A Hierarchy of Semantic Labels for Spanish Dictionaries

Xavier Blanco(✉)

Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Spain
Xavier.Blanco@uab.cat

Abstract. We present a hierarchy of semantic labels for the Spanish language. Each semantic label corresponds to the *genus proximum* (the next kind) of the lexical units that it describes. After having chosen ALGO 'something' as the head of our hierarchy, we distinguish between HECHO 'fact' and ENTIDAD 'entity', each of which are further partitioned into around twenty subclasses. Facts correspond to predicates while entities can correspond either to semantic names (objects) or to quasi-predicates. For disambiguation of predicates and quasi-predicates, we use the notion of an actantial formula (a linguistic expression that specifies the actants of a predicative form). The implementation of this hierarchy in the Spanish electronic dictionary of NooJ would allow us to foresee diverse applications.

Keywords: Semantic label · Actantial formula · Quasi-predicate · Spanish lexicography

1 Introduction

In the frame of the Labelsem project (FFI-2013-44185-P funded by the Spanish Ministerio de Economía y Competitividad), the research group fLexSem (Phonetics, Lexicology and Semantics, Autonomous University of Barcelona) is developing a comprehensive hierarchy of semantic labels for dictionaries of Spanish. A "semantic label" of a dictionary's entry is a lexical unit (simple or complex) or (more rarely) a syntagm that corresponds to the genus proximum (the next kind) of the entry in question. For example, the semantic label for *arañazo* ('scratch'), *corte* ('cut') and *herida* ('wound') in contexts such as *Mi hijo se hizo un arañazo en la cara* ('My son got a scratch on his face'), *¿Cómo evitar que mis padres vean los cortes que llevo en la muñeca?* ('How to hide cuts on my wrist from my parents?') or *Lamerse una herida es una respuesta instintiva en los humanos y en otros muchos animales* ('Wound licking is an instinctive response in humans and many other animals') would be LESIÓN ('injury', 'lesion') because the definitions of *arañazo*, *corte* and *lesión* would begin with *lesión que*… More precisely, the semantic label is a minimal paraphrase for a given lexical unit that corresponds to

This research has been funded by the *Ministerio de Economía y Competitividad* (Spain) in the frame of the project FFI-2013-44185-P *Jerarquía de etiquetas semánticas (español y francés) para los géneros próximos de la definición lexicográfica.*

T. Okrut et al. (Eds.): NooJ 2015, CCIS 607, pp. 66–73, 2016.
DOI: 10.1007/978-3-319-42471-2_6

the communicatively dominant component of its (Aristotelian) definition. Note that we will use capital letters for the labels.

The semantic label constitutes the syntactic head of the entry's formal definition. It must be noted that the *definiendum* almost never corresponds with the entry form (the lemma) but with a propositional form or actant structure that includes all the semantic actants of the lemma. For instance, we do not define *herida* ('wound') but *herida de X por parte de Y en Z con W* ('wound of W from Y in Z with W') (X and Y being animates, Z a part of X, and W a physical object). Other propositional forms accepted by *herida* would be the objects of different definitions.

Let us stress that these labels are actual linguistic signs of Spanish and not a metalinguistic device. This implies that the regular semantic, syntactic and restricted lexical co-occurrence of a given label with the *definiendum* can and must be controlled before attributing a lemma to it. This control plays a key role in the elaboration of the hierarchy as it is the central criterion for the attribution of a label. Moreover, it is a distinctive trait of our hierarchy of labels since most sets of semantic labels are made up of metalinguistic entities. Labelling in this way we obtain both a minimal paraphrase of the lemma's signified and a syntactical substitute in any context [2].

At present, our hierarchy comprises approximately 700 labels (only nominal ones). The total number of labels in the hierarchy cannot be fixed in advanced since we cannot arbitrarily restrict ourselves to particular sets of hyperonyms. The hierarchy is (mainly) inductively built, so we need the *genus* or next kind for each lema in the dictionary. Of course, the usual inheritance mechanism can be used to form quantitatively manipulable sets of lemmas.

