Towards a New Approach to the Study of Neology*

Teresa Cabré Rogelio Nazar

Institute of Applied Linguistics
Pompeu Fabra University
Roc Boronat 138
08018, Barcelona, Spain

Abstract

This paper describes a series of methods for the analysis of new words from the perspective of quantitative linguistics and reports on different experiments on formal neology (both monolexical and polylexical units) as well as semantic neology. Our attempt to extract formal neology is based on the analysis of diachronic corpora, where we compare the relative frequency distribution of words in the corpus with an exponential function which, according to our hypothesis, represents the typical behavior of a neologism. The analysis of semantic neology is also based on diachronic corpora, the difference being that the approach does not take into account the diachronic frequency distribution of the analyzed unit *per se* but rather its co-occurrence with other words, as we interpret that a change in the units that tend to co-occur with a word *a* is indicative of the creation of a new sense in *a*.

1 Introduction

The study of new words has been an active area of research in lexicography and linguistics for several decades and yet the notion of neology continues to be problematic. The meaning of the term *neologism* is a matter of great complexity because the condition of neology is relative at different levels. For a word to be considered new, it has to be perceived as such by the broad mass of speakers of a language, though many words can fulfill that criterion and still not be considered neologistic. A person might not know a particular lexical unit for cultural, sociological or geographical factors. Conditions such as education and age are key to predict if a person will be acquainted with a given vocabulary unit, but also historical

[*] This paper is a summary of the talk given in the 4th Joint Seminar on Neology and Specialized Translation organized by the CVC and Termisti, Institut suprieur de traducteurs et interprètes, Haute École de Bruxelles April 29, 2011, and is a short version of a text that will appear in *Neologica*. This work has been possible thanks to fundings from project APLE (Spanish Ministry of Science and Education: Ref. FFI2009-12188-C05-01, subprograma FILO) "Updating processes of the Spanish lexicon from the press" Period: 2010-2012. Project Leader: Dr. M. Teresa Cabré.

reasons can determine that members of a particular culture are more likely to be familiar with a particular concept. Words that have a specialized meaning are a subset of this class, and it is mostly fortuitous how specialized terms will evolve over time. It might occur that a term continues to be used in a reduced circle of specialists, it might also happen that it gains popularity among wider groups of people or might as well be absolutely forgotten. All of these problems make the scientific research in this particular field a real challenge, mainly because it is not clear what an objective method would be to measure if a given word is or is not a neologism and, if it is, in what degree.

The research team IULAterm has devoted more than two decades to research in the field of neology and started the Observatory of Neology (OBNEO)[1] in an effort that included a collection in a database of thousands of specimens found in the media in Spanish and Catalan and later, in a collaborative enterprise with other institutions[2], of several other Romance languages. During the first years, the task of collection was undertaken manually from printed newspapers and magazines as well as radio and television broadcasts. All the occurrences of these new words were organized in a database according to a classification that evolved to adapt to new data and became progressively more finely grained (see [1] for an up-to-date description). With the advent of the World Wide Web, part of this process could be automatized by downloading the vocabulary from online newspapers filtered with a lemma list derived from a lexicographic reference corpus [7]. The result of this procedure is a list of candidates that has to be manually checked in order to eliminate those words that, although not in the dictionaries, cannot be considered neologisms (i.e. referring expressions of various types, such as proper nouns, technical terminology, etc.). This procedure, known as the lexicographic criterion, is of course not entirely satisfactory because, aside from the problem of false positives, it is focused on formal neologisms only, leaving other phenomena such as syntactic, syntagmatic and semantic neology out of consideration.

In this paper we explore new procedures for the analysis of neology that include the latter types of phenomena, and we replace categoric rules such as to be or not to be listed in a dictionary by another view in which decisions are based on probabilities and frequency distributions. In this view, we disregard all lexicographic material in favor of a diachronic corpus as the only source of information. In the case of formal neologisms, candidates are analyzed according to the evolution of their frequency over time. In the case of semantic neologisms, by contrast, we look for changes in the co-occurrence behavior of the unit since, as we will show later, a new development in the meaning of a lexical unit can be measured

[1] The Observatory of Neology was created in 1989 at the University of Barcelona by M.Teresa Cabré, and in 1993 moved to the Pompeu Fabra University. The OBNEO data bank, BOBNEO, currently contains more than 174.000 records.

