## Producing LRs in parallel with lexicographic description: the DCC Project

## Joan Soler i Bou

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

#### Abstract

This paper is a brief presentation of some aspects of the most important lexicographical project that is being carried out in Catalonia: the DCC (Dictionary of Contemporary Catalan) project. After making a general description of the aims of the project, the specific goal of my contribution is to present the general strategy of our lexicographical description, consisting in the production of an electronic dictionary able to be the common repository from which we will obtain different derived products (the human dictionary, among them). My concern is to show to which extent human and computer lexicography can share descriptions, and the results of lexicographic work can be taken as a language resource in this new perspective. I will present different aspects and criteria of our dictionary, taking the different layers (morphology, syntax, semantics) as a guideline.

## **General outline of the DCC project**

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. Because of its nature, cost, structure, etc. the DCC should be considered with the status of a Catalan National Project. 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, etymological, 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.<sup>1</sup>

# The Descriptive Dictionary of Catalan and the DCC lexicon

### Objectives

The descriptive dictionary of the modern Catalan is being compiled using as primary source a textual corpus of 52 million words. One of the results of the project will be, then, a human usable dictionary with an estimated coverage of 80,000 main entries, but the design and the organisation of the description plan clearly states that this human dictionary will be derived from a multilayered electronic dictionary that can be considered as a Lexical Resource, with complete information on morphology (inflectional and derivative), syntax and semantics.

#### **Implementation and strategy**

The experience acquired in specific LRs production projects like PAROLE-SIMPLE has been largely taken into account in the design of the project. Up until now the relationship between Lexicography and the LRs (Ooi, 1998) production has been practically reduced to the strategies of extracting lexical information from existing dictionaries; projects like ACQUILEX (Boguraev & Briscoe, 1989) have made relevant research on how we can reprocess the information of a dictionary in order to produce computational data, lexical databases and knowledge databases. Other strategies of Lexicon constitution, developed in projects such as WORDNET (Miller, 1998) and EUROWORDNET (Vossen, 1998) have been largely based in lexicographic resources, too.

In the DCC project the strategy for the achievement of a Lexicon does not rely primarily on the reuse of existing

<sup>&</sup>lt;sup>1</sup> 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 "Secretaría de Estado para Universidades 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. The academic and scientific direction of the entire project corresponds to Prof. Joaquim Rafel i Fontanals.

resources. The DCC lexicon acquisition is based on the idea of the integration of lexical description (which is the most difficult and time consuming task) and the encoding of the lexical information. Depending on the richness and the criteria of this encoding we should be able to produce a lexicon organised in such a way that we can derive from it different kinds of results; one of these results being the human dictionary, but others being computational data that could be directly used in different NLP applications, developments, and products.

## The DCC lexicon

#### Characteristics

The idea of a description in parallel is supported by the evidence that most of the assumptions and purposes, at least with regards to descriptive aspects, are very similar between human and electronic lexica, especially those aspects concerning lexicon population (that can be essentially based upon the same assumptions), the structure of entries and general organisation of the data.

As for the different layers of description shared by both lexica, the DCC project has intended to be as much as possible compatible with the implementation of EAGLES standards made by the PAROLE-SIMPLE project.

The way in which the information in the DCC Lexicon is encoded and stored permits the generation of the structure taken as template for each dictionary entry. We can draw this template as two figures representing two organisational levels of the dictionary: the macrostructure level and the microstructure level. The structural elements considered as description items in the templates are filled by an specific description; the amount of these give a general characterisation of the word entry.

The DCC lexicon is supported by a relational database and



Figure 1: the DCC lexicographical entry

the encoding/browsing tool is implemented in scripts interacting with this structure. This data and this tool are a main component of the DCC lexicographic workstation, and cover different needs; among these, the more important are:

- on-line encoding of entries
- linking with the DCC corpus
- consistency checking and validation

- generation of results for different uses (SGML lexicon, human dictionary)

## The DCC Lexicon entry

From the macrostructure point of view, the information directly associated with the entry level follows the schema given in figure 1. Each entry has then a POS category, selected from a short list presented in figure 2, and an inflectional model pointing a process in the morphological analyser and generator.

adi	Adjective
1	
adv.	Adverb
afirm.	Affirmative
compar.	Comparative
conj.	Conjunction
det.	Determinant
f.	Feminine noun
interj.	Interjection
interr.	Interrogative
<b>m.</b>	Masculine noun
neg.	Negative
num.	Numeral
prep.	Preposition
pron.	Pronoun
pron. clít.	Clitic pronoun
pron. rel.	Relative pronoun
quant.	Quantifier
<b>v.</b>	Verb
v. aux.	Auxiliar verb

