sciences	2.189.052	7,66
6 Applied Sciences	4.422.174	15,47
7 Arts. Leisure. Sports and Games	2.735.839	9,57

Typological group	Tokens	%
8 Language and Literature	2.182.574	7,63
9 History and Geography. Biography	3.343.008	11,69
0 Letters	129.762	0,45
Total NON LITERARY	**28.593.136**	**55,79**
A Essay	2.996.755	13,22
P Poetry	2.471.281	10,91
N Narrative	13.579.181	59,92
T Theatre	3.613.327	15,95
Total LITERARY	**22.660.544**	**44,21**
TOTAL:	**51.253.680**	

Table 1

Chronologically, the corpus is not strictly referred to the present-day language, because it covers a period of 150 years, between 1833 and 1988. At the moment the basis for the permanent updating of the CTILC is being designed.

The characterization and chronological classification system designed for the CTILC is based on three chronological sections, which can be seen in the table 2, as well as the number of words (including proper nouns) that belong to each section. These divisions correspond to three different periods of development in the social use of the language, which are well differentiated and identified from the historical point of view.

Group	Subgroups	# of words
Section I (1833-1873)	4 groups (10 years)	2.260.083
Section II (1874-1913)	4 groups (10 years)	8.457.482
Section III (1914-1988)	15 groups (5 years)	41.654.379

Table 2

The number of texts included in the corpus is 3.387. They have been loaded almost entirely, but in some cases, due to the excessive length of the work, only significant passages have been taken into account. The edition that has been loaded is always the first one, and it is used as a reference to localize the textual information (work, page, line, position in the line).

It is a completely tagged and disambiguated corpus at a morphosyntactical level, specially designed for lexicological research. The high level of reliability based on the human validation of the information makes the CTILC be one of the most remarkable corpus in relation to other corpora of reference created in Europe.

A lexicographic database (BDLex), which contains 13 dictionaries published during the same time period that covers the corpus, complements the CTILC.

Second Stage: the writing of the DDesc

Objectives

The main objective of the second stage is to write a descriptive dictionary on the basis of the resources created in the first stage of the project. A descriptive dictionary is generally understood as a lexicographic work which intends the definition of the lexical units of a language, both from the point of view of their content and their actual use, without any kind of restriction based on prescriptive criteria.

This dictionary, which will have a complex nomenclature, is expected to deal with approximately 110,000 lexical entries, 70,000 of which will be main entries.

The writing of the DDesc will produce both a paper dictionary and a machine readable dictionary, which will include in a formalized way all the specifications necessary for its exploitation in the area of linguistic engineering. Thus, among the final products of the DCC project there is one directly related to linguistic resources: a lexicon provided with morphological, syntactic and semantic information; directly reusable as a linguistic resource for the development of applications for the Catalan language. This lexicon will contain information on inflected morphology, argument structures, selective restrictions, semantic frames and lexical relations. All this information will come from the formalization of the tasks of the compilation of the dictionary.

The experience acquired by the DCC working group in the creation of the PAROLE lexicon (20,000 lexical entries tagged at a morphological and syntactic level following the specifications of the EAGLES group), and in its continuation SIMPLE (which will add semantic information to the PAROLE lexicon), will be very useful towards the structural design of the Ddesc, and it will also allow its fundamental criteria be those applied in the creation of a group of lexicons for 12 European languages.

Characteristics

The DDesc will have the CTILC as its main source of information, according to one of the most generally accepted principles of current lexicography: the use of textual corpora. The BDLex will also be used as reference information. The specific lines in the first stages of the writing of the DDesc are the following:

a) Establishment of the frequency-based nomenclature.
b) Lexicographical criteria and specifications. Some criteria deal with the treatment of microstructural information (sense distinction and syntactic patterns within an entry, the nature of the definitions, sense classification, selection and adequacy of examples, treatment of lexical combinations, collocations and compounds, etc); some other refer to the mapping between the linguistic criteria applied in the tagging and lemmatisation of the CTILC and the macrostrutural specifications (differentiation of entries, graphical and formal variants); finally, several criteria are not based in a *linguistic* analysis of the corpus, but they intend to give complementary (i. e. statistical and metalexicographical) information to the entry.
c) Compilation of 650 dictionary entries. This specific task has the double goal of being a test for the phase of full compilation, and also of being a test for the lexicographical criteria.