[2] Through OBNEO, the group IULATERM coordinates a network with other observatories from seven Iberoamerican universities, gathered together in 2002 for the project "Antenas Neológicas" (http://www.iula.upf.edu/obneo/obprojca.htm) with the purpose of undertaking contrastive analyses of the different varieties of European and American Spanish. Another network is NEOROC, established in 2003 by a group of seven observatories from Spain and devoted to the analysis of the different varieties of Spanish within Spain (http://www.iula.upf.edu/rec/neoroc/). In 2004, the network NEOROM was created by another group of observatories from America and Europe with the goal of studying neology in all Romance languages (http://obneo.iula.upf.edu/bneorom/index.php). Finally, in 2007 the NEOXOC network was created by eight universities with the purpose of studying neology from the different varieties of Catalan (http://www.iula.upf.edu/rec/neoxoc/)

by the changes in “the company it keeps”. The paper therefore tries to summarize a series of lines of research started elsewhere [6].

2 Methods

2.1 Monolexical and Polylexical Neologisms

Monolexical and polylexical neologisms are both considered formal neologisms. In general, and for practical reasons, computational linguists define lexical units as orthographic words, distinguishing thus between single words (monolexical units, i.e. separated by blank spaces or punctuation signs) and larger units consisting of combinations of these single words (polylexical units or multi-word expressions). The difference is, naturally, not as trivial as it may seem from this perspective. However, to keep this distinction is natural and in fact important when applying the lexicographic criterion, because polylexical –or syntagmatic– neology detection was seen as adding a second order of complexity or simply as unfeasible. There were two reasons to discourage any attempt in that direction: the first one is that the number of different combinations of words grows exponentially as one considers larger sequences; and the second is that entries in general dictionaries consist of single words, while multi-word expressions are treated as subentries. With these elements at hand, it is difficult to detect neologisms such as *cell phone* or *land line*, combinations of words that do exist individually as entries in the exclusion dictionaries.

When one leaves aside the dictionaries, however, new possibilities emerge for the analysis. Firstly, there is no theoretical reason to treat the combinations of words as something inherently different from single words. In simple terms, sequences of orthographic words (of say, two to five positions) are considered “long words”. Of course there is the fact that the number of different words grow exponentially, but this, in any case, is not a theoretical issue but a computational one. Furthermore, when one applies statistical analysis to data, the panorama changes completely, because the number of combinations of words can be considerably reduced by keeping only those combinations that are strongly associated, as will be explained in the following paragraphs.

For clarity of exposition, let us begin by exemplifying the analysis with single words, but the method is exactly the same in the case of syntagmatic expressions. As already mentioned in the introduction, our analysis takes a time-sliced corpus as the only source of information and is based on the hypothesis that, ideally, the relative frequency of a neologism describes an exponential curve such as the one shown in Figure 1 for an arbitrary period (1975-2006), which is simply an exponential function (Equation 1).

$$f(x) = x^a \tag{1}$$

The point we are trying to make is that neologisms are not to be seen as words that are used for the first time, but words whose frequency of occurrence describes an abrupt rise from zero or almost zero to a sharp peak, and this general principle concerns both monolexical and polylexical units. Let us show some examples (Figure 2) which represent words whose frequency curve show a general resemblance with the ideal neologism depicted in Figure 1. These words are *spam*, *blog* and *sms*, which could be considered neologistic in

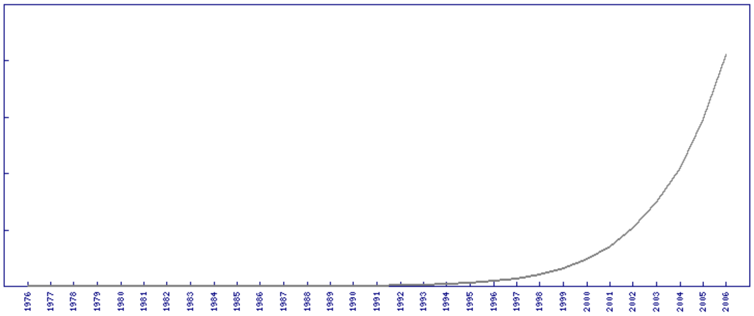

Figure 1: The relative frequency curve of an ideal neologism in a time-sliced corpus.