Figure2: List of POS categories

#### Statistical information

Generic statistical information is also conveyed within the information linked to the entry; each entry is marked with a rank code, which indicates how is the degree of importance in the vocabulary. The 5 ranks of vocabulary have been established on the basis of the Diccionari de freqüències (1996-1998)<sup>2</sup>, and takes into account not only the bare frequency of lemmas in the DCC corpus, but also the distribution of each lemma through the different typological groups of the corpus. This statistical classification of the vocabulary may be useful for the selection of entries in specific applications. The five groups (see figure 3) have been established upon the value of the use of each lemma. The use corresponds to the product between frequency and *dispersion index* (an index between 0 and 1, indicating the distribution of a given lemma through the different parts of the corpus).

<sup>&</sup>lt;sup>2</sup> The *Diccionary de freqüències* is a three-volume dictionary published as one of the first results of the DCC project, and gives the statistical data of the DCC corpus. The publication includes a CD-Rom containing all the data in a reusable format.

Rank	Use	Maximum frequency	Minimum frequency	
1	> 5288,3 = 5288,3	5.140.416	5.794	
2	< 5282,97 > 1225,23	9.878	1.378	
3	<1225,23 >175,75	4.507	200	
4	<175,71 >7,34	1.534	8	
5	=7,33 <7,33	70	1	

Figure 3: table of ranks

The morphological profile of the entry refers to the distribution of morphological categories in the instances of the lemma in the DCC corpus. In some cases this information expresses morphological deviations in the behaviour of lexical elements that are very difficult to explain in other terms. The categories and values taken into account are the following:

**number** (s = singular, p = plural);

**gender** (m = masculine, f = feminine);

**tense** (pr = present, ipf = imperfect, pf = perfect, fu = future);

**mode** (*if* = infinitive, ge = gerundive, pp = participle, *in* = indicative, su = subjunctive, co = conditional, *im* = imperative);

**person** ( $1 \dots 6$  = first to sixth person)

#### **Profile information**

The profile information is given in form of percentile relation between different values of a category: for nouns, the number; for verbs, tense, mode and person; and for adjectives number and gender. In some cases this information would help to explicitly give an account of the different lexical profiles shared for a same kind of morphosyntactic word, as shown in the example of figure 4.

gratitud	f.	Ν	100	0
esclavitud	f.	Ν	96	4
servitud	f.	Ν	78	21
habitud	f.	Ν	62	38
latitud	f.	Ν	57	42

Figure 4: Example of different morphological profiles for nouns ending in *-itud*.

## Morphology in the DCC Lexicon

## Inflectional information

The encoding of the inflectional morphology in the DCC is based on the same strategies as used in PAROLE lexica (this means the GENELEX project). A code of inflectional model of each entry gives information on its morphological behaviour, which is not totally predictable on the basis of the pure canonical form and the grammatical category. This encoded information is a pointer that refers to a word ending table that clearly specifies the link between each ending and the subindex of the stem selected. The code is accompanied by a list of different stems ordered and subindexed in accordance with the tables.

The morphological description of a lexical item, then, consists, on the one hand, in a reference to an inflectional table and, on the other hand, in a set of subindexed stems that combine with the relevant subindexed terminations. Past participles are treated separately with an specific reference to one or more of the 8 tables that identify the different processes of participle generation in Catalan verbs.

Here follows a developed example of the morphological information given for several verbs; the alphanumerical reference to the table of endings; after that, in brackets, the list of alternating stems; a number in square brackets after the stem indicates the model of participle that use the selected stem; if this number is accompanied by a morphological code, this means that the stem is selected only for the formation of the indicated form:

