# Corpus Exploitation Strategies for the Lexicographic Definition Task

Judit Feliu, Àngel Gil; Berta Pedemonte; Cristina Guirado
Institut d'Estudis Catalans

*The main goal of this paper is to formalize and to present some guidelines helping the lexicographer corpus query procedures in order to obtain the most refined results aiming at benefit the definition as much as possible. The paper will briefly introduce the three main language resources (LR) involved in the authors' daily linguistic job, that is, the Catalan descriptive dictionary built on the basis of a Catalan corpus, the corpus itself and the Catalan main dictionaries repository normally looked up. The focus of the paper will be put on the improvement of the strategies followed so far in order to use the corpus query system in an efficiently oriented manner as far as the descriptor selection and the extrinsic part of the definition fulfilment is concerned. Bearing in mind a concrete number of patterns in order to retrieve certain kind of information for each type of unit defined will help the lexicographer to keep coherence among different team members and also, and probably most important, among different but semantically related types of words.*

## 1. Introduction

The first part of this paper is devoted to briefly present the *Diccionari Descriptiu de la Llengua Catalana* (DDLC) and the Catalan corpus used in order to build the dictionary. Also the dictionary repository consulted for the semantic aid will be described. Main focus, however, will be put on the proposal of some strategies offered to the lexicographer when working with corpus to obtain the best results interrogating the corpus. The descriptive dictionary and, in turn, each lexical entry and each sense is provided with a definition and at least one example and the information retained from corpus should be the most appropriate for each sense. The authors will shed some light in order to attain the most refined contexts in each case for the ever hard task of defining a word or for selecting and separating different senses of a lexical entry. Finally some conclusions and future research lines will be also mentioned.

## 2. LR Description

### 2.1. The *Diccionari Descriptiu de la Llengua Catalana* (DDLC)

As stated in Rafel&Soler (2006), the *Diccionari Descriptiu de la Llengua Catalana* (DDLC) is a corpus-based Catalan dictionary currently developed at the Institut d'Estudis Catalans (IEC). The DDLC is conceived as a *descriptive* dictionary, which means that its goal is to give a 'real and complete definition of each lexical item, without any restrictions based on prescriptive criteria'. As for the body of a lexical item, every article must have at least one *sense* accompanied by a lexicographical description (syntactic pattern and semantic constraints, collocations, definition and exemple(s)). Taking this clue idea into account, the authors of the paper have oriented their research to the refinement of the kind of information extracted from corpus samples that are analysed in order to define each of these lexical items.

In the DDLC general framework, the definition field for the case of nouns can be imagined as a 'lexical chain' constituted by two 'slots', one for the main unit of a definition (*descriptor*) (which can be single, such as *place*, or compound, like *group of people*) and the other one for the complementary information, mostly but not uniquely occupied by the extrinsic information (Rafel, 2006). See the following two examples for the different types of the mentioned descriptors which correspond to a hyperonym and to a meronym case analysed in depth later in the following paper sections:

**esclerosi** *f.* **1**. [**N**INCOMPT (*de* N1)] (N1[òrgan del cos]) Enduriment patològic [d'un òrgan]1. *~ renal, ~ arterial. [...] després d'extirpats la bufeta i els càlculs amb drenatge del coledoc o sense, la icterícia persisteix i els malalts moren de llur esclerosi hepàtica irreparable.* [Gallart (1923): 61, p. 18]. *Que la hipertensió pot produir l'esclerosi, no és una afirmació merament hipotètica, sinó un fet del qual tenim proves experimentals abundoses.* [Trias de Bes (1929): 61, p. 32].

**equip** *m.* **1a**. [**N**COMPT (*de* N1)] (N1[activitat]) Conjunt de persones organitzades per a la realització [d'una activitat]1. *~ de gestió, ~ de control, ~ de professors, ~ de dissenyadors, ~ d'investigadors // ~ psicopedagògic, ~ tècnic, ~ assistencial, ~ directiu, ~ sanitari, ~ redactor, ~ mèdic, ~ organitzador, ~ pastoral, ~ comercial // en ~. En la nostra època, es pot parlar d'un equip de recerca científica de llengua francesa, de llengua americana, de llengua alemanya o russa.* [Adrover (1967) [T]: 80, p. 15]i. *L'equip era integrat per seixanta-un guàrdies municipals [...].* [*Avui* (1979): 40, niv. 1, p. 54]i.

## 2.2. The BDLex

The second LR used as a helpful tool in the everyday activity is the DBLex database (Sanromà, 2004). It is a lexicographical corpus built during the period going from 1997-1999 and constituted by a repository of 13 Catalan dictionaries from the 19th and 20th centuries. The dictionaries contained in the BDLex chronologically coincide with the period covered by the corpus range. Moreover, it includes lexicographic publications not based upon previous lexical works and all different registers spoken in the Catalan language are represented. Most of the "inspiration" the lexicographer needs in the project framework in order to write the definitions included in the DDLC comes not only from the IEC normative dictionary, but also from the BDLex database.

## 2.3. The Corpus textual informatitzat de la llengua catalana (CTILC)

The Corpus textual informatitzat de la llengua catalana (CTILC) was designed and built as the main source for the making up of the DDLC even though it can be used in further research activities or LR initiatives.

The CTILC contains 52 million words. Texts coverage goes from 1832 up to 1988. Textual typology includes literary and non literary publications (narration, theatre and poetry and opinion articles on the one side, and manuals, papers, legal texts, newspapers and catalogues on the other side). The CTILC is fully annotated. Each occurrence is lemmatised and morphologically tagged. The lemmas appear accompanied by all possible variant forms and the information can be retrieved, then, by lemma or by form and, obviously, also by author, title, typology or year of publication.

