

# Automatic acquisition for low frequency lexical items

Núria Bel, Sergio Espeja, Montserrat Marimon

IULA – Universitat Pompeu Fabra

Pl. de la Mercè, 10-12. 08002 Barcelona, Spain

E-mail: nuria.bel@upf.edu, sergio.espeja@upf.edu, montserrat.marimon@upf.edu

## Abstract

This paper addresses a specific case of the task of lexical acquisition understood as the induction of information about the linguistic characteristics of lexical items on the basis of information gathered from their occurrences in texts. Most of the recent works in the area of lexical acquisition have used methods that take as much textual data as possible as source of evidence, but their performance decreases notably when only few occurrences of a word are available. The importance of covering such low frequency items lies in the fact that a large quantity of the words in any particular collection of texts will be occurring few times, if not just once. Our work proposes to compensate the lack of information resorting to linguistic knowledge on the characteristics of lexical classes. This knowledge, obtained from a lexical typology, is formulated probabilistically to be used in a Bayesian method to maximize the information gathered from single occurrences as to predict the full set of characteristics of the word. Our results show that our method achieves better results than others for the treatment of low frequency items.

## 1. Introduction

The work we present here handles a specific case of the task of lexical acquisition understood as the induction of information about the linguistic characteristics of lexical items on the basis of information gathered from their occurrences in texts. Research in lexical acquisition is based on the assumption that lexical items have regular patterns of syntactic behaviour and that these regular patterns distinguish classes that ultimately are semantic classes. Most of the recent work in the area of lexical acquisition has been concerned with the identification and use of such patterns. The way they identify and use such information are the basis of two lines of research: in one line, the induction of these patterns from data helps to predict lexical classes and the words that are members of such classes; in the other one, these patterns are sought in data as evidence for classifying words into pre-defined, linguistically motivated classes.

The contribution of our work is to address specifically the problem of handling the case of low frequency lexical items, because the performance of published methods decreases notably when, for instance, only few examples of a word are available. Both lines of research just mentioned have in common that need to take as much text as possible as source of evidence. But the importance of covering such low frequency items (these patterns and words) lies in the fact that a large quantity of them will be occurring few times, if not just once, in any particular collection of texts. Even more, according to Zipf's, this will be the case for most of the words, especially nouns, in any length size corpus (Zipf, 1935). Lexical acquisition for NLP must be able to handle these cases too because lexical coverage is crucial to achieve the proper performance of the processing components that rely in lexical information. For instance, Briscoe and Carroll (1993) observed that half of parse failures on unseen test data were caused by inaccurate lexical information, and Baldwin et al. (2004) identified that in parsing 20,000 strings from British National Corpus (BCN) a 40% of

grammar failures were due to missing lexical entries, with a grammar dictionary of about 10,500 lexical entries that time.

Our work proposes two innovative ideas: first, to compensate the lack of occurrences resorting to linguistic knowledge on the properties of lexical classes. The use of knowledge on lexical classes is not new, but our contribution is to propose a way to use it probabilistically extracting it directly from a lexical typology, that is, without deriving the probabilities from a sample of data. To assess the probabilistic model from the data creates problems with low frequency items, not only words, but also patterns of occurrence that show a low frequency. Second, in order to handle the uncertainty of the data (some patterns of occurrence can characterize more than one class, and there is noise in the identification of the patterns), we take advantage of the formulation of lexical classes in terms of combinations of grammatical features in order to build classifiers for each of these features first, and reconstructing the classes in a second stage.

We have based our proposal on works specifically concerned with the problem of sparse data and lexical learning by Bayesian methods, in particular, Anderson (1991) and Xu and Tenenbaum (2007). We propose a Bayesian method for lexical learning where lexical knowledge is represented as a probabilistic model. A lexical typology is formulated probabilistically to be used to maximize the information gathered from even a single occurrence to predict the full set of properties of the word. In our model, given a hypothesis space (the value yes/no for each grammatical feature) and one or more examples, the system evaluates all hypotheses in order to choose the most likely value for every feature. The system does it by computing their posterior probabilities, proportional to the product of prior probabilities and likelihood. The prior probabilities are the expectation about which hypotheses are more or less plausible, independent of the observed example. The likelihood is the expectation about which examples are likely to be observed given a particular

hypothesis about a feature and value. This expectation is obtained, by means of a hybrid method, from structured knowledge, i.e. a linguistic typology of lexical classes. The final decision about a word is determined by averaging the predictions of all hypothesis weighted by their posterior probabilities and summing the predictions made by each occurrence to take eventually the one which has accumulated the maximal value.