**cantar** *v*. 101a {cant<sub>0</sub>[1]} **abrigar** v. 101b { $abrig_0[1]$ ,  $abrigu_1$ } **donar** *v*. 101c {don<sub>0</sub>[1], dón<sub>1</sub>} admetre v. 201a {admet<sub>0</sub>, admè<sub>1</sub>[5ms], adme<sub>2</sub>[5fs,mp,fp]} **apercebre** *v*. 201b {aperceb<sub>0</sub> [8], apercep<sub>1</sub>} **romandre** v. 202a {roman<sub>0</sub>, romand<sub>1</sub>, romanc<sub>2</sub>, romangu<sub>3</sub>, romà<sub>4</sub>[5ms], roma<sub>5</sub>[5fs,mp,fp]} **absoldre** *v*. 202a {absol<sub>0</sub>[6], absold<sub>1</sub>, absolc<sub>2</sub>, absolgu3} **compondre** *v*. 202a {compon<sub>0</sub>, compond<sub>1</sub>, componc<sub>2</sub>, compongu<sub>3</sub>, compos<sub>4</sub>[7]} confondre v. 202a {confon<sub>0</sub>, confond<sub>1</sub>, confonc<sub>2</sub>, confongu<sub>3</sub>, confó<sub>4</sub>[5ms], confo<sub>5</sub>[5fs,mp,fp]} **adormir** *v*. 301a {adorm<sub>0</sub>[3]} **consumir** *v*.  $301a \{ consum_0[3] \} 305 \{ consum_0[3] \}$ **morin** v.  $301a \{ mor_0[6] \}$ **acollir** v. 301b { $acoll_0[3]$ ,  $acull_1$ } escollir v. 301b {escoll<sub>0</sub>[3], escull<sub>1</sub>} 305 {escoll<sub>0</sub>[3]}

The total number of standard verbal models is 66. The DCC lexicon, moreover, contains information on the inflectional expansion of Catalan regional variants, each variant consisting in a full system of models and tables: 69 the Valence regional variant, 76 for Balearic regional variant and 66 for the Northern regional variant.<sup>3</sup>

As for the nominal models, the encoding system follows the same principle; the total number of tables is 29. The stems subindexed as 0 always corresponds to the form of the lemma, this is why is ommited in surface representation:

**sofà** *m*. 1a **important** adj. 1a **pa** *m*. 1b {pan<sub>1</sub>} **acció** *f*. 1b {accion<sub>1</sub>} **cas** *m*. 3a **ús** *m*. 3b {us<sub>1</sub>} **monòton** *adj*. 11a

<sup>&</sup>lt;sup>3</sup> These are the principal regional variants recognised and treated in the normative grammar of Catalan that the IEC is near to publish.

únic adj. 11b {úniqu<sub>1</sub>} aeri adj. 11c {aèri<sub>1</sub>} ple adj. 11c {plen<sub>1</sub>} groc adj. 11e {grog<sub>1</sub>, grogu<sub>2</sub>} queviures m. 8 vacances f. 8 capaç adj. 9

## Syntax in the DCC Lexicon

The POS inventory of categories, as shown in figure 2, consists in a few elements, that are used as main criteria for entry splitting purposes. All the relevant subcategorising information of any kind (such as, for instance, transitivity, noun class, pronominalisation, etc.) is encoded at the sense level, in such a way that each sense is fully described in terms of its syntactic behaviour. The structure of the sense level is represented in figure 4.

The syntactic pattern of each sense of the entry is indicated in a formalised way in the DCC lexicon. The strategy followed in representing the information is based in the philosophy of describing the subcategorisation patterns in terms of *descriptions* that contain a list of syntactic *positions*. This approach is highly compatible with GENELEX architecture, once again according to the implementation made by PAROLE-SIMPLE project; this means, on the one hand, the specification of a number of positions of a construction that gives the information on the surface form of each reading; and, on the other hand, a number of syntactic alternations (like ergativity, equi



Figure 5: The sense structure

alternation, and so on) that give the information on deeper aspects of a reading of a lexical element.

A syntactic pattern in the DCC lexicon is an expression that describes the relevant characteristics of the insertion

context of a lexical entry; this information is encoded at the sense level, and takes as nucleus (a kind of SELF, in GENELEX terminology) the category of the described lexical item. The positions in the pattern are expressed in terms of categories taken from the POS (N, V, ADJ., etc.) defined in the categorisation schema of the dictionary entries; but here the POS tags are taken as syntactic constituents (in most cases projections of simple categories). Thus, N stands for an argument that should be instantiated by a noun, a noun phrase or a nominative pronoun, V refers to a verb or a verbal phrase. The patterns can contain, also, lexical literal elements when needed (bound prepositions, relative pronouns, and so henceforth).

Syntactic patterns include, also, all the relevant information on each argument position; for instance, in the case of adjectives, the pattern can give information on:

- positional constraints (pre/post-position with regard to the noun),

- selection of copulative predication verb (ser/estar);
- gradability
- prepositional subcategorized arguments

Then, a pattern of one reading of an adjective like *encertat* (right) in phrases like *En Pere ha estat encertat en la selecció del tema de l'examen (Pere has taken the right decision in the selection of the topic for the exam)* can be represented as follows:

#### N<sub>1</sub> està ADJ en N<sub>2</sub>/Vinf

The syntactic patterns also deal with especial properties of the positions, like *alternations* (represented by the slash) and *optionality* (in brackets).