In the project framework, the Institute has also designed and implemented a corpus query interface which is the LR used by all team members for exploiting data. The interface offers the possibility to search a single lemma or form, but also to find the unit with its most frequent collocations in the simple query window. On the complex search, the options expand up to 10 positions on both sides of the core unit and the user can, in addition, query by lemma or by morphological category.

## 3. Corpus exploitation strategies

The daily activity of the authors of the paper has derived into a first attempt to formalize the diversity of corpus queries starting from a wide range of patterns and taking into account the differences observed among each semantic lexical category. This analysis will provide the lexicographer with a catalogue of recommendations for the case of nouns in an initial step and, hopefully, in a near future, with a list of some corpus exploitation strategies on the basis of formalized patterns that extract relevant semantic information. The data retained will be manually analysed in order to fulfil the definition. The definition is the structural item that makes it explicit the sense of the lexical unit and its syntactic and semantic constraints. It is expressed by means of natural language but formalised as much as possible, that is, following some grammatical, discursive and register guidelines. The natural language used in order to

define a lexical unit is then a kind of metalanguage and the type of information it expresses, in our case, will be retrieved from the corpus and formalised afterwards. The strategies on hyperonymy and meronymy will help to find this kind of information and they will become a formalised methodology suggested to the lexicographic team members for internal coherence project improvement.

### 3.1. The hyperonymy pattern

The corpus query system has been initially interrogated in the experiment to show that the traditional *is_a* relation recognition patterns broadly used and recommended by many authors in the literature (Sierra et al, 2006, among others) could be useful and applied in our project. Some relevant semantic results are obtained but, conversely, no data about precision and recall can be even mentioned because of the intrinsic characteristics of the corpus.

*El pardal és un ocell que arriba a ésser sociable.*
(*a sparrow is a bird that becomes very sociable*)

However, when the lexical unit to be defined is very frequent, the use of this single *is_a* pattern is not qualitative good enough and, for this reason, the authors have been working on some different verbal patterns that help to find the most suitable examples for the integrated descriptors (that is, one single word) is concerned in the case of the hyperonymy relation. The patterns explored so on are the following[1]:

*hi ha una raó en els símptomes interns d'una malaltia --per exemple una grip: dolor d'ossos, mal de coll*
(*there is a reason for internal symptoms of disease- for example flu: bone pain, sore throat*)

**grip f**. Malaltia infecciosa aguda i contagiosa, d'origen víric, que provoca febre, mal de cap, símptomes d'afecció a les vies respiratòries, etc. (DIEC)
(**flu f.** Acute infectious and contagious disease, viral, that causes fever, headache, symptoms of disorder in the airways, etc.)

*Tota cultura pot ésser **entesa com a** sistema*
(*Every culture can be **understood as a** system*)

**cultura f.** 1. (...) 2 *1* f. [LC] [AN] [PE] Conjunt dels símbols, valors, normes, models d'organització, coneixements, objectes, etc., que constitueixen la tradició, el patrimoni, la forma de vida, d'una societat o d'un poble. (DIEC)
(**culture f.** 1. (...) 2. Set of symbols, values, norms, models of organization, knowledge, objects, etc.)

*la persona agent, **és a dir** la que realitza l'acció del verb*
(*the agent, **that is to say** wich perfoms the action of the verb*)

**agent m**. 6. En gram., persona que fa l'acció expressada pel verb. (DIEC)
**(agent m.** 6 In grammar, person doing the action expressed by the verb)

And a few more lexical patterns mentioned from the worst obtained results to the best ones, such as: *això és* (that is) and *concebut com* (conceived as) that arise as non productive, and, *definit com* (*la llengua és definida com un sistema estructurat de sons vocals arbitraris* [language defined as a structured and arbitrary sound system], *com ara* (*grups de tumors com ara el melanoma* [tumoral groups such as melanoma]. These patterns have allowed the authors to suggest, always on the basis of data extracted from corpus, what would be the most suitable descriptor for the definition the lexicographer is selecting and to compare the linguistic information retrieved from corpus to the information extracted from the BDLex and the normative dictionary consulted in the everyday activity.

[1] All examples shown are extracted from the CTILC. An approximate literal translation is provided for each example and the dictionary definitions in order to help the reader to understand the Catalan quotes.

### 3.2. The part-whole detection patterns

Not only for the case of the hyperonymy relation may the syntactic patterns help the lexicographer. Conversely, the case for nouns expressing holonymy and meronymy arise more complicated given the fact that in the DDLC working methodology this relation is expressed through two different types of definition according to its nature.

As stated above, the first type of definition includes all those cases in which a single whole is identified with one single part (i. e.: the *cerebellum* can only be a part of the brain). In that case, the descriptor is marked in a dissociated way indicating that there is a part-whole relation between the defined noun and the members integrating the collective lexical unit defined (i.e.: **choir** <dsint>group</dsint> of <dsem>people</dsem> who sing together) (Ilson, 1987). On the second group the authors find those words that maintain a part-whole relation with a few words (i. e.: a *handle* can be for taking a basket, a cup, a suitcase...). The definition of these nouns gathers all their variable elements in one extrinsic item that accompanies an integrated descriptor. In the case of *handle*, the DDLC defines it as follows: part [of an object]$_1$ used for holding it.