Spanish in 2007, which is the last year of the data used in [6], a collection of press articles from the Spanish online newspaper El País[3] from 1976 to 2007. These words were obtained by calculating the similarity between the frequency curves of the ideal neologism defined in Equation 1 and the frequency curve of all the words that appeared in the last year of the collection within a given frequency interval (e.g., words that appeared between 40 and 100 times in 2007). The similarity between curves, as geometric objects, is calculated using the Euclidean distance (Equation 2) where X and Y are the two curves and n the number of points (in this case, years) to be compared. In order to compare two curves, they first have to be normalized (Equation 3), a necessary procedure to analyze units of different frequency.

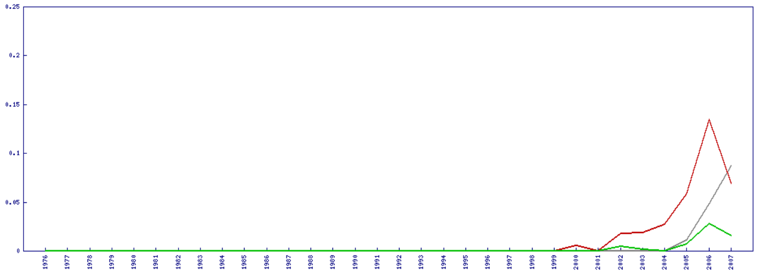

Figure 2: Frequency curve of the words *spam*, *blog* and *sms* in El País between 1976 and 2007.

$$E(X,Y) = \sqrt{\sum_{i=0}^{n}(X_i - Y_i)^2} \tag{2}$$

$$x'_i = \frac{x_i}{max(X)}) \tag{3}$$

With respect to syntagmatic neologisms, as we said earlier, they receive the same treatment as the single words. We define the vocabulary analyzed as any sequence of up to five words (n-grams) in a certain frequency band in the last year of the collection, with the

[3] http://www.elpais.es

exception of those which have a member of a stoplist as the first or last word. The stoplist is defined in this case as the k most frequent words in the whole corpus (here $k = 100$). At first sight it might seem a very lax restriction of the number of possible combinations, however, the real filtering process occurs when the frequency curves are compared with that of our ideal neologism, because it is highly unlikely that a word combination will describe a frequency curve like Figure 1 due to pure chance, as it is the case with *teléfono fijo* ('land line') and *teléfono móvil* ('mobile phone'), both depicted in Figure 3. In most cases, noise in the candidate lists is not due to random error but consists rather of proper nouns of famous people and multi-word referring expressions of various types, mostly topics that made it into the agenda setting of the media (e.g., *Avril Lavigne*, *revista Forbes*, 'Forbes magazine', *frente polisario* 'Polisario Front'). At some point, however, the distinction between genuine syntagmatic neologisms and noise becomes less clear. Should we consider that a candidate such as *climate change* (Figure 4) is appropriate to be included in a dictionary of neologisms? Is it a lexical unit or a referring expression? Certainly we cannot consider it compositional, and its meaning today is not the same as it was in 1975. Yet, lexicographers would not be willing to include such expression in a dictionary, arguing that such unit is better placed in the encyclopedia.

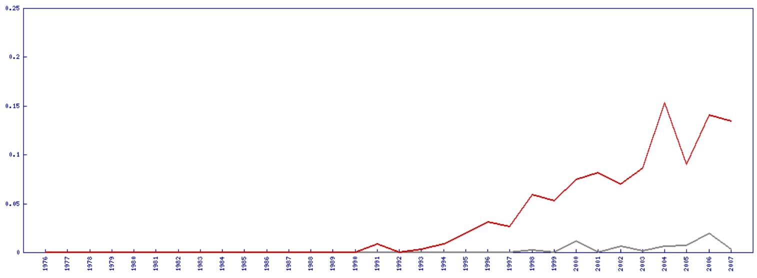

Figure 3: Frequency curves of *teléfono móvil* ('mobile phone') and *teléfono fijo* ('land line')

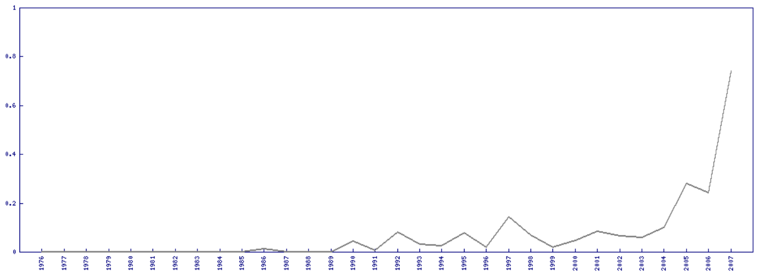

Figure 4: Frequency curve of the expression *cambio climático* ('climate change')