In the case of verbs the pattern includes information of many aspects of verbal syntax, including passivisation, control and correference, pronominalisation of arguments, and so on.

Even the number of individual patterns is, as one can expect, relatively high, a main advantage of this treatment is, apart from the exhaustiveness, the possibility of grouping the patterns taking different criteria, like, for example, the number of arguments. The encoding/edition tool implements many facilities to work with this information. As example of different verbal patterns with two arguments and no especial characteristics (like optional, subject predicative verbs, and so on) can be seen in figure 6.

The DCC lexicon also defines relationships between two patterns, following the same strategy as PAROLE Framesets. This permits the descriptive system to deal with syntactic alternation phenomena like *ergativity*, *equi alternations*, and *symmetry transformations*. Here follow some examples:

#### *a)* Ergativity:

 $[N_1 \ V \ N_2] \Leftrightarrow [N_2 V \text{pron}] \ El \ vent \ ha \ tombat \ l'estaca \ / l'estaca \ s'ha \ tombat$ 

#### b) Equi alternations:

 $[N_i \ V \ que \ Vsubj_j] \Leftrightarrow [N_i \ V \ de \ Vinf_i]: va \ tractar \ que s'animés / va \ tractar \ d'animar-lo$ 

#### c) Subject symmetry transformation:

 $[N_1 \text{ Vpron } a N_2] \Leftrightarrow [N_{1+2} \text{ Vpron}]: en Pol s'assembla a la Clara / En Pol i la Clara s'assemblen$ 

#### *d)* **Object symmetry transformation**:

 $[N_1 V N_2 amb N_3] \Leftrightarrow [N_1 V \text{pron } N_{2+3}]$ : compara aquest dibuix amb aquell altre / comparar aquest i l'altre dibuix / comparar els dos dibuixos

N V N
N V N <sub>no pass</sub>
N V a N
N V a / cap a N
N V ADJ
N V ADJ/Vger
N V ADV
N V de N
N Vpron N
N Vpron <i>a/en</i> N
N Vpron <i>a/fins a</i> N
N Vpron amb N
N Vpron <i>contra</i> N
N Vpron de N
N Vpron en N
N Vpron <i>per</i> N
N Vpron ADJ
N Vpron ADV
N Vpron Int Vind
N Vpron de Int Vind
N Vpron <i>que</i> Vind
N <sub>i</sub> V a Vinf <sub>i</sub>
$N_i$ <b>V</b> pron ( <i>de</i> ) Vinf <sub>i</sub>
N <sub>i</sub> <b>V</b> pron <i>a</i> Vinf <sub>i</sub>
N <sub>i</sub> <b>V</b> pron <i>que</i> Vsubj <sub>i</sub>
$\mathbf{V} \ li_{i} per \ Vinf_{i} (a \ N_{i})$
$\mathbf{V} \ li_{i} \operatorname{Vinf}_{i} (a \ N_{i})$
<b>V</b> pron li <sub>i</sub> ( <i>de</i> ) Vinf <sub>i</sub> ( $a$ N <sub>i</sub> )
Vpron <i>li</i> <sub>i</sub> <i>Int</i> Vinf <sub>i</sub> ( <i>a</i> N <sub>i</sub> )
<b>V</b> pron $li_i N_{conc} (a N_i)$
<b>V</b> pron $li_i$ que Vind (a N <sub>i</sub> )
N Vrefl

Figure 6: Patterns of verbs with two arguments

## Semantics in the DCC Lexicon

Even though the semantic layer is not fully formalised for computational uses, the DCC lexicon includes some information that will allow the generation of relevant data in a computational form.

Semantic and lexical constraints over positions of syntactic patterns are specifically indicated as selectional constraints. Lexical constraints associate an element in the pattern (like the example below: *cavaller* 'knight' in **adobar**), whereas semantic constraints identify: *a*) the semantic class to which an element should belong (like *humà* 'human' in the example of **adobar**), or *b*) the hyperonym shared by all the lexical items that could appear in this position of the pattern (like *roba* 'clothes' in the example of **balder**).

adobar v. [...]  $[N_1 V N_2 N_3]$  ( $N_{1,2}$ [humà];  $N_3$ =cavaller) Armar [algú]<sub>2</sub> [cavaller]<sub>3</sub>. balder *adj*. [N *li*<sub>i</sub> va **ADJ** (*a* N<sub>i</sub>)]; (N[roba])

Lexicographic definitions in the DCC are formed, as usual in modern dictionaries, by two kinds of discourse elements: *intrinsic elements* (which have to accomplish the interchangeability principle) and *extrinsic elements*. Extrinsic elements are represented in square brackets (see *algú* 'someone' and *cavaller* above) and made the needed semantic or syntactic adjustments in the intrinsic definition in order to accomplish the principle of synonymy, introducing semantically necessary elements. As one can see, in the case of the DCC every intrinsic element in the definition is subindexed with a number that makes reference to one of the positions specified in the pattern.