There is no doubt that this mark-up proposal is useful in order to recognize and to retain similar types of nouns expressing this kind of semantic relation, for example, collective nouns. However, the analysis in depth found on most relevant literature about part-whole relations (Winston et al., 1987, mainly) is not neither considered in most important dictionaries nor always reflected in the lexicographic definitions. The DIEC dictionary expresses the member-collection relation by means of the following item: *conjunt de* (*group of*):

**generació f.** Conjunt de les persones que viuen en una mateixa època, dins un mateix període de temps. *La generació actual*. (DIEC)
**(generation f.** Group of people living in the same era, in the same period of time. *The current generation.*)

*Aquestes coses [...] són molt apreciades per nosaltres, que **formem part** de la generació jove catalana.*
(*these things ... are very appreciated by us, who **are part of** the young catalan generation*)

**junta f**. Conjunt de persones nomenades per a dirigir els assumptes d'una col·lectivitat, executar certs serveis, etc. *La junta directiva d'una associació. La junta d'obsequis d'una societat recreativa.* (DIEC)
(**board f.** 2. Group of people appointed to manage the affairs of a community, perform certain services, etc. *The association board*.)

*la Junta Administrativa estava **composta per** quatre membres de l'Hospital de la Santa Creu*
(*the administrative board **comprise** four members of the Hospital de la Santa Creu*)

But it is not always like that, and some inconsistencies arise:

**colla** 1 ***1*** **f.** Nombre de persones aplegades deliberadament per a un fi. *Una colla de carregadors. Una colla de segadors. Una colla de lladres. Anar a colles.* (DIEC)

(***gang*** **f.** Number of people united for a purpose. *A gang of porters. A gang of reapers.*)

*Durant l'agost i el setembre sortíem del poble a punta d'alba **formant part** de la colla de plegadors i plegadores.*
(*During August and September we went out of town at dawn **as part of** the gang of beams and folding*)

*Els fadrins i les fadrines **s'unixen** en colles [...] i se n'ixen al camp [...]*
(*The boys and girls **join** gangs and go to the fields*)

One of the main reasons of this inconsistency is the difficulty the lexicographer faces in order to retain this kind of relations when fixing the definition of one lexical unit. The dictionary making process from corpus will help when combining data together with some strategies for

knowledge discovery. Further research efforts have been devoted to find syntactic verbal patterns expressing the part-whole relations in order to provide a more refined dissociated descriptor for collective nouns. The result of this work is a list of lexical and syntactic expressions that reflect meronymy and hyponymy relations between the corpus elements.

From the syntactic point of view, these patterns can be classified in two major categories:

- Phrase-level patterns: where the pattern includes the part-whole concepts in the same noun phrase. For example, in phrases like *el nus de la corbata* (*the knot of the tie*) or *els dits de la mà* (*the fingers of the hand*), the part and the whole turn up connected by the preposition *de*.
- Sentence-level patterns: where the pattern includes the part-whole concepts in a sentence, joined by a verb (i.e.: *L'apartament consta de dues habitacions, cuina i lavabo. The apartment has two bedrooms, kitchen and bathroom*).

Next two sections are devoted to the analysis of the syntactic patterns used in order to retrieve CTILC contexts expressing the part-whole relation that will be handled by the lexicographer when defining a lexical entry of this kind, both for the case of holonyms and meronyms.

a) Holonyms

Literature on this topic (Feliu, 2004) coincides to highlight a series of verbs that can express the part-whole relation. In this section the sentence-level patterns have been used to locate contexts that express holonymy in the corpus. The selected verbs to do the experiment are: *integrar* (*to integrate*), *compondre* (*to compose*), *constituir* (*to constitute*), *formar* (*to form*) and *incloure* (*to include*).

Our first searches, based on these verbs, had a quite disappointing result. It must be taken into account that a big part of the bibliography of reference had based their research on a specialised corpus and the CTILC is a general one. Consequently, it is necessary to modify the elements of this list in order to increase their effectiveness and the results precision.

Thus, the authors have added a preposition that indicates possession to the verb and repeated the query. Even though it can be observed that, in the case of the Catalan language, the search is refined using a preposition behind the verb, the percentage of valid occurrences was not high enough to consider them a useful pattern. The second change, consisting on the use of the verbal passive form as a mandatory pattern of search, was much more productive and the percentage of valid results increased considerably. See the following examples:

INTEGRATED BY (integrated by):
*[...] el **quartet**, integrat per A. Guinjoan, J. Ferrer, F. Vilanova i J. Guerin, interpretarà obres de Bach, Mozart i Paganini.*
(*[...] the quartet, integrated by A. Guinjoan, J. Ferrer, F. Vilanova and J. Guerin, perform works by Bach, Mozart and Paganini.*)

COMPOST PER/DE (composed by/of):
*Un jurat **compost** per professionals de l' art i específicament del disseny [...]*
(*A jury of professionals from the art and desing specifically [...]*)

*En aquesta edició el **programa** s' ha compost de conferències, debats i cinema.*
(*This year s program 'is composed of lectures, discussions and films.*)

FORMAT PER/DE (made by/from):
*[...] un petit **feix** de llenya, en general format de mates i branques procedents del sotabosc [...]*
(*[...] A small bundle of firewood, usually made from the branches of bushes and undergrowth [...]*)

*Aquest **llibre** està format per tretze gravats amb una explicació de cada un d' ells per als nens.*
(*This book is made by thirteen pictures with an explanation of each of them for children.*)