Beside these lines of action, the project also integrates the design and implementation of a *lexicographical workstation*, which is the frame that will allow the

lexicographers the access to all the relevant information; moreover, this structure also includes tools for controlling and checking the different phases of the compilation.
The lexicographical workstation is the result of the integration of the following modules:

a) An SGML editor and browser
b) Assistance software (for the lexicographer)
c) Internal module for transferring data
d) Connectivity and accessibility to the different platforms of the network

In addition to these internal components, the workstation has also two external components: the two databases that the lexicographer will use for the information retrieval, CTILC and BDLex, corresponding respectively to two different kinds of data (textual and lexicographical). Both external components are the linguistic resources developed within the DCC. The interface of access to them has been specifically designed for lexicographical purposes.

Reuse of the resources

The CTILC

In the brief description of the CTILC given above I pointed out that the textual representation criteria, the morphosyntactic tagset, and the relational structure for the corpus data storing had been conceived and developed as a part of the project. With the purpose of adapting our corpus to the standards, an interface has been set up to convert the CTILC internal data to the EAGLES standard. Because of the experience acquired in the constitution of the Catalan corpus of PAROLE, we have followed the implementation of the EAGLES recommendations made for the PAROLE project.
The main goal of this task has been the reuse of the CTILC data for general and specific purposes, and the possibility of exploiting the resources on different platforms. The specific subtasks have been the following:

a) Evaluation of different formats of SGML text representation
b) Establishment of the Document Type Definition of the output
c) Implementation of the conversion interface
d) Tests of transportability into *conventional* browsers

The PAROLE model was adopted as the basis for the textual representation. This model, based on an implementation of the TEI guidelines, gathers the most important points of the EAGLES CES (*Corpus Encoding Standard*). Instead of defining from scratch a new DTDs system, PAROLE uses the TEI mechanisms (in the same way as EAGLES) to define its own CES:

```
<!ENTITY % TEI.extensions.ent SYSTEM "parole.ent">
<!ENTITY % TEI.extensions.dtd SYSTEM "parole.dtd">
```

In spite of this “new declaration”, the few differences between EAGLES and PAROLE deal with the header of the documents and they are mainly caused by the fact that the final corpus specifications from EAGLES came after the starting date of PAROLE.
Upon these bases the conversion interface produces as a result a morphosyntactically annotated version of each text. One of the most important steps of this conversion is the translation of the morphosyntactical information. The tagging system of the DCC is based on the relation lemma/form; the mapping of this information into the EAGLES recommendations is fully automatic. In the converted text, the morphosyntactical information is included as a value of the attribute “msd” (referred to the element <w>) and the graphical form of the lemma is included as a value of the attribute “lemma” of the same element.
I present in two tables the adaptation of the Catalan tagset according to the EAGLES specifications. As for the values of the *msd* attribute, they can be seen in the table 3; the two columns on the left (*catgram* and *sscatgram*) stand for the morphosyntactical category and subcategory. These two pieces of information fill in the first and the second place of the tag. The rest of columns stand for the features that can be used in the tagging. Each figure in the table points to the position of each feature in the morphological tag, and the equals sign means “not relevant”.
The table 4 specifies the different features that correspond to each category.
This interface has allowed the Catalan PAROLE corpus reuse some texts from the DCC. As a matter of fact, all the morphosyntactically tagged part of this corpus (2,5 M words) comes entirely from the DCC corpus.

