A Taxonomy of Spanish Nouns, a Statistical Algorithm to Generate it and its Implementation in Open Source Code

Rogelio Nazar, Irene Renau
Instituto de Literatura y Ciencias del Lenguaje
Pontificia Universidad Católica de Valparaíso
Av/El Bosque, 1290, Viña del Mar (Chile)
{rogelio.nazar,irene.renau}@pucv.cl

Abstract
In this paper we describe our work in progress in the automatic development of a taxonomy of Spanish nouns, we offer the Perl implementation we have so far, and we discuss the different problems that still need to be addressed. We designed a statistically-based taxonomy induction algorithm consisting of a combination of different strategies not involving explicit linguistic knowledge. Being all quantitative, the strategies we present are however of different nature. Some of them are based on the computation of distributional similarity coefficients which identify pairs of sibling words or co-hyponyms, while others are based on asymmetric co-occurrence and identify pairs of parent-child words or hypernym-hyponym relations. A decision making process is then applied to combine the results of the previous steps, and finally connect lexical units to a basic structure containing the most general categories of the language. We evaluate the quality of the taxonomy both manually and also using Spanish Wordnet as a gold-standard. We estimate an average of 89.07% precision and 25.49% recall considering only the results which the algorithm presents with high degree of certainty, or 77.86% precision and 33.72% recall considering all results.

Keywords: corpus statistics, distributional semantics, Spanish, taxonomy induction

1. Introduction

Influenced by classical logic (Aristotle, on the Categories), modern linguistics since at least structuralist semantics defines a taxonomy of the nouns of a language as the specification of the semantic hierarchy of words in a tree-shape structure, where each link between words establishes their hyponym-hypernym relationships (Lyons, 1977), often called IS-A relations in NLP literature.

A number of properties of such type of structures are well-known, such as the concept of inheritance, i.e., the fact that the “children” nodes inherit the properties of their “parent” nodes. From a lexical point of view, words in children node positions are hyponyms, and words playing the role of parent nodes are hypernyms. Other common properties of taxonomies are asymmetry and transitivity. In some taxonomy X, asymmetry is simply defined as follows: ∀a, b ∈ X, (a → b) ⇒ ¬(b → a), i.e. a cannot simultaneously be the parent and the child of b. And transitivity is analogous to the concept of inheritance, thus ∀a, b, c ∈ X, if (a → b) ∧ (b → c) ⇒ (a → c). This means that taxonomies are directed acyclic graphs, and that it would be an infringement for a taxonomy to contain cycles. This does not imply, however, that the same word cannot have different hypernym chains, which is a different issue. It would be the case, for instance, of polysemous words, as will be shown later in the paper.

Taxonomies can be a useful resource in a number of different applications. Having the ability to replace instances of nouns or noun phrases in corpus by their immediate hypernym allows one to improve the quality of distributional similarity calculations by means of context normalisation (Périnet and Hamon, 2014). Another possible application is on coreference resolution in discourse analysis (Recasens and Hovy, 2009). Bordea et al. (2015) also point out different applications such as e-commerce and online data-bases among others, and it is easy to foresee other examples such as monolingual or bilingual terminology extraction, named entity categorisation, collocation extraction, etc. In our case, in a separate publication (Nazar and Renau, forthcoming), we apply this taxonomy to a corpus-based study of predicate-argument structures in Spanish. In that study, we differentiate the meanings of verbs by taking into account a combination of syntactic and semantic analysis of predicate-argument structures in the syntagmatic context, and the taxonomy of nouns allows us to discriminate different senses or patterns of use of verbs by analysing the semantic class of their arguments (see section 6).

In this paper we describe our work in progress in the development of the taxonomy, which is already freely available for the research community at the project's website1. This version of the taxonomy only contains nouns of general vocabulary and is limited at the moment to single-word lexical units. The application of this algorithm to the development of a specialised taxonomy with terminological units –including multiword expressions– is discussed in a separate publication due to its inherent complexity and differences of approach (Nazar, in preparation). The present paper thus focuses on the description of our statistically-based taxonomy induction algorithm and its implementation as a Perl script running on Linux, applied to general vocabulary words. We also present an example of the application of this algorithm to a Spanish corpus and we assess the the quality of the result in two different ways: precision is estimated through manual evaluation of a sample of the taxonomy by a group of 6 advanced students in linguistics, while recall is estimated automatically using Span-

1 All materials are available simultaneously in two different servers: (http://www.verbario.com/) and (http://www.tecling.com/taxo/) [last access: 3/4/2016]. Other mirror servers are scheduled to appear.
ish WordNet as a gold-standard. At the time of writing, we are replicating experiments in English and French. The article is structured as follows: we first make a short introduction to the state of the art in taxonomy induction (section 2). We describe our methodology (section 3), starting from a brief description of the shallow ontology we use, which includes the most general, top nodes of the taxonomy (subsection 3.1). We then describe each algorithm separately (subsections 3.3-3.5), and finally we explain how we integrate the data obtained from the previous strategies in a single output (subsection 3.6). We describe the evaluation of the process (section 4), we make a short explanation of technical details regarding the material we make available for the scientific community (section 5) and we finish with some conclusions and lines of future work (section 6).