INCLÒS A (included in):
*[...] aquest dret [...] ha estat inclòs a les **Constitucions** de les democràcies occidentals.*
(*[...] This right has been included in the Constitutions of Western democracies.*)

Authors get positive but quite reduced number of results. Only a part of holonymy relations are retrieved with this search and most of the relations belong to collective nouns. Probably it is so due to the corpus characteristics because it has been demonstrated that the same kind of research on domain corpus gives most accurate results. In the future, however, the authors bear in mind to enlarge the number of patterns on this side.

b) Meronyms
In the case of the holonymyc definitions the authors have concentrated the research in the sentence-level patterns. Now, as for the case of meronymyc definitions, the analysis will be focused on the phrase-level patterns. Although it is true that the sentence-level patterns of the former section can also be useful for the location of meronymys, its performance in a general corpus is rather scarce in comparison with the results obtained with the phrase-level patterns.

In Catalan, the main phrase-level pattern to express the part-whole relation is N1 de N2 (N1 of N2). In order to check out if this pattern can help to select the appropriate contexts the authors will put it on trial in a pair of practical cases. Take as an example the words *tall* and *plat*, two nouns that fulfil the two conditions necessary for our analysis. First, they are nouns with a number of occurrences high enough to be considered. Obviously, it becomes a too high time-consuming task and the lexicographer cannot check them all: *tall* turns up 2148 times and *plat* 3632. Second, they are polysemic words that have meronymyc and not meronymyc meanings and, therefore, they will be useful cases for checking out a number of interesting contexts and to validate the patterns in order to select the maximum semantic possibilities. The result of filtering the occurrences of plat / tall + de is the following one:

| | | |
|---|---|---|
| TALL (EDGE / CUT)<br>1-Vora afilada de la fulla [d'una arma, d'una eina].<br>1-*Sharp edge of the blade [of a weapon, of a tool]*<br>2-Tros [d'aliment] tallat d'una part més gran.<br>2-*Piece [of food] cut from a larger part.* | TALL + de<br>971 ocurrences<br>256 useful<br>26% | 1- [...] li haurà ferit l'ànima com el tall d'un cristall.<br>1-[...] *he has hurt her soul as a crystal blade*.<br>2- […] una llesca de pa amb un tall de formatge.<br>2-[...] *a slice of bread with cheese*. |
| PLAT (PLATE)<br>1-Porció d'aliment que cap en un plat.<br>1-*Portion of food that fits in a plate*.<br>2-Peça plana i hortizontal, destinada a contenir alguna cosa, que forma part d'un objecte, d'un instrument.<br>2-*Flat, horizontal piece, intended to hold something, which is part of an object, an instrument*. | PLAT + de<br>836 ocurrences<br>573 useful<br>68% | 1- En Joan empunyà la forquilla per atacar el seu plat d'arròs amb conill.<br>1-*John took up his fork to eat a plate of rice with meat*.<br>2- [...] pesa els dos trossos de ferro posant-ne un a cada plat de la balança.<br>2-[...] *weigh the two pieces of iron on a plate in the balance*. |

So, even though results provide examples of the searched senses, the number of selected occurrences is too high to help the lexicographer. The problem is that this structure can express other relations besides meronymy. For example, N1 of N2 can also express possession (*La casa d'en Lluís*, Lluís' home), origin (*Formatge d'Holanda*, Cheese of Holland), and many other semantic links. Resulting lists are still too heavy to validate and with a high percentage of noise as a result. Consequently, alternative patterns to improve these results must be found.

Two different strategies have been applied in order to make emerge more effective patterns. The first one consists on start from zero without taking into account the structure formerly analyzed. The second one, on the contrary, takes as a basis the pattern N1 de N2 and attempts to find restrictions that, applied to it, may help to refine the contexts initially retrieved.

In order to discover new productive structures, lists of nouns co-occurrences with meronymyc meanings have been analyzed and these turned up very often preceded by a quantitative determiner. In our corpus, most elements belonging to this category are however lemmatized as adjectives and only the numerals have their own morphological category. So, the occurrences from the pattern NUM+nom have been filtered and the results are shown in the following table:

| | |
|---|---|
| NUM + Tall<br>82 ocurrences<br>41 useful<br>50% | 1-[...] una feixuga espasa de dos talls.<br>1-*[...] a heavy two-edged sword*.<br><br>2-Fregeix en una paella tres talls de bacallà.<br>2-*Three cuts of cod fried in a pan.* |
| NUM+Plat<br>132 ocurrences<br>97 useful<br>73% | 1-Et donaven dos plats d'escudella, un plat de carn d'olla, pa i vi per mitja pesseta.<br>1-*I gave two bowls of soup, a dish of meat, bread and wine by half a penny*.<br>2-Els gasos atmosfèrics es poden considerar com si pesessin dins els dos platets d'una balança?<br>2-*Atmospheric gases can be considered as if the plates are weighed in the balance plates?* |

As it can be observed, the authors continue to find examples of all the searched senses but the number of occurrences to revise has dropped considerably and the percentage of valid cases has risen. Therefore, NUM+nom is a valid query pattern for the meronymy detection in the corpus. It must be mentioned, however, that this pattern has an inconvenience: it is only applicable to the countable nouns. For this reason, it is necessary to find another valid pattern for other types of nouns. Following N1+de and attempting to find the way to restrict the results the authors have observed that the simplest way to get it was adding an article in the first position on the left in the query expression: Art + N1 +de. In this way many of the occurrences that do not make reference to parts of a whole have been successfully eliminated. The result of this new filtration is the following one:

| | |
|---|---|
| Article+TALL+de<br>244 ocurrences<br>160 useful<br>65% | 1-[...] has esmussat el tall de la seva espasa [...].<br>1-*You broke the blade of my sword* [...].<br>2-Es rosteixen els talls de bou [...].<br>2-*He fried beef cuts* [...]. |
| ARTICLE+PLAT+de<br>498 ocurrences<br>368 useful<br>73% | 1- [...] li serví el plat de lluç [...].<br>1-*He served a dish of hake* [...].<br>2-[...] el plat dels canelobres [...].<br>2-*Candelabrum dishes.* |

Once again, data shows the same number of located senses, less number of occurrences and a superior percentage of selected meronymys and with the advantage that the pattern Art+N+de is valid for all type of nouns.