CATGRAM	SsCatGram	DEGREE	GENDER	NUMBER	CASE	PERSON	MOOD	TENSE	FUNCTION	FORMATION	POSSESSOR	POLITENESS
ADJECTIVE	QUALI	3	4	5	=	=	=	=	7	=	=	=
ADPOSITION	PREPOSITION	=	4	5	=	=	=	=	=	3	=	=
ADVERB	GENERAL	3	=	=	=	=	=	=	4	=	=	=
ARTICLE	DEFINITE	=	3	4	=	=	=	=	=	=	=	=
ARTICLE	INDEFINITE	=	3	4	=	=	=	=	=	=	=	=
ARTICLE	PERSONAL	=	3	4	=	=	=	=	=	=	=	=
CONJUNCTION	COORDINATIVE	=	=	=	=	=	=	=	=	=	=	=
CONJUNCTION	SUBORDINATIVE	=	=	=	=	=	=	=	=	=	=	=
DETERMINER	EXCLAMATIVE	=	4	5	=	=	=	=	=	=	=	=
DETERMINER	DEMONSTRATIVE	=	4	5	=	=	=	=	=	=	=	=
DETERMINER	POSSESSIVE	=	4	5	=	3	=	=	=	=	7	=
DETERMINER	INTERROGATIVE	=	4	5	=	=	=	=	=	=	=	=
DETERMINER	RELATIVE	=	4	5	=	=	=	=	=	=	=	=
INTERJECTION		=	=	=	=	=	=	=	=	=	=	=
NOUN	COMMON	7	3	4	=	=	=	=	=	=	=	=
NOUN	PROPER	=	=	=	=	=	=	=	=	=	=	=
NUMERAL	CARDINAL	=	3	=	=	=	=	=	6	=	=	=
NUMERAL	ORDINAL	=	3	4	=	=	=	=	6	=	=	=
PRONOUN	PERSONAL	=	4	5	6	3	=	=	=	=	=	8
PRONOUN	DEMONSTRATIVE	=	4	5	=	3	=	=	=	=	=	=
PRONOUN	INDEFINITE	=	4	5	=	=	=	=	=	=	=	=
PRONOUN	INTERROGATIVE	=	4	5	=	=	=	=	=	=	=	=
PRONOUN	RELATIVE	=	=	5	=	=	=	=	=	=	=	=
PRONOUN	POSSESSIVE	=	4	5	=	3	=	=	=	=	7	8
PRONOUN	EN	=	=	=	6	=	=	=	=	=	=	=
PRONOUN	HI	=	=	=	6	=	=	=	=	=	=	=
PRONOUN	HO	=	=	=	6	=	=	=	=	=	=	=
VERB	MAIN	=	7	6	=	5	3	4	=	=	=	=
VERB	AUX	=	7	6	=	5	3	4	=	=	=	=

Table 3: CATGRAM, SsCatGram, and the relevant features for each

Feature	Tag	Value: DTD
CASE	=	WITHOUTC
	N	NOMINATIVE
	A	ACCUSATIVE
	D	DATIVE
	O	OBLIQUE
DEGREE	=	WITHOUTDEG
	P	POSITIVE
	C	COMPARATIVE
	S	SUPERLATIVE
	I	INTENSIVE
	A	APPRECIATIVE
GENDER	=	WITHOUTG
	M	MASCULINE
	F	FEMININE
	I	INDISCRIMINATE
	C	COMMON
MOOD	=	WITHOUTM
	I	INDICATIVE
	S	SUBJUNCTIVE
	C	CONDITIONAL
	N	INFINITIVE
	G	GERUNDIVE
	M	IMPERATIVE
	P	PARTICIPLE
NUMBER	=	WITHOUTN
	P	PLURAL
	S	SINGULAR
	I	INVARIANT
PERSON	=	WITHOUTP
	1	1
	2	2
	3	3
TENSE	=	WITHOUTT
	P	PRESENT
	I	IMPERFECT
	S	PAST
	F	FUTURE
FUNCTION	=	WITHOUTFU
	M	MOD
	S	SPE
FORMATION	=	WITHOUTF
	S	SIMPLE
	C	COMPOUND
POSSESSOR	=	WITHOUTPOS
	S	SINGULAR
	P	PLURAL
POLITENESS	=	WITHOUTPOL
	P	POLITE

Table 4: Features, and their tag- and DTD-values

The BDLex

The BDLex project is a specific action within the DCC programme; its main goal is the constitution of a corpus of dictionaries published in Catalan during the last 150 years. This corpus will be integrated, as a complementary source of information, in the set of resources of the lexicographical workstation for the compilation of the descriptive dictionary. As stated for the CTILC, the internal use of the BDLex does not exclude other possibilities of such an important volume of linguistic data, encoded according to a system which allows a maximum degree of transportability.