2. Related Work

Automatic taxonomy induction has been a topic in computational linguistics since its early beginning. First efforts were focused on exploiting machine readable dictionaries to extract hyponymy or other types of semantic relations (Chodorow et al., 1985; Guthrie et al., 1990, among many others). Later, with the advent of corpus linguistics, similar methods and ideas were extrapolated from lexicographical to textual corpora, mainly by the use of what is now called Hearst’s patterns (Hearst, 1992). I.e., if one finds in a corpus a sequence such as X is a type of Y or other patterns such as X and other (types of) Y, and if X and Y are nouns or noun phrases, then some authors will assume that X is a hyponym of Y (Rydin, 2002; Cimiano and Völker, 2005; Snow et al., 2006; Pantel and Pennacchiotti, 2006).

Earlier methods have now been complemented by word-distribution analysis in very large corpora, inspired in Harris’ (1954) distributional semantics. First research on distributional similarity and semantic clustering (Grefenstette, 1994; Schütze and Pedersen, 1997; Lin, 1998) was followed by more complex algorithms for taxonomy induction and population (Ciaranita, 2002; Alfonseca and Mandhhar, 2002). In this line, graph-based algorithms seem to represent a new trend (Kozareva and Hovy, 2010; Nazar et al., 2012; Velardi et al., 2013).

The large body of work related to automatic taxonomy induction prevents us to make an extended revision in this paper. In general, research is focused either on the extension or adaptation of WordNet (Fellbaum, 1998) including languages other than English, or to the reuse and integration of data from different sources (de Melo and Weikum, 2013; Bansal et al., 2014; Fišer and Sagot, 2015).

WordNet has been extensively used in varied NLP tasks, and commonly as a gold-standard in the context of research on taxonomy induction (Bordea et al., 2015). To some extent, Wordnet differs from what we are trying to achieve in our project, despite the fact that we too use it for the evaluation of our resource. From a theoretical point of view, Wordnet is a top-down ontology of concepts, or more precisely, synsets, defined as sets of words that have the same sense or refer to the same concept. Maziarz et al. (2013) already pointed out that there exists a sort of tension between synsets and lexical units in Wordnet and EuroWordnet, regarding the differences between semasiological and onomasiological perspectives. Unlike WordNet, our approach is semasiological and aimed at the analysis of the vivid dynamics of the lexicon with data obtained with bottom-up, corpus-based methods. We believe this difference make the two resources hard to compare. From a theoretical perspective, only the semasiological approach seems to be useful for the study of lexical semantics. From a practical perspective, one cannot compare a hand-made resource like WordNet with an automatically generated taxonomy like ours. Evidently, a generalised problem in all hand-made taxonomies is the fact that they quickly become outdated and require large amounts of human and technical resources. The automatically generated taxonomy, in turn, must deal with problems of recall, structure and polysemy.

In spite of these limitations, proposals in automatic methods can contribute to complement or even substitute manual taxonomies and make these resources more adaptable to different languages and purposes. This is even clearer specifically in fully statistically-based taxonomies emerging from corpus data in the line of Bullinaria & Levy (2007) or Strohmaier et al. (2012).

Bordea et al. (2015) offer a recent description of state of the art in taxonomy induction and present the results of different teams that participated in the SemEval-2015 Task on Taxonomy Extraction Evaluation. Again, our results are not directly comparable, and the same SemEval organisers admit that there is still much to be done to find an effective way to compare the results of different taxonomy induction algorithms.

More references to related work—most often our own previous work on the subject—are also indicated in the next section as earlier versions of some specific components of the main algorithm have already been published elsewhere.

3. Methods and Materials

The methodology employed in this research consists of a combination of different statistical algorithms that take into account the distributional behaviour of lexical units in order to build a taxonomy. As stated in previous lines, we implement corpus-based language-independent algorithms, i.e., we disregard the use of linguistic information with the explicit purpose of facilitating the replication of experiments in other languages (see section 6).