## 4. Sense disambiguation when the descriptor is missing

In the descriptive lexicographic work, when the lexicographer is charged to define a word with a high ratio of appearance in corpus (for example 9800 occurrences such as the word *planta* meaning 'plant' and 'floor'), it is pretty obvious that it is not possible to read all the contexts in a quite reasonably amount of time. If the corpus is randomly interrogated and results linearly read, some relevant information can be lost on the process. The authors of the paper firmly propose use the already described patterns for the detection of the descriptor. As for *planta* is concerned, when searching the patterns oriented to find a suitable descriptor for the most frequent sense, a sample of the results is:

*I si la planta **és un** ésser i ens o al contrari.*
(*If the plant **is a** being o it isn't.*)

**planta f.** 1**Ésser** vivent que pertany al gran grup dels vegetals, típicament immòbil, de creixement indefinit, autotròfic i mancat de sistemes de relació. (...) 6 **Perímetre** ocupat per un edifici (DIEC).
(**plant f**. 1 Leaving **being** that belongs to the large group of plants, typically inmobil, indefinite growth, autotrophic and relationship systems lacking. (...) 6 **Perimeter** occupied by a building.)

Unfortunately, many other times some of the senses included in the dictionary by means of different descriptors are not explicitly found in corpus. In this case, after having applied all possible strategies to the hyperonymy detection in the corpus, it has not been possible to find examples illustrating the last sense (6) given by the dictionary. Some corpus query attempts have been made with the pattern: *X de/de X (X of/of X)*, because as already known words in context select different disambiguation strategies in the case of polysemic units. This complementary research concentrated on noun combination will help to disambiguate senses as the one you may analyse in the following example where, for a Catalan speaker, it is obviously that the use of *plant* must be equal to *floor*: *La planta d'aquest edifici és rectangular* (*The building floor is rectangular*).

### 4.1. The search of the candidates to extensional elements of the definition

The last group of patterns studied for the enrichment of the definition process is the list of verbs helping the lexicographer to find the complementary information of a definition, sometimes coinciding with the extrinsic information, expressing the constitutive parts, function, finality, and some other characteristics depending on the kind of noun defined. The core verbs analysed in this part of the experiment and a sample of the results obtained is: *tenir, compondre, constituir, fer, formar* o *estar (to have, to compose, to constitute, to do, to form, to be*). *Les **plantes** tenen arrels, tronc i fulles* (***Plants** have roots, trunk and leaves*) / *Aquestes **plantes** tenen de fet sexes separats* (*These **plants** have separate sexes*) / *Neix una **planta** que fa fruits* (*It comes a **plant** that makes fruits*) / *La majoria de **plantes** fan granes* (*Most **plants** make seeds*) / *Una **planteta** constituïda per arrels, tany i fulles.* (*A small **plant** formed by root, stem and leaves*) / *L'aigua baixa del cel i dóna vida a les **plantes**, que constitueixen l'aliment de l'animal.* (*Water falls from heaven and gives life to **plants**, wich are the animal food*).

Finally, the research will concentrate on the data exploitation in order to retrieve relevant information and the methodology followed to prove that some formalization and the use of lexical well-established patterns for, at least, the noun category will help the lexicographer in the definition writing task, especially when selecting the descriptor and the extension

information retained in the definition. Besides, the verbs presented above constitute a helpful list to determine the extrinsic noun arguments. From the observation of the articles database (nouns) already written up in DDLC, it has been possible to detect that the noun arguments usually appear introduced by other linguistic elements that will be highlighted: the relative *on* (where), the possessive adjective, the conjunction *perquè* (because), the adjectives *resultant* (resulting), *destinat* (directed to), *propi* (characteristic) and the syntactic structures *de* + Noun, *de* + Verb and *per (a)* + Verb. This syntactic information will be also taken into account when querying the corpus following the new lexicographic strategies proposed.

### 4.2. Patterns application *versus* Dictionary information: is the experiment really working?

In the last part of this paper a small experiment will be developed in order to check out whether it is possible to obtain subcorpus that allows to illustrate all senses and syntactic patterns of a particular noun combining the different strategies to search the descriptors and external arguments of the nouns.

Recapitulating, the principal patterns proposed for the selection of descriptors would be the following ones: Noun +{*integrar/compondre/constituir/incloure/és_un/per exemple/entès com/és a dir/això és/definit com*}. For the selection of the extrinsic arguments the combinations proposed are: Noun + {*destinat a/on/perquè/de* +Noun/*de* +Verb/*per (a)* Verb/possessive/*resultant/mitjançant/propi de/tenir/estar/fer*}

The experiment has consisted in selecting two very frequent nouns, *quadre* and *banc*, that constitute two complete articles already written up in the DDLC, and to create a subcorpus based on the searches proposed in order to contrast the results obtained with a subcorpus with the same number of occurrences retrieved from a random selection and, later, to see which differences arise between the results of both corpora.

