The Database of Catalan Adjectives

Roser Sanromà*, Gemma Boleda**
*IEC, Institut d'Estudis Catalans
Carme 47, 08001 Barcelona
**TALP Research Center - Technical University of Catalonia
Jordi Girona 1-3, 08034 Barcelona
E-mail: rsanroma@iec.cat, gboleda@lsi.upc.edu

Abstract

We present the Database of Catalan Adjectives (DCA), a database with 2,296 adjective lemmata enriched with morphological, syntactic and semantic information. This set of adjectives has been collected from a fragment of the Corpus Textual Informatitzat de la Llengua Catalana of the Institut d'Estudis Catalans and constitutes a representative sample of the adjective class in Catalan as a whole. The database includes both manually coded and automatically extracted information regarding the most prominent properties used in the literature regarding the semantics of adjectives, such as morphological origin, suffix (if any), predicativity, gradability, adjective position with respect to the head noun, adjective modifiers, or semantic class.

The DCA can be useful for NLP applications using adjectives (from POS-taggers to Opinion Mining applications) and for linguistic analysis regarding the morphological, syntactic, and semantic properties of adjectives. We now make it available to the research community under a Creative Commons *Attribution Share Alike 3.0 Spain* license.

1. Introduction

We present the Database of Catalan Adjectives (DCA), a database with 2,296 adjective lemmata enriched with morphological, syntactic and semantic information. These lemmata have at least 50 occurrences in one fragment of the Corpus Textual Informatitzat de la Llengua Catalana (CTILC; Rafel, 1994) of the Institut d'Estudis Catalans, a semi-automatically annotated and hand-corrected corpus of written genre. The fragment contains 14.5 million words from texts written between 1969 and 1988.

The database includes both manually coded and automatically extracted information regarding the most prominent properties described in the literature regarding the semantics of adjectives, such as morphological origin, suffix (if any), predicativity, gradability, adjective position with respect to the head noun, adjective modifiers, or semantic class (see Hamann 1991 and Raskin and Nirenburg 1991 for general overviews from different perspectives, and Picallo 2002 for Catalan adjectives). As we will see below, some of the information in the DCA has been automatically extracted from the CTILC fragment and some of it manually coded.

The database was initially built in Sanromà (2003). Boleda (2007) enriched it and used it for several experiments on the adquisition of semantic classes for adjectives. We now make it available to the research community under a Creative Commons *Attribution Share Alike 3.0 Spain* license. The DCA and its documentation can be downloaded from the ACL data and code repository.[1]

We next review the information that is included in the database.

2. Information

2.1 Derivational type

The DCA contains information about the derivational type of the adjectives. From a morphological point of view, and taking into account the lexical category of the morphological basis (in the case of the non primitive adjectives) and the morphological category (in the case of the participles), Catalan has six adjective derivational types: basic (nonderived), denominal, deverbal, participial, deadverbial, and deadjectival.

The DCA includes the derivational type of each adjective, excluding the deadverbial and deadjectival types, because they are very rare. Among morphologically derived adjectives, two types are further distinguished depending on the existence or not, from a synchronic point of view, of the morphological basis. Class I derivatives are words whose morphological basis is an existing word. In some cases the derivative and its basis are very close: *dental* 'dental' (<**dent** 'tooth' + suffix *al*). In other cases, the latinizing derivative and the present basis form are distant: *radical* 'radical' (<**arrel** 'root'). Class II derivatives are words with a derivative suffix but no relation with any Catalan morphological basis, as for example *agrari* 'agricultural' or *assequible* 'affordable'. The existence of other derivatives with the same suffix, as *hospitalari* 'hospitable' (<**hospital** + suffix *ble*) or *acceptable* (<**acceptar** 'accept'), and the existence, on occasion, of other words derived from the same radical (*agrícola* 'agricultural') allow us to consider them derivatives.

Thus, the information about derivational types considers four main types plus a subdivision within two of the types (I/II for denominal and deverbal). The distribution of the

[1] Accessible from http://aclweb.org/aclwiki.

adjectives among these derivational types is shown in Table 1.