As already stated in the introduction, in this section we first make a short description of a hand-made shallow ontology that is used as a basic structure for the most general nodes (subsection 3.1) and then we describe a series of statistical techniques to populate such ontology by hyponymy-hyponymy relation discovery (subsections 3.2-3.6). We apply our method on a collection of ca. 45,000 nouns of general vocabulary extracted from corpus and we estimate our taxonomy will double that number shortly.

3.1. The Starting Point: a Basic Ontology

Hanks (Jezeck and Hanks, 2010; Hanks, In process) has presented a shallow ontology (named CPA Ontology) consisting of a hierarchical organisation of around 250 of the most general concepts of the language, which he calls “semantic types”. These are conventional labels such as “Hu-
man”, “Food”, “Device”, “Emotion”, “Activity” and so on. As already explained, we use this ontology as an initial basic structure. We are aware that linguistic and cultural differences can have a great impact on how ontologies are structured, but we nonetheless consider that the CPA ontology is general enough to be used in a large amount of languages sharing cultural aspects, as at least most European languages do.

The procedure consists first in translating the semantic types to the target language and then populating this basic structure with lexical units, extending it progressively with new branches using the algorithms we describe next. This process involves some degree of manual work, which consists of adding the most general words to the ontology, i.e. the high-frequency nouns with the most general or abstract meanings. For instance, we manually paired our node *Recipient* with its equivalent “Container”, one of the semantic types in the CPA ontology. In fact, the whole process could have been conducted automatically applying machine translation to Hank’s semantic types, but being only a few hundred nodes, we decided it would be better to proceed with this step by hand.

After stating the Spanish equivalents of the English semantic types, all the hyponyms are added automatically with our algorithms. In the case of *Recipient*, a hyponym such as *botella* (‘bottle’) is added, and this one in turn also becomes a hyponym of other nouns such as *licorera* (‘decanter’). The following would be an example of the connection between the CPA Ontology and the noun *botella*, taken from corpus and automatically connected with *Recipient*:

\[
\rightarrow \text{Todo ("Anything")}
\rightarrow \text{Entidad ("Entity")}
\rightarrow \text{Objeto Físico ("Physical Object")}
\rightarrow \text{Inanimado ("Inanimate")}
\rightarrow \text{Artefacto ("Artifact")}
\rightarrow \text{Recipient ("Container")}
\rightarrow \text{botella ("bottle")}
\]

### 3.2. Statistical Analysis of Dictionary Entries

As mentioned in section 2, machine readable dictionaries have been used in the past for the extraction of hyponymy relations. The problem with this method, as reported in the literature, is that it is labour intensive, it must be specific for every particular dictionary, and it is also error prone. A more general methodology, independent of the specific dictionary and language, has been reported in previous work (Nazar and Janssen, 2010; Renau and Nazar, 2012). In these studies, multiple dictionaries were used to calculate simple co-occurrence of words in the entries with the words in the definitions. It follows that the most frequent noun in the set of definitions of another noun is likely to be the correct hyponym. E.g., the most frequent noun within all dictionary definitions of the word *motocicleta* (‘motorcycle’) will most likely be *vehículo* (‘vehicle’). This observation lead to the development of full networks of hyponymic links, identifying hubs or nodes that are repeatedly selected as hyponyms by many other words.

The main drawback of those previous attempts was that there was no basic structure to build on, and it is difficult to build a taxonomy from scratch. In the case of the present paper, however, the scenario is different because we integrate the resulting hyponym-hypernym pairings with the CPA Ontology. As a result of the application of this first step, the CPA ontology, which initially contained a set of 250 semantic types, has grown to a total of 2,290 categories, each one representing a parent node populated with hyponyms. This initial tree then continues to be populated by the algorithms subsequently described in this section.

### 3.3. Distributional Similarity

Again as explained in section 2., semantic clustering based on paradigmatic relations is among the oldest and most extended techniques in distributional semantics. The underlying assumption is that two words that tend to occur in very similar contexts must be semantically similar. Importantly, these words may not be syntactically related (they are not seen in presence of each other) but they are used in the same positions, i.e., they are paradigmatically related.

Our earlier attempt using this type of distributional similarity was aimed at semantic clustering of nouns (Nazar and Renau, 2015a) using bigrams taken from Google Books N-Gram Corpus. That approach had the problem of being computationally too expensive due to its quadratic complexity, which makes it difficult to scale from small samples of words to a full lexicon. This time, however, and because now we build on the CPA ontology, instead of a clustering procedure now we have framed the task as a categorisation problem, which has helped us achieve better efficiency and accuracy.

If $W$ is the theoretical set of the nouns of a language and $C = \{c_1, ..., c_{|C|}\}$ the set of semantic categories in which every noun $w_i$ can be classified, then for each pair $(w_i, c_j) \in W \times C$ we obtain a score which will confirm or disprove the membership of noun $w_i$ to category $c_j$, and this score represents their distributional similarity.