*Banc* presents a total of 3737 occurrences in the CTILC. 684 occurrences form the subcorpus obtained from the proposed searches which represents a 18.3% of the total number of occurrences. *Quadre* presents a total of 3852 occurrences in the CTILC. 478 occurrences form the subcorpus obtained from the proposed searches which represents a 12.4% of the total number of occurrences. Next a schema of the lexical entry with the definition (and an approximate English translation) and two tables with the real data that are documented in DDLC for these words and the results obtained with the different subcorpus are shown:

| **BANC. DDLC senses and patterns** | **Random subcorpus** | **Guided subcorpus** |
|---|---|---|
| **1a**. [**N**COMPT] Seient llarg i estret on caben diverses persones. [Long and narrow seat where several persons fit]. | Sense: yes Patterns: yes | Sense: yes Patterns: yes |
| **1b**. [**N**COMPT] Tauló que reforça el buc d'una embarcació i que serveix de seient. [Plank that reinforces the hull of a vessel and that serves as seat]. | Sense: yes Patterns: yes | Sense: yes Patterns: yes |
| **1c**. [**N**COMPT] Banc**1a** petit que serveix per a facilitar l'accés a un lloc més alt. [Small bench that is for facilitating the access to a higher place] | Sense: yes Patterns: yes | Sense: yes Patterns: yes |
| **•1d**. [**N**COMPT (*de* $N_1$)] ($N_1$[humà]) Banqueta**1c**. [Sitting placed out of the playing area from where [the coach, the players, the technicians of a team]1 see the match] | Sense: no Patterns: no | Sense: yes Patterns: yes |
| **1e**. [**N**COMPT] Caixabanc**1**. [Box thought to serve as bench, with the usable lid as a seat] | Sense: no Patterns: no | Sense: no Patterns: no |
| **2a**. [**N**COMPT (*de* $N_1$)] ($N_1$[tipus]) Societat que es dedica a les múltiples operacions comercials produïdes pels diners, considerats com a mercaderia. [Company that dedicates itself to the multiple | Sense: yes Patterns: yes | Sense: yes Patterns: yes |

| | | |
|---|---|---|
| commercial operations produced by the money, considered as merchandise]. | | |
| **2b**. [**N**COMPT] Local on es desenvolupa l'activitat comercial d'un banc[2a]. [Premises where the commercial activity of a bench is developed]. | Sense: yes Patterns: yes | Sense: yes Patterns: yes |
| **•2c**. [**N**COMPT] Banca**1c**. . [banking] | Sense: yes Patterns: yes | Sense: yes Patterns: yes |
| **3a**. [**N**COMPT (*de* $N_1$)] ($N_1$[sorra, roca, material organogen]) Formació geològica subaquàtica consistent en una acumulació [de sorra, roca, materials organògens]1 [Subaquatic geological formation consistent in an accumulation [of sand, rocks]1]. | Sense: yes Patterns: yes | Sense: yes Patterns: yes |
| **3b**. [**N**COMPT (*de* $N_1$)] ($N_1$[roca sedimentària, carbó]) Formació geològica consistent en una acumulació [de roca sedimentària, de sediments orgànics, de carbó]1 que forma un estrat. [Geological formation consistent in an accumulation [of sedimentary rock, of organic sediments, of coal]1 that forms a stratum]. | Sense: yes Patterns: [NCOMPT de N1] | Sense: yes Patterns: yes |
| **3c**. [**N**COMPT] Altiplà marí aïllat de la plataforma continental vorejat per una depressió o gorja profunda que ocupa la part mitjana o externa de la plataforma. [Sea high plateau isolated from the continental platform] | Sense: no Patterns: no | Sense: no Patterns: no |
| **4a**. [**N**COMPT (*de* $N_1$)] ($N_1$[núvols]) Conjunt [de núvols del mateix gènere]1 de poc gruix que s'estenen aproximadament a un mateix nivell. [Collection [of clouds]]. | Sense: yes Patterns: [NCOMPT de N1] | Sense: yes Patterns: yes |
| **4b**. [**N**COMPT *de* $N_1$] ($N_1$[peixos]) Gran quantitat [de peixos]1 que es traslladen junts. [Great amount [of fish] that they move together] | Sense: yes Patterns: yes | Sense: yes Patterns: yes |
| **•4c**. [**N**COMPT *de* $N_1$] ($N_1$[mol·lusc, coral]) Lloc on viu una gran quantitat [de mol·luscs, de corals]1. [Place where a great amount lives [of molluscs, of choirs]]. | Sense: yes Patterns: yes | Sense: yes Patterns: yes |
| **•4d**. [**N**COMPT *de* $N_1$] ($N_1$=abelles) Conjunt [d'abelles]1 pertanyents a una sèrie de ruscs col·locats ordenadament per a l'explotació apícola. [Group [of bees] belonging to a series of beehives placed orderly for the apiarian exploitation]. | Sense: no Patterns: no | Sense: yes Patterns: yes |
| **5a**. [**N**COMPT (*de* $N_1$)] ($N_1$[artesà]) Taula de treball de fusta, resistent i pesant, [dels artesans]1. [Desk of wood, resistant and heavy, [of the artisans]] | Sense: yes Patterns: yes | Sense: yes Patterns: yes |
| **5b**. [**N**COMPT (*de* $N_1$)] ($N_1$[recipient]) Tauló amb forats [per a posar-hi recipients]1. [Plank with holes [to put recipients there].] | Sense: yes Patterns: [NCOMPT de N1] | Sense: yes Patterns: [NCOMPT de N1] |
| **5c**. [**N**COMPT] Banc**1a** que forma part del llit damunt el qual es posen les posts per a aguantar la màrfega i l'altra roba. [Part of the bed on which the boards are put to endure the pallet and the other clothes]. | Sense: yes Patterns: yes | Sense: yes Patterns: yes |
| **6**. [**N**COMPT (*de* $N_1$)] ($N_1$[matèria orgànica]) Lloc on hi ha emmagatzemades [matèries orgàniques]1 per a un ús futur. [Place where there is stored [organic matters] for a future use]. | Sense: no Patterns: no | Sense: yes Patterns: [NCOMPT de N1] |
| **•7**. [**N**COMPT] Cos inferior d'un retaule gòtic. [Inferior part of a Gothic altarpiece] | Sense: yes Patterns: yes | Sense: no Patterns: no |