The label with the greatest semantic extension is ALGO ('something'), followed by HECHO ('fact') and ENTIDAD ('entity'). An HECHO [1] is always a semantic predicate, while an ENTIDAD can be a semantic name or a predicate (quasi-predicate). In this paper, we will restrict ourselves to discussing the semantic labels of a sample containing 1,000 lemmas corresponding to facts and 1,000 lemmas corresponding to entities.

It is worth emphasising that the hierarchy of semantic labels is language-dependent. As a result, it cannot be directly used for translation or for multilingual search operations. However, different mechanisms of connections or equivalences between hierarchies can be proposed in order to consider translinguistic applications. That is relatively straightforward between Spanish and French, since we are building our hierarchy according to the methods and results obtained for the French language by the researchers of the ATILF Laboratory (*Analyse et Traitement Informatique de la Langue Française, Université de Lorraine*) and of the OLST (*Observatoire de Linguistique Sens-Texte, Université de Montréal*) [5, 9, 10]. Moreover, up until now we have kept the same number of classes for the first two levels of the hierarchy.

2 Facts *vs* Entities

First of all, we have to choose a label for the root of our hierarchy. This label has to be the most extensive meaning in Spanish. As indicated above, we choose the label ALGO ('something') since every Spanish lexical unit can accept (stylistic considerations aside)

ALGO as a *genus* (*un armario es algo que…* 'a cupboard is something that…', *un roble es algo que…* 'an oak is something that…'). Note that even for humans the replacement by *algo* is possible and even natural in some contexts (*Un abogado es algo distinto* 'A lawyer is something different', *Ese tío es algo especial* 'This guy is something special') even if, obviously, replacing ALGO by ALGUIEN ('someone') would dramatically improve acceptability when defining lexical units denoting humans.

Next, a basic distinction must be drawn between HECHO ('fact') and ENTIDAD ('entity') [1]. A fact 'takes place', whereas an entity 'is'. A lexical unit of Spanish denoting a fact can be embedded under *Sé que…* or *Creo que…* (*Sé que Juan tiene cáncer* 'I know that John has cancer', *Sé que Juan dio un paseo* 'I know that John went for a walk', *Sé que Juan dio una charla* 'I know that John gave a talk'…). *Cáncer*, *paseo* and *charla*, in these contexts, are facts.

Facts are inscribed in time and, consequently, they combine with grammatical meanings as 'present', 'past', 'future', 'simultaneous', 'previous', etc. They also combine with aspectual meanings like 'semelfactive', 'iterative', 'distributive', 'punctual', 'durative', 'habitual', 'perfective', 'progressive', etc. Moreover, they accept other grammatical meanings such as 'intensive', 'negative', 'causative', 'inchoative', 'interrogative', etc.

By contrast, entities are inscribed in space and they present dimensional values (length, weight…). The grammatical meanings applied to entities are different to those applied to facts. Entities, for instance, accept size ('augmentative', 'diminutive'), and sometimes (if they are living beings) sex ('masculine', 'feminine'), etc. An important point to be taken into account is that an entity always corresponds to a noun, whereas the opposite is not true. A noun can certainly correspond to a fact (see above the examples of *cáncer*, *paseo*, *charla*), and adjectives and verbs are often facts (never entities), while adverbs are facts that select only other facts in their actant structure (see Sect 3).

2.1 Subclasses of Facts

At the moment, our hierarchy comprises sixteen subclasses of facts, namely (in alphabetical order): ACCIÓN ('action'), ACONTECIMIENTO ('event'), ACTITUD ('attitude'), ACTIVIDAD ('activity'), CANTIDAD ('quantity'), CARACTERÍSTICA ('feature'), COMPORTAMIENTO ('behaviour'), CONJUNTO DE HECHOS ('set of facts'), COSTUMBRE ('habit'), ESTADO ('state'), FENÓMENO ('phenomenon'), PERÍODO ('period'), PARÁMETRO ('parameter'), RELACIÓN FACTUAL ('factual relationship'), PROCESO ('process') and SITUACIÓN ('condition'). The English translations of these labels are not to be taken as labels themselves. They are approximative and are only given for the convenience of the reader.