In order to be able to produce a distributional similarity score, words need to be analysed within a particular corpus. We used the text of Spanish Wikipedia as a corpus, not because we have a particular interest on it, but because it is large and freely available. In fact, we believe that any other large enough corpus should have the same effect. We processed this corpus extracting only the text of the pages, with the pages sorted in random order and excluding all metadata and code that make the hierarchical organisation of the content. This resulted in a single block of plain text of ca. 700 million tokens. In order to improve computational efficiency, we divided the distributional similarity calculation in two steps. Normally, distributional similarity is calculated by building word-vectors, where dimensions are defined as the words that tend to co-occur with any given target word. Instead of doing that directly, as a first step we built category-vectors, where the dimensions are the sum of all dimensions in the vectors of the words contained in such category, thus

\[
\text{For clarity’s sake, we capitalize semantic types (concepts in the ontology) and write lexical units in italics.}
\]

\[
\text{3 (https://books.google.com/ngrams/) [last access: 3/4/2016]}
\]

\[
\text{4 Indeed, at the time of writing we are replicating all our experiments using the esTenTen corpus (Kilgarriff and Renau, 2013), which is considerably larger than Wikipedia.}
\]
our set $C$ of semantic categories is redefined here as a set of category-vectors. The efficiency gain is of course explained by the fact that there are always less categories than category members ($|C| < |W|$), and then the number of comparisons is significantly reduced. Thus, given an input noun $w_i$, we calculate the overlap coefficient (equation 1) between noun-vector $\vec{w}_i$ and every category-vector $\vec{c}_j \in C$.

$$O(\vec{w}_i, \vec{c}_j) = \frac{|\vec{w}_i \cap \vec{c}_j|}{\min(|\vec{w}_i|, |\vec{c}_j|)} \tag{1}$$

If $O(\vec{w}_i, \vec{c}_j) > t \Rightarrow w_i \in c_j$ ($|c_j| = 100$ and $t = 4$, both arbitrary parameters). For $w_i$, we can expect that this will be true for more than one category. Thus, we will obtain a new set $H_i$ defined as a short list of pre-selected categories for $w_i$ (a list of length $l$, another arbitrary parameter which in our experiments is equal to 7). Next, $\forall c_j \in H_i$, we now calculate a Jaccard similarity coefficient (equation 2) between word-vector $\vec{w}_i$ and each word-vector $\vec{c}_j,k$.

$$J(\vec{w}_i, \vec{c}_j,k) = \frac{|\vec{w}_i \cap \vec{c}_j,k|}{(|\vec{w}_i| + |\vec{c}_j,k| - |\vec{w}_i \cap \vec{c}_j,k|)} \tag{2}$$

We assume that if $J(\vec{w}_i, \vec{c}_j,k) > t/2 \Rightarrow \vec{w}_i \sim \vec{c}_{j,k}$, and after this comparison there will be a proportion $p$ of similar vs. different cases\(^5\) in every category $c_j$. Then, if $p > t \Rightarrow w_i \rightarrow c_j$, i.e., if noun $w_i$ proves to be distributionally similar to many members of category $c_j$, then $w_i$ is hyponym of $c_j$.

### 3.4. Co-occurrence Graphs

In the previous subsection we applied distributional analysis to find paradigmatic relations between nouns and used that information to associate nouns with semantic categories. In this section, instead, we present a different, complementary approach, based on the study of syntagmatic relations.

This other strategy is based on our observation, already reported in previous work (Nazar et al., 2012; Nazar and Renau, 2012), that words that engage in a hyponym-hypernym relation tend to show a type of asymmetric co-occurrence. That is to say, the presence of a noun in a given sentence can be seen as a predictor of the presence of its hypernym in the same sentence while the opposite is not true. For instance, it is more probable that motocicleta (‘motorcycle’) will co-occur with its hypernym, vehículo (‘vehicle’), than the other way around.

An intuitive way to visualise these asymmetric relations is using co-occurrence graphs. Figure 1 shows one of these graphs, a case obtained from real data using the same corpus mentioned in subsection 3.3. The interpretation is that, given the input word motocarro (‘three-wheeler’), represented as a node in the top of the figure, its asymmetric co-occurrence with other nouns is drawn as a directed graph. In this case, motocarro tends to co-occur with furgoneta (‘van’), motocicleta and vehículo. The noun motocicleta, as already mentioned, also tends to co-occur with vehículo. The node furgoneta, in turn, co-occurs with vehículo and camión (‘truck’), and the latter also co-occurs with vehículo. It seems that all these co-occurring relations end up in the correct hypernym, as if it would be a case of natural selection. The way we see these graphs is as if, for any given input node, the node with more incoming arrows is selected as the most likely hypernym.