| **Banc** / DCC article | Total senses: 20 | Total syntactic patterns: 28 |
|---|---|---|
| Guided subcorpus (684 occurrences) | Documented Senses: 17 (85%) | Documented patterns: 23 (82.1%) |
| Random subcorpus (684 occurrences) | Documented Senses: 14 (70%) | Documented Patterns: 17 (60.7%) |

| QUADRE / DDLC senses and patterns | Random subcorpus | Guided subcorpus |
|---|---|---|
| **1a**. [**N**COMPT] Pintura executada sobre un suport mòbil. [Painting executed about a mobile support] | Sense: yes Patterns: yes | Sense: yes Patterns: yes |
| **1b**. [**N**COMPT] Il·lustració emmarcada, destinada a ésser penjada a la paret. [Framed Illustration, destined to being hanged on the wall]. | Sense: yes Patterns: yes | Sense: yes Patterns: yes |
| **2a**. [**N**COMPT (*de* $N_1$)] ($N_1$[tema]) Conjunt de dades [sobre un tema]1 organitzades de manera que puguin ésser copsades d'un cop d'ull. [Data set [about a subject]1 organized so that they can be grasped of a look] | Sense: yes Patterns: yes | Sense: yes Patterns: yes |
| •**2b**. [**N**COMPT (*de* $N_1$)] ($N_1$[ambient, situació]) Descripció que evoca [un ambient, una situació]1. [Description that evokes [an environment, a situation]]. | Sense: yes Patterns: yes | Sense: yes Patterns: yes |
| •**2c**. [**N**COMPT] Text literari curt de contingut realista. [Short literary text of realistic contents]. | Sense: yes Patterns: yes | Sense: yes Patterns: yes |
| **3**. [**N**COMPT (*de* $N_1$)] ($N_1$[malaltia]) Conjunt de símptomes i signes [d'una malaltia]1. [Collection of symptoms and signs [of an illness]] | Sense: yes Patterns: yes | Sense: yes Patterns: yes |
| **4**. [**N**COMPT] Figura que té forma de quadrilàter[1]. [Figure that has a quadrilateral form] | Sense: yes Patterns: yes | Sense: yes Patterns: yes |
| **5a**. [**N**COMPT *de* $N_1$] ($N_1$[empresa, institució, exèrcit]) Conjunt dels comandaments [d'una empresa, d'una institució, d'un exèrcit]1. [Group of the commands [of a company, of an institution, of an army]] | Sense: yes Patterns: yes | Sense: yes Patterns: yes |
| •**5b**. [**N**COMPT *de* $N_1$] ($N_1$[professional]) Equip [de professionals]1 que desenvolupen una tasca concreta. [Team [of professionals] that develop a concrete task]. | Sense: yes Patterns: yes | Sense: yes Patterns: yes |
| •**6**. [**N**COMPT *de* $N_1$] ($N_1$[impressió]) Espectacle susceptible de provocar [determinades impressions]1. [Show susceptible of provoking [determinate impressions]] | Sense: yes Patterns: yes | Sense: yes Patterns: yes |
| •**7**. [**N**INCOMPT (*de* $N_1$)] ($N_1$[cultura, activitat, centre, període de temps]) Àmbit**1a**. [Domain] | Sense: yes Patterns: yes | Sense: yes Patterns: yes |
| **8a**. [**N**COMPT] Agrupament de personatges que romanen quiets a escena en una mateixa actitud. [Grouping of characters who remain still in scene in the same attitude] | Sense: no Patterns: no | Sense: no Patterns: no |
| **8b**. [**N**COMPT] Subdivisió d'un acte corresponent a un canvi de decoració. [Subdivision of an act corresponding to a change of decoration] | Sense: yes Patterns: yes | Sense: yes Patterns: yes |
| •**8c**. [**N**COMPT] Companyia**3**. [Company] | Sense: yes Patterns: yes | Sense: yes Patterns: yes |
| **9a**. [**N**COMPT] Conjunt d'indicadors disposats a la vista del pilot d'un vehicle de motor per tal de facilitar-ne el control. [Set of indicators disposed in view of the pilot of a motor vehicle in order to facilitate the control]. | Sense: no Patterns: no | Sense: yes Patterns: yes |
| **9b**. [**N**COMPT] Suport que incorpora els elements necessaris per al control d'un circuit elèctric. [Support that incorporates the elements necessary for the control of an electrical circuit]. | Sense: yes Patterns: yes | Sense: yes Patterns: yes |
| **10**. [**N**COMPT] Formació militar en forma de quadrilàter amb soldats encarats cap a les quatre direccions. [Military formation in quadrilatera form with soldiers pointed towards the four directions]. | Sense: yes Patterns: yes | Sense: yes Patterns: yes |
| **11**. [**N**COMPT *de* $N_1$] ($N_1$[vehicle]) Bastiment tubular [d'una bicicleta, d'una motocicleta]1. [Tubular frame [of a bicycle, of a motorcycle]] | Sense: no Patterns: no | Sense: yes Patterns: yes |
| •**12**. [**N**COMPT] Ornament de fil que es cus a les mitges i que va des del turmell fins al tou de la cama. [Ornament of thread that is sewn on the stockings] | Sense: yes Patterns: yes | Sense: yes Patterns: yes |