To a certain extent, it is possible to resort to linguistically motivated criteria in order to distinguish these subclasses. So, for instance, lexical units labelled as GOLPE 'blow' can be viewed as facts with a puntual character that occur at a given moment, and do not present an internal temporal structure. States would be atelic, since they do not have an inherent limit. Therefore, sentences such as **Está sabiendo la respuesta*, **He is knowing the answer* are ungrammatical. Actions are performed by an agent and thence they can be volitional: *Juan me dio un golpe (a propósito, sin querer)*, 'John hit me (on

purpose, unintentionally)' but **Juan sabe la respuesta a propósito,* *'John knows the answer on purpose'.

However, even if these criteria can prove useful in many circumstances, specific contextual effects often blur the applicability of the linguistic tests on which they are based. Not only metaphorical or idiomatic uses, but also technical ones can indeed modify the acceptability of a sentence. Therefore, the main criterion for the attribution of a semantic label to a given lexical unit remains the possibility or impossibility of using the label in a question as *genus proximum* in the lexicographical definition. Another possible formalisation of this lexical relation is the lexical function **Gener** (generic concept) [5]. For instance, the **Gener** value of *incremento* 'increase' is PROCESO 'process'. Note that **Gener** is not the same as hyperonymy, because hyperonymy is a semantic relation while **Gener** is a lexical one. The former will resist translation to another language, whereas the latter may not. The value of **Gener** for a given keyword must accept the attributive construction: *(un) incremento es un proceso (que)* … 'increase is a process (that)…'.

Each one of these subclasses of facts has, in turn, its own subclasses. For example, PROCESO includes PROCESO FÍSICO 'physical process', PROCESO FISIOLÓGICO 'physiological process' and PROCESO SOCIAL 'social process'. The noun *regeneración* (in a context as *La regeneración de los tejidos periodontales* 'Regeneration of periodontal tissues') would be labelled PROCESO FISIOLÓGICO.

2.2 Subclasses of Entities

Our hierarchy comprises twenty subclasses of entities, namely (in alphabetical order): ACUMULACIÓN 'accumulation', ALGO QUE ESTÁ EN DETERMINADA RELACIÓN CON ALGO 'something that stands in a certain relation with', ALGO QUE ESTÁ EN DETERMINADO ESTADO 'something that is in a certain state', ALGO QUE SE CONSUME 'something that is consumed', ÁMBITO DE ACTIVIDAD 'area of activity', BIEN 'property', CONJUNTO 'set', CREACIÓN 'creation', ENTIDAD GEOLÓGICA 'geological entity', ENTIDAD INFORMACIONAL 'informative entity', ENTIDAD SOCIAL 'social entity', ENTIDAD VISUAL 'visual entity', LUGAR 'place', LUGAR ABSTRACTO 'abstract place', MATERIA 'matter', OBJETO FÍSICO 'physical object', OCUPACIÓN SOCIAL 'social occupation', SER IMAGINARIO 'imaginary being', SER VIVO 'living being' and SUMA DE DINERO 'amount of money'.

Some of these classes encompass a very large number of lexical units. Such is the case of SER VIVO 'living being' that includes the labels HUMANO 'human', ANIMAL 'animal' and VEGETAL 'vegetal'. Even leaving aside terminology, the number of lexical units referring to humans is very large. Subclasses of humans can be precisely characterized by means of predicates that select them in a specific way. We observe, for example, that *(Contratar, despedir) a un camarero* ('To hire, to fire a waiter') is acceptable, but not **(Contratar, despedir) a un sacerdote* (*'To hire, to fire a priest'). *Camarero* and *sacerdote* will then be in two different subclasses of INDIVIDUO QUE PRACTICA UN OFICIO 'individual that has a profession'. The methodology of the "classes of objects" [3, 7] is based upon this property of some predicates.