\(^5\) It should be noticed that proportion $p$ requires a normalisation procedure because the number of member in each category can be very dissimilar. There are different ways to do this, but in our case we just used $p = \sqrt{d/s}$, where $s$ is the number of similar cases and $d$ the number of different ones.
3.5. Analogical Inference

The algorithm described in this subsection is different from the previous ones in that it does not take corpora as input. Instead, it builds a table of correspondences between semantic categories and morphological features, which are simply defined as sequences of 3-5 letters at the end of each word. Thus, it takes the output of the previous algorithms as input and then associates, for instance, that terms ending with character sequences such as -itis with characters such as dermatitis or endometriosis with diseases. This way, it will be able to associate new terms such as dermatitis or endometriosis with this category.

The idea of associating morphological features with semantic classes has been reported in the past (Ciaramita, 2002; Nazar et al., 2012; Bansal et al., 2014), as well as detecting a string overlap (i.e., the fact that hypernyms are sometimes included in their hyponyms, as in virus and adenovirus, although this is more often the case in multiword expressions). The main difference here lies in the fact that in this case the association gets registered automatically, as reported in Nazar et al. (2012). Perhaps the best way to understand this difference is in the case of multiword expressions, despite the fact that they are not analysed in this paper. When the association between these features and the semantic classes is automatic, there is no need to find a string overlap between hyponym and hypernym. That is to say, this algorithm will be robust enough to associate expressions such as Maffucci syndrome or Reiter syndrome with diseases even when there is no string overlap, because the lexical unit syndrome has been frequently found in terms pertaining to the class of diseases.

3.6. Integration of Results

The last algorithm is in charge of the task of integrating the results of each of the previous algorithms into a single and coherent taxonomical structure. The integration of results has been solved with a fairly simple voting scheme, which results in a sort of “consensus” taxonomy, far more reliable than the results of each method taken in isolation.

The rule is as follows: if for an input noun $w_i$ there are more than two algorithms that coincide in placing $w_i$ under the category $c_j$, then the link $w_i \rightarrow c_j$ is presented with a high degree of certainty. If, instead, only two algorithms coincide in this, then such link only has a low degree of certainty. Else, the link is ignored.

We see now that this procedure is more simple and effective than our previous integration attempt (Nazar and Renau, 2015b), which we have now abandoned.

4. Results and Evaluation

The results of our experiments can be assessed with a qualitative evaluation by selecting a sample of nouns and inspecting for each of them the obtained ascending hypernymy chains until reaching the top node of the ontology. This type of evaluation is useful to check if the different senses of a potentially polysemous word are registered in the taxonomy, and most importantly, if the prototypical senses are attested in the resource. Consider, for illustration, the case of the word fresa (‘strawberry’), which resulted in the chains shown in Table 1. In chain 1, the word is connected to arbusto (‘bush’), which is in turn linked to the semantic type Planta (‘Plant’). Strictly speaking, the first connection is incorrect because fresa, as ‘strawberry’, is not a kind of bush but an herbaceous plant, regardless of the fact that it is true that arbusto is a kind of plant and the rest of the chain is then correct. Chain 2, which defines fresa as a fruit, is correct and identifies the prototypical meaning of the word. Chains 3 and 4 are also correct and identify secondary, metaphorical meanings: a kind of tool and a kind of color. It is important to point out that, in the case of chain 2, fruto (‘fruit’) is classified as a type of “Food” and not as a type of “Plant Part”, which is another existing node in CPA Ontology. This is because our taxonomy emerges naturally from corpus and does not obey to rigid structures such as those that may appear in a hand-made taxonomy. Being strawberries very popular as a kind of food, they can be considered as such. Ultimately, according to the Diccionario de la Real Academia Española (DRAE)

Table 1: Different taxonomy chains for the Spanish noun fresa (‘strawberry’)

| Chain | Hypernymy Chain |
|-------|-----------------|
| 1     | fresa → arbusto → Planta |
| 2     | fresa → fruto → Artefacto |
| 3     | fresa → Herramienta |
| 4     | fresa → Color |

http://dle.rae.es/?id=ISjyrn6—ISkwvnu

(last access: 3/4/2016)
5. Code and Resources

The website of the project (see URL's in footnote 1) currently hosts a search interface to query the database and obtain results. In addition, regular dumps are made available for download because the taxonomy is growing as it processes more corpora. The taxonomy has a SNOMED-like structure, i.e., there is one table that associates the names of the nodes (the lexical units) with a unique numerical identifier. A second table presents rows as comma-separated list of numbers indicating the whole ascending hypernymy chain of each node. In this database, id numbers in table 1 are inversely correlated with their frequency of occurrence in table 2, in order to obtain minimal redundancy of the data.