The Lexicographical corpus which constitutes the BDLex is the result of the selection of 13 dictionaries (see table 5), that have been considered as the most relevant for the studied chronological period, and more interesting from the point of view of the compilation of the descriptive dictionary. In a second phase of the project, the set of selected dictionaries could be increased with the inclusion of new lexicographical works. The machine readable version of these dictionaries has been obtained through different procedures: scanning-OCR, manual keying-in, reuse of existing machine readable dictionaries, etc.

The global extension of the corpus is 607,705 dictionary entries, distributed among the dictionaries as shown in the table 5.

Dictionary title	# of entries
Amengual, Joan Josep. *Nuevo diccionario mallorquin-castellano-latin*. Palma: Imprenta y Librería de Juan Colomar, 1858. 2 v. (1878 p., 748 p.)	35491
Ecrig, José. *Diccionario valenciano-castellano*. Valencia: Imprenta de J. Ferrer de Orga, 1851. (900 p.)	73290
Esteve, Joaquim; Bellvitges, Josep; Juglà i Font, Antoni. *Diccionario catalan-castellano-latino*. Barcelona: en la oficina de Tecla Pla viuda, 1803-1805. 2 v. (419 p., 429 p.)	28637
Fabra, Pompeu. *Diccionari general de la llengua catalana*. Barcelona: Llibreria Catalònia, 1932. (1782 p.)	51945
Febrer i Cardona, Antoni. "Diccionári menorquí españól francês y llatí". [manuscrit, principis segle XIX]. (670 p.).	28992
Ferrer Pastor, Francesc. *Diccionari general*. València: Impremta Fermar, 1985. (1599 p.)	56615
Figuera, Pere Antoni. *Diccionari mallorqui-castella*. Palma: Imprenta y Llibreria de Esteva Trias, 1840. (626 p.)	30430
Hospitaler, Josep. *Vocabulario castellano menorquin y vice-versa*. Mahon: Imp. De Miguel Parpal, 1869. (300 p. aprox.)	6232
Institut d'Estudis Catalans. *Diccionari de la llengua catalana*. Barcelona: Enciclopèdia Catalana, 1995. (1908 p.)	67944
Labèrnia, Pere. *Diccionari de la llengua catalana: ab la correspondencia castellana y llatina*. Barcelona: Estampa dels hereus de la V. Pla, 1839-1840. 2 v. (988 p., 1040 p.)	45822
Martí i Gadea, Joaquim. *Diccionario general valenciano-castellano*. Valencia: Imprenta de José Canales Romá, 1891. (1975 p.)	113108
[Tarongí i Cortès, Josep]. *Diccionari mallorquí-castellá* per un mallorquí, aficionat á sa seva llengua. Palma de Mallorca: Imprenta d'en Bartomeu Rotger, 1878. [Inacabat] (784 p.)	20594
Vallès, Emili. *Pal·las: diccionari català-castellà-francès: amb vocabularis castellà-català francès-català*. Barcelona: Horta, [1927]. (1296 p.)	48605
Total	607705

Table 5

To give an idea of the size of the corpus we present the frequency of the different structural elements considered in the lexicographical database (table 6).

Structural element	Frequency
Dictionary entries	481808
Word entries	485355
Feminine mark	100574
Plural mark	4349
Inflectional mark (verbs)	867
Pronunciation	124
Graphical indication	40
Etymology	405
Grammatical category	440024
Topic labels	19192
Usage labels	43329
Definition	645386
Example	86203
Subentry	93645
Correspondence (Spanish)	352431
Correspondence (Latin)	191910
Correspondence (French)	25090
Note	76590

For the optimum exploitation of the Corpus, it was necessary to encode the relevant information in order to allow queries based on common patterns to all the dictionaries. This was to minimize the (methodological, graphical, etc.) differences existing among the dictionaries; very often these differences implied that the same information (the graphical form of a word, the grammatical category, etc.) could appear under different forms. In order to cope with these needs, the project developed tools and designed strategies in two concrete lines:

a) Normalisation systems: encoding links between the different forms of an item by referring it to a

normalised form.

b) Structural encoding: disambiguation and tagging of the logical information of each dictionary.