Deriv. type	#Adj	%Adj	Examples
basic	519	22.6%	*petit* 'small' *indirecte* 'indirect' *agredolç* 'sweet and sour' *llunyà* 'far' (deadverbial) *blavós* 'bluish' (deadjectival)
denominal I	745	32.4%	*revolucionari* 'revolutionary' *esfèric* 'spherical' *mexicà* 'Mexican' *afortunat* 'lucky' *sociològic* 'sociological'
denominal II	116	5.1%	*acústic* 'acoustic' *agrari* 'agricultural' *fluvial* 'fluvial' *morbós* 'morbid' *fisicoquímic* 'physico-chemical'
deverbal I	323	14.1%	*alarmant* 'alarming' *volador* 'flying' *atractiu* 'attractive' *preexistent* 'pre-existing' *intercanviable* 'interchangeable'
deverbal II	78	3.4%	*docent* 'teaching' *despectiu* 'disdainful' *regressiu* 'regressive' *potable* 'drinkable' *assequible* 'affordable'
participle	515	22.4%	*aïllat* 'isolated' *deprimit* 'depressed' *conegut* 'known' *compromès* 'compromised' *prefabricat* 'prefabricated'

Table 1: Distribution of derivational types.

2.2 Syntactic groups and associated information[2]

We also code information on the syntactic behaviour of adjectives. In Catalan and other languages, adjectives have two main syntactic functions: They act as predicates in copular environments or as noun modifiers. The default function of the adjective in Catalan is that of modifying a noun; the default position is the post-nominal one (see examples in (1)), as in most Romance languages (Picallo, 2002).

(1) a. taula gran
table big
'big table'
b. arquitecte tècnic
architect technical

However, some (but not all) adjectives can appear pre-nominally, mainly when non-restrictively used (so-called "epithets"). For instance, (2a) is possible, but not (2b).

(2) a. gran taula
big table
'big table' (nonrestrictive interpretation)
b. *tècnic arquitecte
technical architect

Adjectives can also function as predicates, be it in a copular sentence (example (3a)) or in other predicative contexts, such as adjunct predicates (example (3b)).

(3) a. Aquest cotxe és molt maco
This car is very nice
'This car is very nice.'

b. La vaig veure borratxa
Her I-did see drunk
'I saw her drunk.'

We code the relative frequencies of the three syntactic functions: pre-nominal modifier (AN, for adjective-noun), post-nominal modifier (NA), and predicate (VA). We assign rank values to each of these positions according to the percentage of occurrences in each position in the corpus, with the six percentage divisions shown in Table 2:[3]

Rank	From	To
1	75%	100%
2	50%	75%
3	25%	50%
4	5%	25%
5	1%	5%
6	0%	1%

Table 2: Ranks for adjective grouping.

Based on these rank values, adjectives are grouped according to the range of syntactic positions they appear in and to their predominant position. Each group has a label formed by: first, the number of positions that the adjective occupies; secondly, the abbreviated form of the dominant position or positions (NA, VA or AN) accompanied by the rank value; last, in the case of adjectives which occur only in two positions, the abbreviated reference to the second position. For example, the group 3NA1 agglutinates the adjectives that occur in the three basic positions and whose highest rank corresponds to the postnominal position (NA). Similarly, the group 2VA1NA consists of the adjectives that occur only in the VA and NA positions (that is, adjectives that do not occur in pre-nominal position), and whose highest rank corresponds to the postverbal position.

[2] The information on syntactic groups has been extracted from a smaller fragment of the corpus (around 7 million words; texts from 1979 to 1988).

[3] This information was extracted with the CQP tool of the IMS Open Corpus WorkBench (http://cwb.sourceforge.net).

The result of applying these criteria is a new distribution of the adjectives into 21 groups. Their distribution, sorted according to frequency, is shown in Table 3.