---

Table 2: Evaluation of precision of the results of the algorithm

| Category | High Certainty | All results |
|----------|----------------|-------------|
|          | Ok  | Total | P    | Ok  | Total | P    |
| animal   | 76  | 104   | 73.08| 104 | 165   | 63.03|
| bone     | 60  | 64    | 93.75| 98  | 110   | 89.09|
| colour   | 90  | 114   | 78.95| 177 | 300   | 59   |
| dance    | 58  | 58    | 100  | 104 | 112   | 92.86|
| device   | 322 | 361   | 89.20| 735 | 1276  | 57.6 |
| disease  | 330 | 348   | 94.83| 621 | 763   | 81.39|
| doctrine | 75  | 82    | 91.46| 212 | 272   | 77.94|
| furniture| 36  | 39    | 92.31| 59  | 68    | 86.76|
| machine  | 96  | 100   | 96   | 158 | 206   | 76.7 |
| mammal   | 62  | 66    | 93.94| 115 | 121   | 95.04|
| specialist| 51  | 100   | 51   | 89  | 100   |
| vehicle  | 44  | 58    | 75.86| 58  | 94    | 61.67|
| weapon   | 55  | 66    | 83.33| 69  | 88    | 78.41|
| wine     | 16  | 19    | 84.21| 55  | 78    | 70.51|
| Average  | 89.07|       |       | 77.86|

Table 3: Evaluation of precision with the semantic category “fruits” for the purpose of measuring inter-coder agreement

| Rater | High Certainty | All results |
|-------|----------------|-------------|
|       | Ok  | Total | P    | Ok  | Total | P    |
| E1    | 144 | 153   | 94.12| 204 | 247   | 82.59|
| E2    | 129 | 153   | 84.31| 168 | 247   | 68.02|
| E3    | 123 | 153   | 80.39| 170 | 247   | 68.83|
| E4    | 128 | 153   | 83.66| 178 | 247   | 72.06|
| E5    | 127 | 153   | 83.01| 181 | 247   | 73.28|
| E6    | 131 | 153   | 85.62| 187 | 247   | 75.71|
| Average| 85.19|       |       | 73.41|

Table 4: Evaluation of recall in results presented with high degree of certainty by the algorithm, using WordNet as gold-standard

| Category | WN  | Ours | ⋂ | WN’s R | Our R |
|----------|-----|------|---|--------|-------|
| animal   | 153 | 117  | 22 | 18.8   | 14.38 |
| bone     | 69  | 64   | 21 | 32.81  | 30.43 |
| color    | 67  | 96   | 33 | 34.38  | 49.25 |
| dance    | 39  | 59   | 24 | 40.68  | 61.54 |
| device   | 1037| 365  | 158| 43.29  | 15.24 |
| doctrine | 14  | 82   | 0  | 0      | 0     |
| furniture| 79  | 41   | 8  | 19.51  | 10.13 |
| disease  | 165 | 348  | 57 | 16.38  | 34.55 |
| machine  | 96  | 65   | 16 | 24.62  | 16.67 |
| mammal   | 315 | 103  | 77 | 74.76  | 24.44 |
| specialist| 289 | 51   | 20 | 39.22  | 6.92  |
| vehicle  | 204 | 62   | 34 | 54.84  | 16.67 |
| weapon   | 55  | 66   | 21 | 31.82  | 38.18 |
| wine     | 13  | 18   | 5  | 27.78  | 38.46 |
| Average  | 32.78|     |    | 25.49  |       |

---

Note: The website of the project ([last access: 3/4/2016](http://multiwordnet.fbk.eu/english/home.php)) currently hosts a search interface to query the database and obtain results. In addition, regular dumps are made available for download because the taxonomy is growing as it processes more corpora.

---

Footnote 1: The taxonomy has a SNOMED-like structure, i.e., there is one table that associates the names of the nodes (the lexical units) with a unique numerical identifier. A second table presents rows as comma-separated list of numbers indicating the whole ascending hypernymy chain of each node. In this database, id numbers in table 1 are inversely correlated with their frequency of occurrence in table 2, in order to obtain minimal redundancy of the data.
and reduce file size. Alternatively, a navigable html version of the data is also offered for the less technically advanced user.

6. Conclusions and Future Work

The work we have presented is an ongoing methodological proposal to create a taxonomy from corpus data, using a set of algorithms which only apply quantitative strategies. As such, we consider that it is relatively easy to replicate the procedure in other languages, and that is one of the lines of research we are conducting at the moment.