The label OBJETO FÍSICO 'physical objet' is another example of a class that subsumes a considerable number of subclasses. The larger of these subclasses is ARTEFACTO 'artefact', which introduces the important difference between natural objects and artificial ones. Interestingly, artefacts often present verbs of realization or fulfillment as specific collocational values: **Real**$_1$ (*coche*) = *conducir* 'to drive a car', **PreparFact**$_0$ (*coche*) = *poner gasolina* 'to fill up the car', **Real**$_2$ (*bus*) = *ir en* 'to ride on a bus', **PreparReal**$_2$ (*taxi*) = *parar* 'to hail a taxi', etc.

3 Semantic Apparatus

It is important to highlight that a semantic label is not attributed to a form (that can be ambiguous and, therefore, require more than a semantic label) but to a lexical unit. Several methods can be used in order to individualize a lexical unit in a dictionary. In our case, we resort to an actantial formula accompanied by an example.

The actantial formula of a lexical unit is a linguistic expression that includes the form of this lexical unit and its actants (identified by variables: X, Y, Z… and semantically labelled if necessary). For instance, the actantial formula of *acusación* corresponds to *ENUNCIADO que la PERSONA X emite* contra la PERSONA Y a propósito del HECHO Z 'STATEMENT that the PERSON X makes against the PERSON Y concerning the FACT Z'. An example could be: *Se tomó mi observación como una acusación personal* 'He took my remark as a personal accusation'. In the example *El fiscal retiró la acusación contra el ex-diputado* 'The prosecutor withdrew the accusation against the congressman', the actantial formula would be *ACTO JURÍDICO de la PERSONA X contra el INDIVIDUO Y debido a su ACCIÓN Z presentada ante la AUTORIDAD JUDICIAL W* 'JURIDICAL ACT of the PERSON X against the INDIVIDUAL Y because of his ACTION Z'. For the sentence *El Ayuntamiento se personó como acusación particular en aquel caso* 'The city council entered its appearance as private prosecutor in this case', the actantial formula would be *PERSONA X que presenta la acusación Z (contra el INDIVIDUO Y debido a su ACCIÓN Z)* 'PERSON X who presents the accusation Z (against the INDIVIDUAL Y because of his ACTION Z)'. Since our description has a semantic nature, we do not specify in the actantial formula the syntactic actants of the described lexical unit.

In principle only predicates (that denote facts) can have actantial formulae, "pure" semantic names (that denote entities) are accompanied only by an example: *avena*, *La avena ayuda a adelgazar* 'Oats help one to lose weight'. However, many lexical units denoting entities do have an actantial formula that they inherit from the particular situations to which they are related. For example, *tripulación* 'crew' denotes a set of human beings but inherits the semantic actants of tripular 'to crew' and has then the actantial formula: *CONJUNTO DE INDIVIDUOS X del MEDIO DE TRANSPORTE Y* 'SET OF INDIVIDUALS X of the MEANS OF TRANSPORT Y'. *Bocadillo* 'sandwich' presents the formula *PREPARACIÓN ALIMENTÍCIA del INDIVIDUO X hecha con el pan Y y el ALIMENTO Z* 'ALIMENTARY PREPARATION of the INDIVIDUAL X made with the bread Y and the NOURISHMENT Z' that it inherits from the fact that it is prepared in a certain way for the purpose of nourishing X.

4 Methodology and Results

In order to build our hierarchy we are applying a top-down approach as well as a bottom-up one.

We use a top-down approach because our concept of the semantic label is based on the Meaning-Text Theory [4, 6] (more precisely, on the Explanatory and Combinatorial Lexicology) and on its lexicographical developments, such as the DiCo (*Dictionnaire de Combinatoire*) [11] and the *Réseau Lexical du Français* (RLF) [8]. The French hierarchy of semantic labels developed in the frame of these projects is a great advantage when outlining the general structure of our Spanish hierarchy, at least until the third level of labels.