As just said, our classifier is crucially based on the representation of lexical knowledge as a system of classes, where classes are defined in terms of particular combinations of grammatical features. Because some of these features characterise more than one class creating uncertainty, our proposal is to classify words first according to these features, leaving for a later task to map the combination of features into classes. The system assigns a positive or negative value for each feature declared in the system, and, depending on these values, the word is assigned a class. The better we assign the feature values, the better we will infer the lexical class.

For evaluating our proposal we have worked with Spanish nouns. Nouns are the part of speech which presents the largest number of low frequency items. The objective of the evaluation was to check whether the system predicts the correct properties for low frequency items. We have worked with nouns because most of them occur few times in a corpus. In our corpus of about one million words, a 16% of our test set of 289 nouns occur just once, an more than a 50% occur from one to ten times. Our results demonstrate that our approach achieves state of the art results with such low frequency nouns, while other methods, as we want to demonstrate, are unable to handle them properly.

The paper is organized as follows. In section 2, we briefly review the state of the art in methods for lexical acquisition in order to motivate and justify the choices of our proposal. In section 3, we introduce the fundamentals of the grammatical feature based lexical typology used for this experiment. There, classes and features used for classifying Spanish nouns are justified. We also describe in detail the patterns used as cues for classification. This information will be used to compute a probabilistic model out of the linguistically formulated information in the terms described in section 4. The way in which we use the probabilistic model to classify words according to the information obtain from occurrences in texts is described in section 5. Section 6 is for presenting the details of the evaluation, the results of the experimentation and a comparison with a similar exercise done with Decision Trees. The conclusions and future work are presented in section 7.

## 2. State of the art

The task of lexical acquisition is to assign a word certain properties according to the information gathered from its

occurrences in texts. The first problem is to filter noise: most of the techniques for lexical acquisition look for particular patterns of occurrence or cues in texts, but, as Brent's (1993) pointed out, "the cues occur in contexts that were not aimed at". This so called noise can come from different sources. Noise can be due to errors in processing the text, but there is a more systematic source of noise, which is characteristic of linguistic data, and that must be taken into account. Most of the methods for lexical acquisition, independently of the level of analysis of the input, base their decisions on counting the observation of particular co-occurrences as cues, i.e. a verb will be transitive if it is observed just before a noun phrase, as in "he saw her daughter". But in texts, continuity, that is two words occurring one after the other, does not necessarily mean a relationship. This is the traditional object of constituent analysis: to decide whether two continuous elements are directly related or not.

The first idea to distinguish noise from real cues was to discard as noise the cues that do not appear frequently enough, and it is due to Brent (1993), who used the Binomial Test Hypothesis to discriminate noise. The key point of his approach is that if a particular cue is more frequent than other ones, it must be a true property of the word to be acquired and hence it is not noise. The problem of this approach is that some pertinent cues occur too few times to be distinguished from noise. For instance, in a corpus of 3,334,563 tokens, an adjective like 'applicable' appears 440 times, and a 37% of these co-occurring with its bound preposition 'to'. In the same corpus, the adjective 'favorable' occurs 60 times, and only a 5% co-occurring with its bound preposition 'to', while 'generous' that occurs 7 times is never found with its bound preposition 'with'. Such big differences create real problems to frequency based methods.

Most of authors working with frequency criteria have tried to reduce noise using parsed texts (Briscoe and Carroll, 1997; Korhonen, 2002) or using linguistic generalizations that could offer a better distributed evidence (Chelsey and Salmon-Alt, 2006 used constituents and Preiss et al. 2007 used grammatical relations). Although they used different methods and materials, their results have in common an improvement in precision scores (percentage of properties correctly acquired of all properties acquired), between 80% and 90% depending on the authors and the part of speech, but not in recall (percentage of correct properties acquired with respect to those that should be acquired according to the test material), that in the case of the experiments with nouns by Preiss et al. (2007) drops to a 47.2%, working with more than 150 occurrences per word, but only using a frequency based threshold to discriminate noise. The system seems to fail in discriminating those with fewer occurrences from noise. Thus, the low recall scores can be interpreted as the failure of the frequency based methods to handle items with lower frequency. The problem is that,

as predicted by the Zipf's principle, linguistic data presents a distribution where there is a long tail of elements showing up very little in every level of representation<sup>1</sup>.