Group	Example	#Adjs.	%Adjs.
3NA2	*important* 'important'	459	20,0%
2NA1VA	*general* 'general'	446	19,4%
3NA1	*anterior* 'previous'	408	17,8%
1NA1	*social* 'social'	379	16,5%
2NA2VA	*mort* 'dead'	98	4,3%
3VA2	*clar* 'clear'	95	4,1%
2NA1AN	*especial* 'special'	71	3,1%
2VA2NA	*situat* 'situated'	71	3,1%
3NAVA3	*estrany* 'strange'	62	2,7%
3AN2	*petit* 'small'	44	1,9%
3NAAN3	*alt* 'high'	39	1,7%
2VA1NA	*capaç* 'able'	38	1,7%
3NAVAAN3	*trist* 'sad'	20	0,9%
2NA2AN	*següent* 'following'	18	0,8%
3VA1	*impossible* 'impossible'	14	0,6%
3AN1	*gran* 'big'	12	0,5%
2AN2NA	*esmentat* 'mentioned'	11	0,5%
2AN1NA	*darrer* 'last'	5	0,2%
3VAAN3	*pitjor* 'worst'	3	0,1%
1AN1	*mal* 'bad'	2	0,1%
2VA1AN	*reclòs* 'secluded'	1	0,0%

Table 3: Adjective groups.

Both the rank of the adjective in each position and the syntactic group are included in the DCA. In addition, further fields in the database code more fine grained information (e.g., distinguishing occurrences with and without an intervening adverb).

2.3 Distributional information

The DCA contains information on other textual correlates (henceforth, *features*) of semantic and distributional properties of adjectives, such as gradability. As an example of how the information was extracted from the corpus, for gradability the presence of certain adverbs (*molt* 'very', *tan* 'so') or the suffix *-íssim* (as in *grandíssim* 'very large') were taken into account. In general, feature values correspond to proportions (number of ocurrences with a particular feature divided by total occurrences of the adjective in the corpus); the exceptions are described in the documentation of the resource.

The properties and the corresponding features are listed in Table 4. For more details on the feature definition, see Boleda (2007: Section 6.2.1.3). Note that some of the features overlap with the information described in the previous section (e.g., non-restrictivity is similar to the AN rank information). However, most of them provide complementary information.

Property	Feature description	#Features
Non-restrictivity	Pre-nominal modification of head noun	1
Predicativity	Predicative function (with copulas *ser*, *estar*, or both)	4
Gradability	Presence of gradability and comparison markers (adverbs, suffixes)	4
Syntactic function of head noun	Head noun is subject, object, or complement of a preposition	3
Distance to the head noun	Linear distance (number of words)	1
Binaryhood	Presence of a preposition to the right of the adjective (higher value for adjectives with two arguments)	1
Agreement properties	Gender, number	2

Table 4: Semantic features. The last columns show the number of features that correspond to each property.

2.4 Semantic class

Finally, the DCA includes the semantic class of 210 of the adjective lemmata as assigned by a committee of three experts. The 210 adjectives are a stratified sample of the database (balancing for frequency, derivational type, and suffix); within the sampling criteria, the lemmata were randomly chosen.

The semantic classification distinguishes between basic, event-related, and object-related adjectives, closely following the Ontological Semantic classification proposed in Raskin and Nirenburg (1998). These classes have the following properties:

Basic adjectives:[4] These are the prototypical adjectives, and denote attributes or properties that cannot be decomposed (*bonic* `beautiful', *sòlid* `solid'). They are predicative, gradable, and in Catalan they can appear both pre- and post-nominally.

Event-related adjectives: These do not denote typical properties, but rather a property based on a relationship to an event. For instance, the semantics of *tangible* ('tangible') includes some sort of pointer to an event of *touching*. These adjectives are typically deverbal, tend to bear arguments, hence they appear more often in predicative positions and do not often appear in pre-nominal position.

Object-related adjectives: Similarly, object-related adjectives denote a property based on a relationship to an object. For instance, *pulmonar* 'pulmonary' can be paraphrased as 'related to the lungs' and lungs are the external object the property is related to. These adjectives

[4] This term is not standard; some works in descriptive grammar use 'qualitative' for roughly the same class.

are typically denominal, not gradable, and can only act as predicates under very restricted circumstances. In Romance languages such as Catalan, French, or Spanish, they can only modify nouns post-nominally. Adjectives of this type exhibit a strong adjacency constraint, appearing immediately after the noun. In particular, if a noun is modified by more than one adjective, a relational adjective will come first.