Nevertheless, even when working with closely related languages, such as French and Spanish, the top-down approach cannot reach a satisfactory degree of precision. As a result, we need to adopt a bottom-up strategy that consists mainly of the manual labelling of a large number of Spanish lexical units. By "manual" labelling we mean the assignment, by a lexicographer, of an actantial formula and a semantic label to a disambiguated lexical unit. We plan to label 20,000 Spanish lexical units extracted from our Spanish Electronic Dictionary of Spanish (integrated in the NooJ linguistic development environment [12]), of which we have labelled approximately 8,000 up to the present day. This labelling of assorted and relatively usual Spanish lexical units allows us to progressively build up the different levels of our hierarchy (up to eight at this moment) and postulate the necessary labels.

Labelling is performed without previously ordering the lexical units. Peer-to-peer revisions (especially overall revisions of the lemmata attributed to a given semantic label until a precise moment, as well as revisions that focus on overrepresented labels) ensure, from our point of view, a fair degree of accuracy and homogeneity. Nevertheless, only systematic tests performed when our quantitative goal is attained will be able to ensure real robustness.

It is worth highlighting that our methodology is, by no means, the quickest way to semantically label a dictionary. It could even be said that it is a particularly arduous one. But it is important to bear in mind that our final goal is not only (and not mainly) the labelling of the dictionary but the development of a hierarchy of classes that accurately represent the lexical semantics of Spanish. In our opinion, that can only be done by a team of trained lexicographers applying their know-how to a large sample of the lexicon. Subsequently, there would be no need to continue with this procedure.

One way to significantly increase the lexical coverage would be simply to look for lexical units that correspond to a certain label. If *clavel* 'carnation', *dalia* 'dahlia' and *gardenia* 'gardenia' are labelled FLOR 'flower', nothing would prevent us compiling a list of flowers and then checking for their presence in the dictionary. This strategy is particularly suited for populating the hierarchy with multiword lexemes, since they are normally unambiguous. Of course, it will work much better for deep classes belonging to technical domains than for shallow classes or for labels incorporating evaluative meanings.

5 Further Applications

The obvious field of application for a hierarchy of semantic labels, as we conceive it, is the electronic lexicography. By itself, the mere implementation of an extensive system of semantic labels inside a large coverage electronic dictionary opens the field to a large range of applications. This is especially true when working with a fully-fledged linguistic development environment like NooJ. The possibility to annotate semantically very large corpora, to combine this information with the available morphological and syntactic descriptions, and to include it, when needed, in finite-state machines and regular grammars offers a world of possibilities for the processing of natural language. Let us point out that semantic labels are very often the only real semantic information that PLN systems can access. And semantic information is crucial for so many applications.

It goes without saying that semantic information is particularly relevant for search engines and for machine translation. In this latter field, the combined use of semantic labels and actantial formulae can be a reliable method of disambiguation and, therefore, of precise translation.

Let us return to the example of *acusación*. In the sentence *El Ayuntamiento se personó como acusación particular en aquel caso* the subject of *personarse* can only be a human being. The selected actantial formula for acusación will be then: *PERSONA X que presenta…'*, which is linked to the translation equivalent 'prosecutor'. In this context, *acusación* can be safely translated by 'prosecutor': 'The city council entered its appearance as private prosecutor in this case'. However, 'prosecutor' would be completely inappropriate for translating forms associated with the two other formulae of *acusación*, that do not correspond to human beings but to ENUNCIADO and ACTO JURÍDICO respectively: *'He took my remark as a personal prosecutor'. *'The prosecutor withdrew the prosecutor against the congressman'. The combined use of the actantial formula of *personarse* and of the semantic label (next kind) of *acusación* allows us to select the appropriate translation equivalent.

Finally, from a more theoretical point of view, we think that the hierarchy of semantic labels can be used to accurately describe a variety of diachronic semantic changes that have up to now been referred to in a rather loose way. We are currently working in this direction.