It should be added that in the same way that there is a group of words whose frequency curve in the time line resembles that of our ideal neologism, there are also other groups of words with different behavior which, if of interest, could also be analyzed using the same method, with the only requirement of changing the kind of ideal curve. Figure 5, for instance, shows the case of the word *spa* which, although with an undoubtedly ascending frequency, was already being used relatively often in the eighties. A different case is that of the ephemeral neologisms, those which were successful for a limited period of time but then fell into desuetude, as it apparently occurred to the word *multiétnico* ('multiethnic'), in Figure 6, and other cases of expressions that refer to specific historic events. A somewhat special case is that of *kale borroka* (Figure 7), a word that literary means 'street fight' in Basque and refers to violent actions of resistance of organized groups of Basque youth. The maximum frequency counts of the expression in the newspaper coincides with two different periods of eruption of this type of violence. With these examples we would like to express the idea that new words can have different frequency curves. In our case, however, because we were only interested in neologisms that were successful (at a given point in time), we decided to focus only on those that resemble our ideal ascending curve.

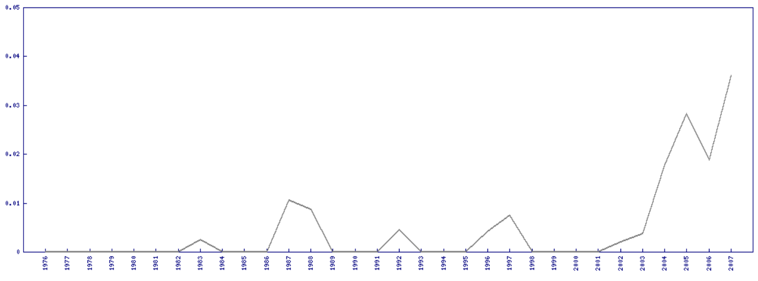

Figure 5: Frequency curve of the expression *spa*

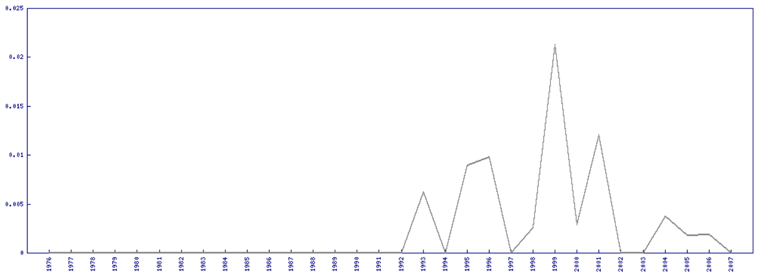

Figure 6: Frequency curve of the expression *multiétnico* ('multiethnic')

2.2 The Study of Semantic Neology

The study of the new meanings of words has something in common with the previous case of the syntagmatic expressions in which they were both defined by the exclusion list criterion. The difference with the previous is that now it cannot be possible to adopt the same method, because the creation of a new meaning is not necessarily signaled by a change in the frequency of use of a word.

The great difficulty of the detection of semantic neology from this point of view is that these units do appear as entries in the dictionaries. However, the real reason to view the task as impossible lies in an implicit notion of lexical unit as a string of characters. Of course we

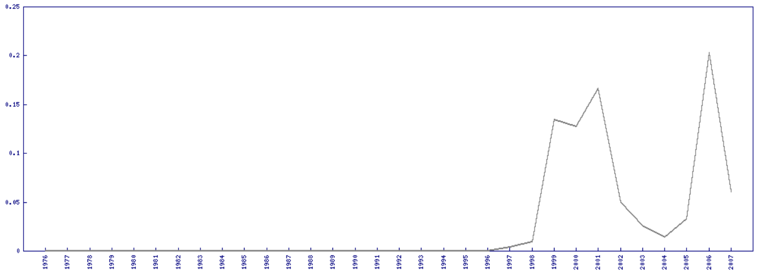

Figure 7: Frequency curve of the expression *kale borroka*

can also admit other definitions, such as a distributional one, where a lexical unit is viewed as a form that that has a tendency to co-occur (with significant frequency) with a group of other words, a notion that makes possible to distinguish different lexical units even when they are formally identical.