The genus (*descriptor*) of the definitions is always specifically marked. This will allow to execute processes of coherence checking, verifying hierarchical relations in the dictionary, and so on. The descriptor is the element of substantial definitions that ensures the endocentrism (and functional coherence) between *definiens* and *definendum*. The marked descriptor usually consists in the hyperonym of the defined word, but in some cases it can correspond to different semosyntactic relations (like causativisation, part-whole relations, etc.) that are expressed by complex descriptors like:

<Conjunt de X> <En forma de X> <Fer fer X> <Fer tornar X> <No fer> <Tros de x> etc

Moreover, the links between "related senses" (such as metonymic alternations) are also indicated in the lexicon. These links are very similar to the semantic regular alternations treated in recent lexicon theories (Pustejovsky, 1995).

al) Container  $\Rightarrow$  Quantity of the containee. [Ncompt de N<sub>1</sub>] (dues ampolles de vi; dos prestatges de llibres)

a2) Drink  $\Rightarrow$  Bottled drink. [Ncompt] (posa'm una cervesa; em queden dues llimonades a la nevera)

*a3)* Container / Containee. [Ncompt] (*em quedo tot el prestatge; he de traslladar encara dos despatxos; tot el teatre es va posar dempeus; les dues escoles han anat d'excursió al mateix lloc*)

*b1*) Animal, fruit / Food. [Nmass] (*no m'agrada el pollastre*; *pastís de poma*; (*un plata d'*) *enciam*)

*b2)* Animal, tree / Substance (made of). [N *de* Nmass] (*sabates de cocodril; armari de noguera*)

*c1)* Vegetable product / Plant (producer). [Nmass] (*una plantació d'arròs*)

*c2)* Vegetable product, flower / Plant (producer). [Ncompt] (*una plantació de cacauets*)

*d*) Fruit / Seed (contained in). [Ncompt] (*no mengis tantes ametlles*)

*e1*) Plant, part of a plant, fruit / Drink (made of). [Nmass] (*t'agrada el cafè*?)

*e2)* Commercial activity / Products. [Nincompt] (*la joieria és cara*)

f) Company, *activity / Place*. [*el/la* Ncompt] (*he d'anar al banc*)

g) Substance / Physical object (made of). [Ncompt] (he comprat dos bronzes; un esmalt)

*h)* Feeling / token of a feeling. [Ncompt] (l'havien fet objecte de burles i menyspreus).

## **Final remarks**

Modern lexicography and modern dictionaries have stressed the similarities between human and electronic lexica. Computational Lexicography have made research assuming that is possible to (automatically) acquire data for NLP from existing dictionaries. Moreover, projects like PAROLE-SIMPLE have started to achieve electronic lexica via *ad hoc* encoding, in a similar way that "traditional" lexicographers do.

Lexicographic description is expensive and timeconsuming. Given that the object to describe is the same, we can encode the linguistic information in such a way that could be used for any specific purpose. The exhaustiveness of human dictionaries with regard to the number of entries, and the exhaustiveness of computational lexicons with regard to linguistic descriptions should encounter a common place that should quantitatively and qualitatively increase the reliability of both activities. The lexicographical phase of the DCC project is making an attempt to find strategies of description that could be used in the near future in both directions.

#### References

- Boguraev, B. & Briscoe, T. (1989) (Eds.): *Computational Lexicography for Natural Language Processing*, Longman, London.
- Fellbaum, C. (1998): *Wordnet: an Electronic Lexical Database*, MIT Press, London.
- Ooi, V. B. Y. (1998): *Computer Corpus Lexicography*, Edinburgh University Press, Edinburgh.
- Pustejovsky, J. (1995): *The Generative Lexicon*, MIT Press, Cambridge (Massachusetts).
- Vossen, P. (1998): *EuroWordnet: an Electronic Lexical Database*, Kluwer Academic Publishers, Dordrecht.