6 Conclusion

We firmly believe that research in the area of semantic labels can be profitable for different areas of linguistics. Moreover, the more a hierarchized set of semantic labels is used in a varied range of applications, the more robust and reliable it will become. Of course, we are still far from having reached the point at which we can contemplate full-fledged real-world applications. Aside from achieving much better lexical coverage, we have to solve a number of procedure problems.

One of the questions that we need to address is how to adapt our hierarchy to label not only nouns but the other parts of speech as well [2]. Let us remember that the semantic label is a *genus proximum*. As a consequence, it is necessary that the semantic label

could replace in any context (stylistic considerations aside) the lexical units labelled by it. Since a noun cannot replace an adjective, nor a verb or an adverb, we need the corresponding sets of adjectival, verbal and adverbial labels. This sets are not to be rebuilt from scratch, but rather to be derived from the nominal labels resorting to the derivational paradigmatic lexical functions $\mathbf{A}_0$, e.g. $\mathbf{A}_0$ (*fuerza* 'strength') = *fuerte* 'strong'; $\mathbf{V}_0$, e.g. $\mathbf{V}_0$ (*muerte* 'death') = *morir* 'to die'; $\mathbf{Adv}_0$ e.g. $\mathbf{Adv}_0$ (*cuidado* 'care') = *cuidadosamente* 'carefully' [9, 11].

Finally, let us mention that semantic granularity is a question that requires further investigation, especially when considering particular applications. At the same level of depth, there are labels that represent semantic contents much more intuitively than others. For instance, DEPORTE 'sport' or COLOR 'colour' are perceived as more natural and easier to work with than DISPOSITIVO 'device' or ALGO DE CARÁCTER NEGATIVO 'something having a negative character'. While considering bilingual applications, it is worth asking if these perceptions will always be similar for both languages.

References

1. Blanco, X.: Etiquetas semánticas de HECHO como género próximo en la definición lexicográfica. In: Calvo, C., Lépinette, B., Anscombre J.-C. (eds.) Lexicografía en el ámbito hispánico, pp. 159–178. Universitat de València (2010)
2. Blanco, X.: Les étiquettes sémantiques comme genre prochain : le cas des verbes. Verbum **XXIX**(1–2), 113–125 (2007)
3. Gross, G.: Manuel d'analyse linguistique. Approche sémantico-syntaxique du lexique. Presses universitaires du Septentrion, Villeneuve-d'Ascq (2012)
4. Mel'čuk, I., Milićević, J.: Introduction à la linguistique. Hermann, Paris (2014). vol. 1
5. Mel'čuk, I., Polguère, A.: Lexique actif du français. L'apprentissage du vocabulaire fondé sur 20 000 dérivations sémantiques et collocations du français. De Boeck & Larcier, Bruxelles (2007)
6. Mel'čuk, I.: Semantics. From Meaning to Text. John Benjamins Publishing Company, Amsterdam (2012)
7. Lepesant, D.: Principles for a Semantic Classification of Verb Predicates. Language Research, Special Issue December 2003. Language Education Institute, Seoul National University, pp. 21–38 (2003)
8. Polguère, A.: From Writing Dictionaries to Weaving Lexical Networks. Int. J. Lexicography **27**(4), 396–418 (2014)
9. Polguère, A.: Classification sémantique des lexies fondée sur le paraphrasage. Cahiers de lexicologie **98**, 197–211 (2011)
10. Polguère, A.: Étiquetage sémantique des lexies dans la base de données DiCo. Traitement Automatique des Langues (TAL) **44**(2), 39–68 (2003)
11. Polguère, A.: Towards a theoretically-motivated general public dictionary of semantic derivations and collocations for French. In: Heid, U., Evert, S., Lehmann, E., Rohrer, C. (eds.) Proceedings of EURALEX 2000, pp. 517–528. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Stuttgart (2000)
12. Silberztein, S.: La formalisation des langues. L'approche de NooJ. ISTE Editions, London (2014)