The group or vector of collocatives can occur at different distances from the target word (say, a context of 10 or 20 words at left and right of the target word) and not necessarily in adjacent positions. Normally, the group tends to be relatively stable over time. Thus, when we observe that there is a change in the collocatives of a given lexical unit conforming a well defined new group of words, then we have found a possible sign of the development of a new meaning.

In order to study the new meanings of words we are applying a clustering technique based on co-occurrence graphs ([5]). Initially applied for Word Sense Induction and Disambiguation, what this algorithm does is to accept a target word (single or multi-word expression) and a corpus where this unit is expected to occur (in this case, our diachronic corpus). In essence, the clustering process is based on a graph where nodes are words or word n-grams and arcs between nodes are weighted according to the frequency of their co-occurrence. The goal is to determine the significance of the frequency of co-occurrence in the same contexts of two nodes i and j from graph of the target word x. The motivation is that one of the results of this co-occurrence graphs is that they tend to show different hubs (regions of highly interconnected nodes) in case x is a polysemous word. The exact graph clustering procedure is shown as pseudocode in Algorithm 1, where j is every node in the graph of the target word t and k is an arbitrary threshold which refers to the number of documents that two clusters have to share in order to be considered similar (in our experiments, $k = 33\%$).

In the example shown in Figure 8, also mentioned in ([6]), we have the expression *palabra de honor* which in the corpus analyzed can have the meaning of 'word of honor' or 'strapless gown', depending on the context. The second use could be considered neologistic, at least in the vocabulary of the press. The frequency of use of the expression, however, does not indicate a change. It is by applying the clustering technique that we can separate both kinds of contexts of occurrence, as shown in the two subgraphs or clusters depicted in Figures 9 and 10. In each of these two subgraphs we can see different lexical units that accompany this expression in each of its two senses. One of them is populated with proper nouns of politicians and words such as *credulidad* ('credulity'), *proclamar* ('to proclaim'), *inocencia* ('innocence'), etc., while the other contains fashion related vocabulary such as *Gucci*, *Swarosky*, *tonos*,

Algorithm 1 Co-occurrence graph clustering algorithm

```
for all $j \in G(t)$ do
    1. create cluster with all documents where $j$ occurs
    2. next if first cluster
    3. compare new cluster with previously created ones
    4. if one has overlap $> k$, collapse both
    5. if more than one have overlap $> k$, destroy new cluster
    6. last if no more documents
end for
Pairwise comparison of clusters:
while two clusters have overlap $> k$ do
    collapse both
end while
```

('tones') and so on. What is interesting is that, as each context of occurrence is dated, we can observe that while the time span of the first subgraph is 1979-2004, the other one by contrast is 1998-2007. The fact that one of the clusters is more recent than the other is what suggests that we are confronted with a semantic neologism and not just another polysemous expression.

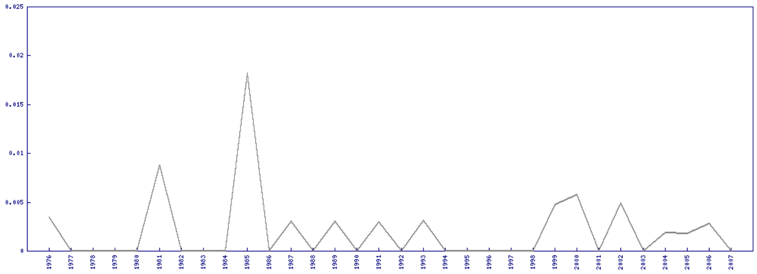

Figure 8: Frequency curve of the expression *palabra de honor*

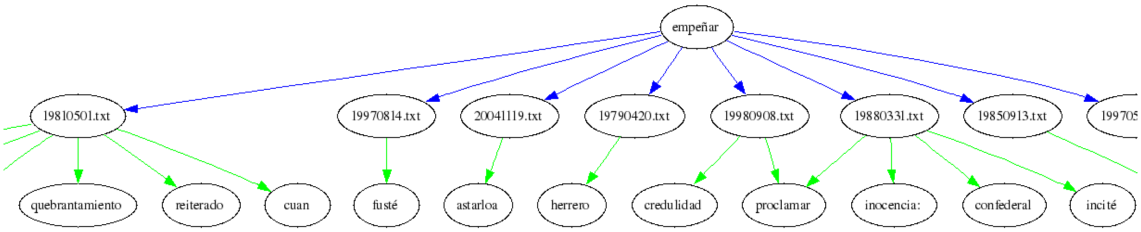

Figure 9: A subgraph from the co-occurrence graph of expression *palabra de honor* corresponding to its sense of 'word of honor'