Another approach has tackled the problem of lexical acquisition by using distribution similarity judgements instead of pure quantitative decisions to decide about what is relevant information and what is noise. The idea behind is that linguistic classes define differences in the distribution of certain cues, i.e. the class of transitive verbs will show up in passive constructions, while the intransitive verbs will not. While most of the frequency based systems just mentioned work in a predictive way: the patterns induced from the data will show a number of different classes made of a set of lexical items, the works we are going to comment work by classifying the data gathered, i.e. the cues that distinguish classes are defined a priori and serve to indicate that an item belongs to a class. These works mostly use supervised methods with machine learning techniques. The learner is supplied with examples of the cues that linguistically motivate a number of proposed classes. The final exercise is to confirm that the data characterized by the linguistically motivated cues support indeed the division into the proposed classes. This is the approach taken by Merlo and Stevenson (2001), who selected very specific cues ad-hoc for classifying verbs into a number of Levin (1993) based verbal classes. Other authors have tried to use more general features, such as the *pos* tags of neighboring words (Baldwin, 2005), or general linguistic information as Joanis et al. (2007) who used the frequency of filled syntactic positions or slots, tense and voice of occurring verbs, etc., to describe the whole systems of English verbal classes.

The results of these systems based on predefined classes and cues show a better treatment of low frequency items, or at least not a significant difference attributable to differences in the number of occurrences. Merlo and Stevenson (2001) and Joanis et al. (2007) demonstrate that their results are not affected by differences in frequency. They achieved an accuracy, i.e. the number of correct classifications among all the classifications, around a 70%. These results further support the idea that for a proper treatment of low frequency items, we need to base the decisions in information that is independent of its occurrence in a collection of data, because as Korhonen (2002) demonstrated, the probability of a property observed in a corpus is very different to the conditional probability of this property given a particular word. As we explain in section 3, we decided to produce a probabilistic version of the knowledge embodied in a lexical typology by means of conditional probabilities. This information obtained from the symbolic knowledge is not biased by the zipfian distribution, and should overcome the problem

---

<sup>1</sup> Works such as Chelsey and Salmon-Alt (2006) and Preiss et al. (2007) confirm that Zipf's distribution also characterises different levels of linguistic abstraction: constituents, grammatical relations, etc.

of sparse data, as it happened when Korhonen (2002) used probabilistic information derived from WordNet classified verbs for smoothing probabilities obtained from data and achieved an improvement in recall scores, from 51.8% to 71.2% for English verbs.

The works by Merlo and Stevenson (2001) and Joanis et al. (2007) identified another important aspect that we have introduced in our proposal. In the distribution of cues per lexical class, these authors found that there are classes that share cues. In addition to the uncertainty of deciding whether a cue is noise or not, there is uncertainty in the selection of cues that describe a class. This observation is very much in line with current analysis of lexical classes in terms of combinations of grammatical features, as we will see in the next section. We have addressed this uncertainty by breaking down the classification process. We propose to first classify words into having or not having every grammatical feature. Dorr and Jones (1996) leads us to think that if the features are properly identified, the mapping to classes would be trivial.

### 3. Classes, features and linguistic cues

According to the linguistic tradition, words that can be inserted in the same contexts are said to belong to the same class. Thus, lexical classes are linguistic generalizations drawn from the characteristics of the contexts where sets of words tend to appear. Lexical acquisition can be approached as a classification of the contexts where words occur as those that characterize a particular class. As said before, some contexts can characterize more than one class, as if lexical classes were defined in terms of orthogonal patterns of properties, those that in several linguistic theories are known as grammatical features.

For the research we present here, we have taken the lexicon of a HPSG-based grammar developed in the LKB platform (Copestake, 2002) for Spanish (Marimon et al. 2007a and 2007b), similarly to the work of Baldwin (2005). In the LKB grammatical framework, lexical types are defined as a combination of properties in terms of grammatical features. The lexical typology for nouns, for instance, can be seen as a cross-classification, comprising noun countability vs. mass distinctions, and subcategorization information also expressed in terms of grammatical features.

A classifier was built for each of the features that form the cross-classified types. For nouns, mass and countable, on the one hand, and, on the other hand, for subcategorization information three further basic features: *trans*, for nouns with thematic complements introduced by the preposition *de*, *intrans*, when the noun has no complements; and *pcomp* for nouns having complements introduced by a bound preposition. The complete type can then be recomposed with the assigned features. "Temor"

(fear) and “adicción” (adiction) will be examples of *countable*, *trans* and *pcomp\_a*. The combination of features assigned will be the final type which is a definition of the complete behaviour of the noun with respect, for instance, optional complements.

As linguistic cues for identifying these features, we have used 23 patterns or contexts that can be indicative. We have followed the methodology of works such as Merlo and Stevenson (2001) and Baldwin and Bond (2003) that based their classifiers in a linguistically motivated cue selection process. The contexts where different types of nouns are expected to occur are less clearly defined than the ones for verbs, in which most of the authors have worked. Linguistic cues, that is, contexts that were taken as the expected syntactic behavior of nouns given a particular feature, are the following.

