Editing, Accessing and Browsing The *diccionari descriptiu de la llengua catalana* (ddlc)

Teresa Sadurní i Villaronga
Josep M. Domènech i Gibert
Institut d'Estudis Catalans
C. del Carme, 47
E – 08001 Barcelona
tsadurni@iec.cat
jdomenech@iec.cat

Abstract
This paper describes the interface for the support of the dictionary-making process and access to a corpus-based contemporary Catalan dictionary, the *Diccionari Descriptiu de la Llengua Catalana* (DDLC). The DDLC started as the pre-eminent lexicographical project at the Institut d'Estudis Catalans (Barcelona) in 1998. One of its most remarkable characteristics is that it has been conceived from the very beginning as an electronic database, where the "written form" of the dictionary, in addition to other reports or query results, is automatically generated from the data structure. We mainly focus on four of the interface modules: *a)* functionality for adding and editing DDLC entries; *b)* data validation and internal tracking of the dictionary-making process by statistical means; *c)* browsing the DDLC entries, and *d)* querying the database. The DDLC interface is accessible at http://dcc.iec.cat/ddlc. Registration is required for external users, and the functions available vary depending on user privileges.

1 The DDLC

The DDLC (*Diccionari Descriptiu de la Llengua Catalana*) is based on a corpus of more than 50 million words, the *Corpus Textual Informatitzat de la Llengua Catalana* (CTILC), which was developed at the IEC as a previous stage for the dictionary compilation (cf. Rafel, 1994).

The DDLC is currently being compiled. Around 27000 entries have been written so far (March 2005), representing 25% of the complete dictionary. Although some of the data is still under revision, the entries are already available on the internet, which allows external users to send feedback at any time. Given that the interface was conceived as an intranet structure and later it was published over the internet, different user privileges have been established to access the various resources. Thus, for example, DDLC lexicographers have some built-in queries and are able to add and edit entries, whereas an external user is only allowed to browse them.

This interface is a part of the lexicographical workstation, which also includes access to other linguistic resources, such as the CTILC (textual corpus), online dictionaries, lexicographical repositories, the DDLC specifications, etc.

The demonstration is structured in 4 parts, which match the 4 main interface modules:

(1) Adding and editing DDLC entries

(2) Data validation and internal tracking of the dictionary-making process by statistical means

(3) Browsing the DDLC entries

(4) Querying the database

2 Adding and editing DDLC entries

The DDLC was conceived from the very beginning as an electronic dictionary, and therefore all the data are stored in a database. This gives us the flexibility to produce different kinds of outputs, depending on the requirements we need. It also increases the internal coherence of the data and its further validation, as well as provides a user-friendly way for lexicographers to add new entries and especially to edit them.

We will first see some of the functionalities available to add and edit the dictionary entries, in this case considered as a sum of different structural elements (syntactic pattern, constraints on this pattern, collocations, definition, examples, derived forms, phrases, etc).

The interface has been built up taking these structural elements into account and the lexicographers have access both to the specific piece of information itself and to the related information. They can browse the whole entry any time, since the "written form" of the dictionary is automatically generated from the data structure.

We start by selecting the entry of the dictionary itself, and the data available for browsing or editing is structured in the following way:

- Header: headword, part of speech; inflectional information; usage range: an index from 1 to 5 that shows the relative importance of the word within the corpus; morphological profile: usage of the different inflectional forms.
- Senses: definition, syntactic pattern, semantic constraints on the pattern, collocations, examples, syntactic conversions.
- Phrases: headword, part of speech; senses: same structure as the main entry, except syntactic conversions.
- Variations on the headword.
- Derived forms: headword; part of speech; type of process; sense on which the derived form can be applied; collocations; examples.
- Additional information: senses found in the lexicographical repository which have not been documented in our corpus.
- Subsidiary entries: derived forms or formal variations on the headword. These entries only contain a link to the main entry.

For further explanation of the data structure, cf. Rafel, Soler (2006).

3 Data validation and internal tracking of the dictionary-making process by statistical means

This functionality is used exclusively by the coordination staff in the DDLC project in order to follow the dictionary data and compilation process up. It includes features such as:

• Task management: assignment of entries; retrieval and change of the status of the entries, etc.

• Entry management: list of currently written entries; status of an entry; list of entries written as a group, etc.

• Compilation process follow-up: annual report (amount of entries written per month, average per month, average per year, etc.); statistical information of the lexicographer (total amount of entries per lexicographer, total amount of entries per year, total amount of entries per lexicographer and year, average of entries per lexicographer and month, average of entries per lexicographer and year, etc.); list of entries per lexicographer (can be limited to a certain period of time); statistical information of the data (total amount and average for each piece of information), etc.

• Queries: definitions lexicon; information about links within definitions; information about examples (list of publications or authors – included in the corpus – quoted in the dictionary – sorted alphabetically or by the amount –, publications not exemplified in the dictionary, statistical information according to typology, period of time, etc.); distribution of the usage range within the entries, etc.

4 Browsing the DDLC entries

This functionality is available for both internal and external access and it corresponds to the platform for publishing the DDLC over the internet. Internal users can browse all the entries that have been created so far, including the ones which are being written, whereas external users can only access the entries that have been delivered as finished by the lexicographer. We must remark here that the process of data validation and internal coherence check-up is done in parallel with the dictionary compilation, and therefore some entries could be modified later.

Entries may be accessed either by typing the entry or by typing a pattern to get a list of entries that match the pattern. The dictionary entry is generated automatically live from the data structure.

This online version of the DDLC makes use of the electronic format in several ways: to retrieve the necessary information to understand some elements of the structure, to access another sense in a synonymic definition, to check which other entries have a link to the current one, to obtain a list of other senses with the same genus or pattern, or to obtain the full reference of the publications quoted as examples.

5 Querying the database

Querying the database is only available for members of the DDLC staff. It is a set of built-in queries that sometimes can be parameterized, especially for sorting the results or filtering information.

Nowadays, this functionality includes queries such as:

• definitions with the string x within the extrinsic element

• definitions with the genus x

• definitions that contain the string x

- senses with the syntactic pattern x
- senses with the pattern constraint x
- patterns with the modifier x
- constraints applied to the part of speech x
- phrases that contain the string x
- derived forms that contain the string x and/or that follow the process y
- collocations that contain the string x

This functionality is both used in the process of writing the entries (since most of these queries were built to cover the needs of the lexicographers) and for further validation of the data in order to give coherence to the whole work.

Our demonstration shows a sample of the currently available functionalities. We mainly focus on the adding and editing entries module because it is the most important part in the interface and that which displays the structure of the dictionary.

References

Rafel, J. (1994), 'Un corpus general de referència de la llengua catalana', Caplletra 17, pp. 219-250.

Rafel, J. (dir.) (1996-1998), *Diccionari de freqüències*. 3 vols + 2 CD-ROM. Barcelona, Institut d'Estudis Catalans.

Rafel, J. (2000), 'El Corpus Textual Informatitzat de la Llengua Catalana i l'activitat lexicogràfica de l'Institut. Aspectes descriptius i aspectes normatius', in *Jornades per a la Cooperació en l'Estandardització Lingüística*. Barcelona, Institut d'Estudis Catalans, pp. 197-205.

Rafel, J., Soler, J. (2006), 'A Descriptive Dictionary of Contemporary Catalan: the DDLC Project', in *Euralex 2006 Proceedings*.