We are aware that ours is still work in progress and there are many problems to be addressed. We can summarise the following lines of future work:

1) To increase precision making a more detailed error analysis. For the moment, results show that many of the mistakes are related to the polysemy of the nodes of the taxonomy and inclusion of other relations such as meronymy. A strategy to depurate our taxonomy is currently being tested (Nazar et al., submitted).

2) To establish a better methodology to calculate recall instead of using WordNet as gold-standard. There are different aspects to be taken into account. For example, if results show the prototypical meaning of the word, or the most common uses of the word, etc.

3) To increase the number of nouns in the taxonomy until we cover the vast majority of the Spanish language.

4) To replicate the methodology in other languages. At the moment we are working with English and French.

5) To apply the same method to a specialised vocabulary in order to get domain-specific taxonomies, including multiword expressions.

6) To include more taxonomy induction strategies. There is one algorithm in particular that has already been tested and will be published elsewhere (Nazar & Renau, forthcoming), in which we used a very restricted notion of context such as a fixed position in word ngrams. Consider the case of a sequence such as *es un jugador de* (*“is a * player”*). One can expect that names of games or sports will appear in the position of the asterisk: *es un jugador de golfe, rugby, tenis...* (*“is a golf/rugby/tennis... player”*). Of course there will also be cases such as *es un jugador de talento/nivel/Estados Unidos/equipo...* (*“is a player of talent / is a level player / is a player from United States / is a team player...”*), but the key aspect here is that we do not relate words because they share a single ngram but a very large number of different ones.

We would like to finish this paper referring to a parallel project in which we are applying our taxonomy to the creation of syntagmatic verb patterns which are connected to the different meaning of words, following Hanks' (2013) Corpus Pattern Analysis. For example, in a sentence such as *El abuelo se murió de un infarto* (*“Grandfather died of a heart attack”*), the verb *morir* (*“to die”*) has a different meaning than in the case of *El abuelo se murió de vergüenza* (*“Grandfather felt very embarrassed”*). The main difference between both senses or patterns of use of the verb *‘to die’* in both sentences is indicated by the fact that the direct objects have different semantic types (*“Disease”* and *“Emotion”, respectively*). Labelling verb arguments with semantic types allows us to identify these structures and discriminate verb meanings automatically (Nazar and Renau, forthcoming).

7. Acknowledgements

This research is supported by two grants from the Chilean Government: Conicyt-Fondecyt 11140686, “Inducción automática de taxonomías de sustantivos generales y especializados a partir de corpus textuales desde el enfoque de la lingüística cuantitativa” (lead researcher: Rogelio Nazar) and Conicyt-Fondecyt 11140704, “Detección automática del significado de los verbos del castellano por medio de patrones sintáctico-semánticos extraídos con estadística de corpus” (lead researcher: Irene Renau). We would like to thank the anonymous reviewers of the paper for their useful comments and the students who helped us with the evaluation.

8. Bibliographical References

Alfonseca, E. and Manandhar, S. (2002). Extending a lexical ontology by a combination of distributional semantics signatures. In Lecture Notes in Computer Science, pages 247–3.

Aristotle. (n.d.). Categories.

Bansal, M., Burkett, D., de Melo, G., and Klein, D. (2014). Structured learning for taxonomy induction with belief propagation. In Proc. of ACL, Baltimore, Maryland, USA, June.

Bordea, G., Buitelaar, P., Faralli, S., and Navigli, R. (2015). Semeval-2015 task 17: Taxonomy extraction evaluation (texteval). In Proc. of the 9th International Workshop on Semantic Evaluation (Semeval 2015), pages 902–910, Denver, Colorado, June.

Bullinaria, J. and Levy, J. (2007). Extracting semantic signatures. In Behavior Research Methods, 39(3):510–526.

Chodorow, M., Byrd, R., and Heidorn, G. (1985). Extracting semantic hierarchies from a large on-line dictionary. In Proc. of the 23rd annual meeting on ACL (Chicago, Illinois, USA), pages 299–304.
Ciaramita, M. (2002). Boosting automatic lexical acquisition with morphological information. In Proc. of the ACL-02 Workshop on Unsupervised Lexical Acquisition - Volume 9, ULA '02, pages 17–25, Stroudsburg, PA, USA.

Cimiano, P. and Völker, J. (2005). Text2onto: A framework for ontology learning and data-driven change discovery. In Proc. of the 10th International Conference on Natural Language Processing and Information Systems, NLDB’05, pages 227–238, Berlin, Heidelberg. Springer-Verlag.

de Melo, G. and Weikum, G. (2013). Taxonomic data integration from multilingual wikipedia editions. Knowledge and Information Systems, 39(1):1–39.