The most frequent cue that can be related to the feature *countable* is plural morphology. We mention frequency because, although some more discriminative contexts can be identified, they will not be very frequent, and thus are not useful. It can happen that a context is clearly a sign of having a particular feature, but if not very frequent, it will not be found, and the system will have no information to decide. Thus, more frequent cues, although less conclusive, can be a better choice than very informative but scarce ones. As for *mass*, the used cues were to be the head of a noun phrase without determiner occurring immediately after a verb, and the co-occurrence of the noun in singular with certain quantifiers<sup>2</sup>. Nevertheless, we should mention that mass nouns in Spanish can also appear in the contexts of countable ones, as in the case of “cerveza” (beer) when in constructions such as “tres cervezas, por favor” (three beers, please), and it is reflected in the typology.

To find frequent enough discriminative contexts that could be described at the level of morphosyntactic tags for identifying a noun occurring with its complements was harder and deserved some feature analysis work. For the feature *trans*, we first introduced nominalization suffixes such as “-ción”, “-sión” and “-miento”, that were, however, not enough. We find out that the results improved when checking the presence of the determiner of the potential complement. More complements were found to be determined, as in “aceleración de la economía” (‘acceleration of the economy’), than not, while modifiers tend to be non determined, that is zero determined, as in “mesa de juego” (‘table of games’). We have also used as a cue for transitive nouns the presence of two PPs introduced by the preposition *de* (‘of’) as in “la colección de coches de mi hermano” (‘the collection of cars of my brother’). Finally, to find the bound preposition

<sup>2</sup> The quantifiers are: *más* (‘more’), *menos* (‘less’) and *bastante* (‘enough’). But these cues, based on a collection of lexical items, are less productive than other characteristics such as morphological number or presence of determiners, as they appear very scarcely in texts.

of complements, we used a pattern for each possible preposition found after the noun in question.

#### 4. A probabilistic version of a lexical typology

These five features for characterizing nouns we have introduced in the previous section account for eleven types, as shown in Table 1, which conform the typology that we used as the base for the computation of the probabilistic model.

TYPE / SF	mass	count	intrans	trans	prep
<b>n_int_mass</b> e.g. acidez (‘acidity’)	yes	no	yes	no	no
<b>n_int_mass_count</b> e.g. aceite (‘oil’)	yes	yes	yes	no	no
<b>n_int_count</b> e.g. mesa (‘table’)	no	yes	yes	no	no
<b>n_trans_count</b> e.g. traductor (‘translator’)	no	yes	no	yes	no
<b>n_trans_mass</b> e.g. abaratamiento (‘cheapening’)	yes	yes	no	yes	no
<b>n_ppde2_count</b> e.g. colección (‘collection’)	no	yes	yes	yes	no
<b>n_ppde2_mass</b> e.g. aceleración (‘acceleration’)	yes	yes	yes	yes	no
<b>n_ppde_pcomp_count</b> e.g. aproximación de X a W (‘approach of X to W’)	no	yes	no	yes	yes
<b>n_ppde_pcomp_mass</b> e.g. temor de X a W (‘fear’)	yes	yes	no	yes	yes
<b>n_ppcomp_mass</b> e.g. ‘espacio entre’ (‘space between’)	yes	yes	no	no	yes
<b>n_ppcomp_count</b> e.g. ‘comerciante en’ (‘trader’)	no	yes	no	no	yes

Table 1. The typology of Spanish nouns.  
A simplified version

If we assume that linguistic cues found in texts are evidence that a noun has a particular feature, we can predict in which contexts the nouns having a certain feature will be likely to occur. Looking at Table 1, we see that we can not only predict the contexts directly related to a feature, but we can also predict the contexts or cues that just coincide when belonging to a particular class. Hence, we can compute the probability of appearing in different contexts, all those that are cues of all the features that conform the classes in the typology. For instance, in Table 1 we can predict that it is more likely that a noun having the feature *transitive*, occurs with an linguistic cue of the feature *countable* (plural, for instance) than with a linguistic cue of the feature *mass* (absent determiner, for instance), as there are 6/6 cases for *trans=yes* and *count=yes*, and only 3/6 cases for *trans=yes* and *mass=yes*. As we will see in the next section, we will take this information also for gathering information from

every occurrence for all the classification exercise about having every feature or not. Therefore, the linguistic classes can provide us with likelihood information to be used as a substitute of the computations made by observing the data directly, which is what a supervised machine learning method does.