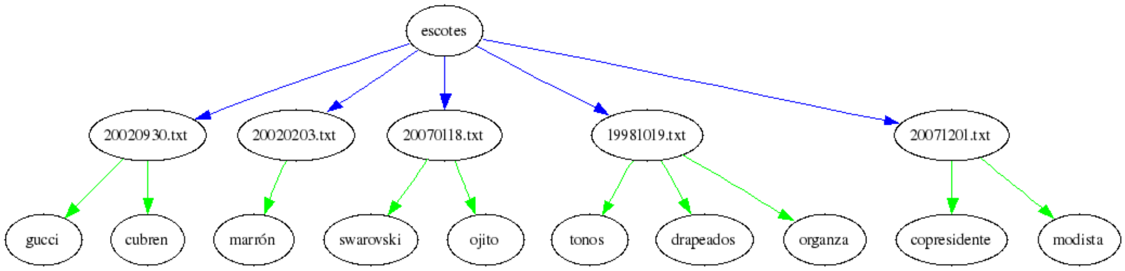

Figure 10: Another subgraph from the same co-occurrence graph of *palabra de honor*, now corresponding to the sense of 'strapless gown'

3 Conclusions and Future Work

We have presented a description of some of our new developments in the field of automatic neology detection including syntagmatic and semantic neology. Our research is in an experimental phase and we are still far from reaching conclusive results. Thus, rather than proposing a unified theory of neology, our motivation in this paper has been to explore different possibilities without losing awareness of the fact that we may adopt different methodologies in the future. In essence, however, we believe that the statistical analysis of diachronic corpora is the most natural approach. We are making slow but steady progress in different simultaneous fronts. In parallel, we are doing research in the area of syntactic neology, also known as grammatical conversion [2], as well as the extraction of semantic neology with another method that enables us to extract co-occurrence statistics from n-gram corpora instead of text [4], motivated by the publication of Google N-grams (Books) [3].

As for future work, we will compile a new corpus of recent Spanish press articles. Of course this can raise criticisms in the sense that a press corpus is biased and not representative of general language, but the same notion of general language is difficult to define, and what exactly would constitute an ideally balanced corpus is still an open question. Perhaps it could be a good idea to constitute a sort of "control" corpus parallel to the press articles, consisting of a large sample of text from different sources. In our opinion, the only feasible way to keep a massive text collection up-to-date would be to fully automatize the process, however the difficulty of such implementation would be how to obtain a reliable estimation of the date in which a given text was written. This problem does not exist when one works with a defined corpus such as a newspaper because each article has an explicit date of publication. However, when the sources are unknown, guessing the date of production of a text can be another scientific challenge.

References

[1] M.T. Cabré and R. Estopà. *Les paraules noves. Criteris per detectar i mesurar els neologismes.* Vic/Barcelona: Eumo Editorial/Universitat Pompeu Fabra, 2009.

[2] M.T. Cabré, M. Janssen, R. Nazar, and J. Vivaldi. Detección semiautomática de neologismos: los neologismos gramaticales. In *II Congreso Internacional de Neología de las Lenguas Románicas*. CINEO 2011, forthcoming.

[3] J. B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, Google Books Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden. Quantitative analysis of culture using millions of digitized books science. *Journal of Department Finance*, 331(6014):176–182, 2010.

[4] R. Nazar. Neología semántica: un enfoque desde la lingüística cuantitativa. http://www.iula.upf.edu/materials/111214nazar.pdf.

[5] R. Nazar. *A Quantitative Approach to Concept Analysis.* PhD thesis, Pompeu Fabra University, 2010.

[6] R. Nazar and V. Vidal. Aproximación cuantitativa a la neologa. In *Actes del I Congrés Internacional de Neologia de les Llengües Romàniques*, volume 6, pages 865–878. CINEO 2008, 2010.

[7] J. Vivaldi. *Sextan: prototip d'un sistema d'extracció de neologismes. In M. T. Cabr, J. Freixa, E. Sol (ed.). La neologia en el tombant de segle: I Simposi sobre Neologia, pp. 85108.* Barcelona: Observatori de Neologia, 2000.