Some adjectives have different senses, each of which belongs to a different class (see examples in Table 5 below). These types of regular polysemy are coded in the database through the assignment of complex classes (basic-event, basic-object, event-object). For further details on the classification and the criteria, see Boleda (2007) and Boleda et al. (2008). The distribution of the adjectives into classes, as well as some examples, is provided in Table 5.

Sem. class	#Adjs.	%Adjs.	Examples
basic	107	50.1	*ample* 'wide' *recent* 'recent' *silenciós* 'silent'
event	37	17.6	*imperceptible* 'imperceptible' *revelador* 'revealing' *vivent* 'living'
object	30	14.3	*barceloní* 'Barcelonian' *marxià* 'Marxian' *respiratori* 'respiratory'
basic-event	7	3.3	*cridaner* 'vociferous / loud-coloured' *embolicat* 'wrapped / embroiled' *sabut* 'known / wise'
basic-object	23	10.1	*amorós* 'affectionate / related to love' *familiar* 'familiar / related to family' *socialista* 'socialist(ic)'
event-object	6	2.3	*comptable* 'countable / related to counts' *digestiu* 'digestive / related to the digesting process' *docent* 'teaching / related to teachers or the teaching task'

Table 5: Distribution of semantic classes.

3. Conclusion

In this paper we have presented the Database of Catalan Adjectives, containing morphological, syntactic and semantic information obtained from corpus data and manual annotation. Because it contains a large set of adjectives (over two thousand) and a rich number of features (from derivational morphology to gradability), it is useful for quantitative approaches to linguistic research, as well as for NLP purposes. Some of its potential uses are the following:

1. exploiting the information on frequency distribution, both of adjectives and their properties;
2. studying word formation mechanisms within the adjectival category;
3. detecting new adjective meanings not deducible from the morphological components;
4. relating word formation processes and semantic classes through the syntactic behaviour of adjectives;
5. analyzing the role of polysemy in the semantics of adjectives;
6. enriching the description of adjectives in traditional dictionaries and in lexical databases used in Natural Language Processing systems.

4. Acknowledgements

This research has been partially funded by the Spanish government (projects TIN2006-1549-C03-02, TIN2009-14715-C04-04, HUM2007-60599/FILO), two PhD grants by the Generalitat de Catalunya and Fundación Caja Madrid and one post-doc contract by the Spanish Ministerio de Educación y Ciencia (JCI-2007-57-1479) to Gemma Boleda.

5. References

Bally, C. (1944). *Linguistique générale et linguistique française*. Berne: A. Francke.

Boleda, G. (2007). Automatic acquisition of semantic classes for adjectives. Ph.D. thesis, Pompeu Fabra University.

Boleda, G., S. Schulte im Walde, T. Badia (2008). An Analysis of Human Judgements on Semantic Classification of Catalan Adjectives. *Research on Language and Computation* 6(3): 247-271.

Hamann, C. (1991). Adjectivsemantik/Adjectival Semantics. In von Stechow, A. and Wunderlich, D. (Eds.), *Semantik/Semantics. Ein internationales Handbuch der Zeitgenössischen Forschung. An International Handbook of Contemporary Research*, pages 657-673. Berlin/New York: de Gruyter.

Levi, J. N. (1978). *The Syntax and semantics of complex nominals*. New York: Academic Press.

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

Picallo, C. (2002). L'adjectiu i el sintagma adjectival. In Solà, J. (Ed.), *Gramàtica del català contemporan*i, pages 1643-1688. Barcelona: Empúries.

Raskin, V. and Nirenburg, S. (1998). An applied ontological semantic microtheory of adjective meaning for natural language processing. *Machine Translation*, 13(2-3), 135-227.

Sanromà, R. (2003). Aspectes morfològics i sintàctics dels adjectius en català. Master's thesis, Universitat Pompeu Fabra.