Fellbaum, C. (1998). WordNet: An Electronic Lexical Database. Bradford Books.

Fišer, D. and Sagot, B. (2015). Constructing a poor man’s wordnet in a resource-rich world. Language Resources and Evaluation, 49(3):601–635.

Grefenstette, G. (1994). Explorations in Automatic Thesaurus Discovery. Kluwer Academic Publishers, Norwell, MA, USA.

Guthrie, L., Slator, B., Wilks, Y., and Bruce, R. (1990). Is there content in empty heads? In Proc. of the 13th International Conference on Computational Linguistics, COLING’90 (Helsinki, Finland), pages 138–143.

Hanks, P. (In process). CPA ontology. http://www.pdev.org.uk/#onto. [last access: 1/3/2016].

Harris, Z. (1954). Distributional structure. Word, 10(23):146–162.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In Proc. of the 14th Conference on Computational Linguistics, COLING’92, pages 539–545, Stroudsburg, PA, USA.

Jezek, E. and Hanks, P. (2010). What lexical sets tell us about conceptual categories. Lexis: E-Journal in English Lexicology, (4):7–22.

Kilgarriff, A. and Renau, I. (2013). esTenTen, a vast web corpus of peninsular and american spanish. Procedia - Social and Behavioral Sciences, 95:12 – 19. 5th International Conference on Corpus Linguistics (CILC2013).

Kozareva, Z. and Hovy, E. (2010). A semi-supervised method to learn and construct taxonomies using the web. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP ’10, pages 1110–1118, Stroudsburg, PA, USA. ACL.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In Proc. of the 17th International Conference on Computational Linguistics - Volume 2, COLING ’98, pages 768–774, Stroudsburg, PA, USA.

Lyons, J. (1977). Semantics, volume 2. Cambridge University Press.

Maziarz, M., Plasecki, M., and Szpakowicz, S. (2013). The chicken-and-egg problem in wordnet design: synonymy, synsets and constitutive relations. Language Resources and Evaluation, 47(3):769–796.

Nazar, R. and Janssen, M. (2010). Combining resources: Taxonomy extraction from multiple dictionaries. In Proc. of the International Conference on Language Resources and Evaluation, LREC, 17-23 May 2010, Valletta, Malta.

Nazar, R. and Renau, I. (2012). A co-occurrence taxonomy from a general language corpus. In Proc. of EURALEX 2012, pages 367–375, Oslo, Norway.

Nazar, R. and Renau, I. (2015a). Agrupación semántica de sustantivos basada en similitud distribucional: implicaciones lexicográficas. In Lingüística y diccionarios, pages 273–288. Universidad da Coruña, A Coruña.

Nazar, R. and Renau, I. (2015b). Ontology population using corpus statistics. In JOWO@IJCAI, volume 1517 of CEUR Workshop Proc. CEUR-WS.org.

Nazar, R. and Renau, I. (forthcoming). A quantitative analysis of the semantics of verb-argument structures. In Collocations and other lexical combinations in Spanish, pages 92–108. OSU Press, Ohio.

Pérez, P. and Pennacchiotti, M. (2006). Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In Proc. of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of ACL, ACL-44, pages 113–120, Stroudsburg, PA, USA.

Recasens, M. and Hovy, E., (2009). A Deeper Look into Features for Coreference Resolution, pages 29–42. Springer Berlin Heidelberg, Berlin, Heidelberg.

Renau, I. and Nazar, R. (2012). Hypernymy relations from definiens-definiendum co-occurrence in multiple dictionary definitions. Procesamiento del Lenguaje Natural, 49:83–90.

Rydin, S. (2002). Building a hypernymy lexicon with hierarchical structure. In Proc. of the ACL-02 Workshop on Unsupervised Lexical Acquisition - Volume 9, ULA ’02, pages 26–33, Stroudsburg, PA, USA.

Schütze, H. and Pedersen, J. O. (1997). A cooccurrence-based thesaurus and two applications to information retrieval. Inf. Process. Manage., 33(3):307–318, May.

Snow, R., Jurafsky, D., and Ng, A. Y. (2006). Semantic taxonomy induction from heterogenous evidence. In Proc. of the 21st International Conference on Computational Linguistics, Sydney, Australia.

Strohmaier, M., Helic, D., Benz, D., Körner, C., and Kern, R. (2012). Evaluation of folksonomy induction algorithms. ACM Trans. Intell. Syst. Technol., 3(4):74:1–74:22, September.

Velardi, P., Faralli, S., and Navigli, R. (2013). Ontolearn reloaded: A graph-based algorithm for taxonomy induction. Computational Linguistics, 39(3):665–707.