| **Quadre** / DCC article | Total senses:19 | Total syntactic patterns: 23 |
|---|---|---|
| Guided subcorpus (478 occurrences) | Documented Senses: 18 (94.7%) | Documented patterns: 22 (95.6%) |
| Random subcorpus (478 occurrences) | Documented Senses: 16 (84.2%) | Documented patterns: 20 (86.9%) |

We observe, then, that in spite of not obtaining a perfect result the guided corpus shows some more relevant data than a random corpus both concerning the number of documented senses and the number of syntactic patterns retrieved. It is also convenient to add that the not documented senses are in all the marginal sense cases, the less frequent, and that almost all the senses not documented in the guided corpora appear in the BDLex (as a matter of fact all the senses with the exception of the last one of *banc*). The lexicographer might specifically search this semantic information in order to check out if these senses effectively turn up or they do not appear in the corpus.

## 5. Final words and future research lines

The authors have initiated a research line oriented to show that the writing of articles from a subcorpus delimited by patterns of search based on the detection of descriptors and external arguments can speed up the process of writing a dictionary like the DDLC. It has been demonstrated that these strategies are valid for considerably reducing the number of occurrences that the lexicographer should analyse in order to write up an article.

Main contribution of the experiment is the resulting corpus exploitation strategies catalogue for the definition task, summarised as follows:

| **Hyperonymy** | **Part-whole** | **Extension** |
|---|---|---|
| Noun + és un/ per exemple / entès com / és a dir / definit com | NOM + integrat per / integrat de / compost per / compost de / constituït de / inclòs a / inclòs en<br>Numeral + NOM<br>Article + NOM + de | NOM + >5 tenir / compondre / constituir / fer / formar / estar<br>NOM + 1 de +1 nom / verb<br>NOM + 1 on / perquè / resultant / destinat / propi<br>NOM + 1 per + >2 verb<br>NOM - 1 possessive |

This catalogue will be from now on taken into account in the lexicographic daily activity. Authors take for granted that the consequently time reduction will not only benefit the dictionary fulfilment in terms of production but also in terms of internal coherence among different team members.

Future research will be devoted to apply the patterns, as just mentioned, to the daily activity but, also, and which is even more appealing from the lexicographic research point of view, new patterns will be studied for some other morphological categories, mainly verbs and adjectives, in order to proceed in a similar way when defining a lexical unit in the ongoing descriptive dictionary built on the basis of corpus.

## References

Feliu, J. (2004). *Relacions conceptuals i terminologia*. Barcelona: Institut Universitari de Lingüística Aplicada.

Giboreau, A.; et al. (2007). 'Defining sensory descriptors: Towards writing guidelines based on terminology'. In *Food Quality and Preference*, 18, 265-274.

Girju, R.; Badulescu; A.; Moldovan, D. (2003). 'Learning Semantic Constraints for the Automatic Discovery of Part-Whole Relations'. In *Proceedings of the Human Language Technology Conference (HLT)*. Edmonton, Canada.

Ilson, R. (1987). 'Towards a Taxonomy of Dictionary Definitons'. In Ilson, R. (ed.) (1987) *A Spetrum of Lexicography*. Amsterdam: John Benjamins.

Krishnamurthy, R. (2008). 'Corpus-Driven Lexicography'. In *International Journal of Lexicography*, 21(3), 231-242.

Rafel, J. (2006). 'Elements extrínsecs en les definicions lexicogràfiques: teoria i aplicació'. Bernal, E.; DeCesaris, J. (eds.) *Palabra por palabra. Estudios ofrecidos a Paz Battaner*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, 201-217.

Rafel, J.; Soler, J. (2006). 'A Descriptive Dictionary of Contemporary Catalan: the DDLC Project'. In *Proceedings XII EURALEX International Congress*, vol. I, 443-455.

Sanromà, R. (2004). 'Constitució i explotació d'una base de dades lexicogràfica (BDLex)'. In Battaner, P.; DeCesaris, J. (eds.) (2004). *De lexicografia. Actes del I Symposium Internacional de Lexicografia (Barcelona, 16-18 de maig de 2002)*, Barcelona: Universitat Pompeu Fabra, 741-754.

Sierra, G.; et al. (2006). 'Towards the Building of a Corpus of Defintional Contexts'. In *Proceedings XII EURALEX International Congress. Atti*. Vol. I, 229-240.

Soler, J. (2006) *Definició lexicogràfica i estructura del diccionari*. Barcelona: Institut d'Estudis Catalans.

Winston, M. E.; Chaffin, R.; Herrmann, D. (1987). 'A Taxonomy of Part-Whole Relations'. In *Cognitive Science*, 11, 417-444.