The main aim of the processes called normalisations is the association of a series of elements of the same nature, taken as they appear in the printed dictionaries; they are associated to a reference list of normalised forms shared by all the BDLex. The elements that have been normalised are the following:

- graphical form of the lexical entries
- grammatical categories
- topic labels
- usage labels

The establishment of the normalised form for each lexical entry has been a main task in the encoding of the dictionaries, and has required the implementation of a highly automated system. The BDLex dictionaries present a high degree of graphical variation. These variations are mainly due to the fact that some dictionaries were published before the normative codification of the Catalan language was made (Institut d'Estudis Catalans, around 1914). This fact implies that, very often, the same lexical entry could have many different graphical forms; in other cases, many lexical entries (even though they could be specific of one dictionary) appear in a form that does not follow the general principles of Catalan orthography. To give an example, we can find in the dictionaries the following lexical entries: *afaytar, afeytar, afeytarse, afeytar, afaitar* o *afeitar-se*, all being graphical variants of *afaitar* (to shave); the normalisation consists in linking each form to the normative one, which is taken as the canonical form in the BDLex. The number of these canonical forms is around half a million.

The rest of normalisations have implied a lower level of complexity, due to the fact that they refer to discrete lists of normalised forms. In the case of grammatical categories and usage labels, the normalised forms have been obtained by the comparison of the sets of labels used by the dictionaries. As for topic labels, a two-level conceptual taxonomy of normalised forms have been constructed in order to allow a granular or general grouping of the information.

The formal representation of the normalisations is encoded through an attribute *norm* linked to each element that has to be normalised. In the case of word entries, for instance, this attribute is associated to the element *orth* of the entry.

The structural encoding of the machine readable dictionaries has followed the model of the TEI guidelines on dictionary texts. The SGML declaration of TEI is specially addressed to published dictionaries, so several aspects have been more detailed within the project in order to facilitate the computational treatment of the dictionary data. The possibility of making extensions of the TEI DTDs has allowed the accomplishment of a double goal: on the one hand, to give a maximum level of reusability to the resources; on the other hand, to collect with a maximum degree of explicitness the logical structure of the dictionaries. The minor changes made in the TEI DTD can be summarised in the redeclaration of the following elements:

```
<!ELEMENT %n.entry;     - -     (form+, sense*)
+(seg,sb,spx,note,pron,etym,gramGrp,mt,cs,vr,rd,mn)
>
<!ATTLIST %n.entry;
        type            CDATA           "main"
        n               CDATA           #IMPLIED
        TEIform         CDATA           'entry'>
<!ELEMENT %n.sense;     - -     (def*,eg*,trans*)
+(form,note,etym,gramGrp)>
<!ATTLIST %n.sense;
        level           NUMBER          #IMPLIED
        n               CDATA           #IMPLIED >
<!ELEMENT %n.eg;        - - (q+,def*,eg*,trans*)
+(etym,num,seg,gramGrp,lbl,formula,p,note)>
<!ELEMENT %n.trans;     - -     (tr)  +(note)>
```

This redeclaration has been made with the purpose of restricting the possible contexts of the elements to allow a more reliable parsing of the dictionaries, and so on. All the structural information is encoded in order that it can be used in the exploitation of the database. At present all the BDLex dictionaries are available in TEI conformant format. An interface of access to the corpus designed within the DCC project is already integrated in the lexicographical workstation.

Final remarks

In this general presentation of the current status of the DCC project we can stress the following items as general conclusions that are especially important for language resources in minority languages:

a) Financial aspects: as stated above, the DCC project is being funded by the two main political administrations in Catalonia (Catalan and Spanish governments). The collaboration in this kind of project is absolutely necessary.

b) Availability and accessibility: as prior criteria, the Institut d'Estudis Catalans wants to facilitate as much as possible the accessibility and availability of the CTILC to the different users interested in corpus data. For this reason, the CTILC has been built as a multifunctional database that, apart from internal uses, can be reused and adapted to the different kind of purposes of Linguistic Engineering.

c) Standardisation and transportability: language resources, specially those referred to minority languages, in spite of its real *commercial* interest, must be as much as possible reusable for different working groups with different purposes. This is why projects with a large course of development must adapt their resources to the current standards; this has been a major goal of the DCC project.