Furthermore, our method to calculate the likelihood has been tuned to take into account certain known characteristics of linguistic data. First, some cues are just optional contexts for a word. For instance, a word that has a feature such as “bound preposition” can appear with it, but it is not obligatory. Second, the low level tool (Regular Expressions based on lemmas and part of speech tags) used to find the cues is limited, and, following with the same example, it will not find the preposition that heads the complement if it is not almost immediately after the word in question. In order to tune the correlations between cues (*LC*) and features (*SF*), that is:  $P(LC|SF)$ , we have used a function that lowers the likelihood: a word that has a particular syntactic feature is expected to be found in a particular context (a particular *lc*), but, as said before, we can have missed it. Our function assigns to each *yes* in Table 1 a *yes|no* value, in order the likelihood to take into account the possibility of having missed the cue. In other words, a word having a particular feature should be observed in a particular context, but in case it is not observed, the hypothesis is still valid.

## 5. Assigning features to words

In what follows, we present how the probabilistic information mentioned in previous section is used when observing the occurrences of a word in texts, and the computation of how much these contexts amount for assigning a particular feature.

For each syntactic feature  $\{sf_1, sf_2, \dots, sf_n\}$  of the set *SF* represented in the lexical typology of reference, we define the goal of our system to be the assignment of a value,  $\{no, yes\}$ , according to the result of a function  $Z: \sigma \rightarrow SF$ , where  $\sigma$  is a word’s signature, the set of its occurrences in a given corpus. The decision on value assignment is achieved by considering every occurrence as an accumulation of evidence in favor or against having every particular syntactic feature. Thus, our function  $Z'(SF, \sigma)$ , shown in (1), given every syntactic feature and value of *SF*,  $sf_{i,x}$ , and a particular word signature  $\sigma$  containing  $z$  different vectors,  $\sigma = \{v_1, v_2, \dots, v_z\}$ , will sum the information coming from all the vectors with respect to  $sf_{i,x}$ .

$$(1) Z'(sf_{i,x}, \sigma) = \sum_j P(sf_{i,x} | v_j)$$

In order to assess  $P(sf_{i,x}|v_j)$ , we use (2). It is the application of Bayes Rule for solving the estimation of the probability of a vector conditioned to a particular feature and value.

$$(2) P(sf_{i,x} | v_j) = \frac{P(v_j | sf_{i,x})P(sf_{i,x})}{\sum_k P(v_j | sf_{i,k})P(sf_{i,k})}$$

For solving (2), we have assumed that the prior  $P(sf_{i,x})$  is computed on the basis of the typology too, assuming that the feature that is more frequent in the Table 1 will correspondingly be more frequent in the data.

For computing the likelihood  $P(v_j|sf_{i,x})$ , as each vector is made of  $m$  components: the linguistic cues  $v_z = \{lc_1, lc_2, \dots, lc_m\}$ , we proceed as in (3) on the basis of  $P(lc_l|sf_{i,x})$ , data that we have assessed, as explained in section 4, out of the lexical typology, for every *lc*.

$$(3) P(v_j | sf_{i,x}) = \prod_{l=1}^m P(lc_l | sf_{i,x})$$

Finally,  $Z$  as in (4) is the function that assigns the feature values to signatures, what is done in a higher scoring basis. In the theoretical case of having the same probability for *yes* and for *no*,  $Z$  is undefined.

(4)

$$Z = \left\{ \begin{array}{l} Z'(sf_{i,x} = yes | \sigma) > Z'(sf_{i,x} = no | \sigma) \rightarrow yes \\ Z'(sf_{i,x} = no | \sigma) > Z'(sf_{i,x} = yes | \sigma) \rightarrow no \end{array} \right\}$$

## 6. Evaluation

### 6.1 Methodology and Data

We have worked with a part of speech tagged corpus (*Corpus Tènic de l’IULA*) which consists of domain specific texts. The section used for our evaluation was of 1,091,314 words of texts in the domain of economy.

We evaluated by comparing with Gold-standard files that we got from the manually encoded lexica of the SRG grammar. The usual accuracy measures as *type precision* (percentage of feature values correctly assigned to all values assigned) and *type recall* (percentage of correct feature values found in the gold-standard) have been used. F1 is the usual score combining precision and recall. Note that ambiguity has not been treated at all, being the evaluation against a unique correct feature, and we have tried to get rid of very ambiguous nouns. The baseline algorithm used has been a simple majority-class classifier, as computed from the gold-standard files that assigns the most frequent value for each syntactic feature.

Due to the difficulties in comparing our approach to other works in the domain, we have used for evaluation our own work on lexical acquisition with the same materials but using a C4.5 Decision Tree (DT) classifier (Quinlan 1993)





