The Living Lexicon: Methodology to set up Synchronic Dictionaries[1]

Rogelio Nazar and Jenny Azarian
Institute for Applied Linguistics, Pompeu Fabra University, Barcelona (Spain)

In this paper, we want to investigate the subset of the vocabulary of a given language or dialectal variant which is in actual use in the discourse of a linguistic community in order to set up a synchronic dictionary. The aim of this article is, thus, to develop a methodology for acquiring the nomenclature of synchronic dictionaries in a systematic way. To do this, we consider two kinds of operations: addition of entries –the birth of words, or Neology- and removal -the death of words, Desuetude, as we call it here. The methodology consists in contrasting dictionaries of a language (or dialectal variant) to find the intersection of the vocabulary, and to compare the vocabulary of the dictionaries with the vocabulary of a diachronic corpus. Such a methodology enables us to answer the following research questions: 1) what proportion of the vocabulary is shared by most dictionaries, 2) what proportion of units of each dictionary is no longer in use and 3) what proportion of the vocabulary units in use today is still not registered in the dictionaries. These three questions are central to the definition of the ideal headword. In a pilot experiment in Peninsular Spanish, we combine the study of the main dictionaries of this language variant with diachronic studies using corpus statistics on Spanish newspaper archives.

Correctness rests upon usage.
(Bloomfield 1962: 67).

1. Introduction

Changes in languages and their dialectal variants are constant and unavoidable. As Crystal (1996) said, 'linguistic changes are an intrinsic feature of language, deep-rooted in its social milieu'. The life cycle of the words which constitute the vocabulary of a linguistic community – both the birth and death of words – must be seen as a sign of the vitality of a language, since it shows the need for naming new realities, for creating new words at the expense of old ones, the need for *recycling* the vocabulary in use. Language can be considered an entity in motion driven by many intrinsic forces: creativity, changes of a social, political and economical nature, art, spiritual tendencies, and, moreover, the recent effects of cultural globalization. Linguistic change is reflected in the usage of the language, since the selection of the vocabulary is an aspect of this usage. According to Crystal, 'only by paying close attention to linguistic changes can we guarantee […] an encounter with language which is realistic, relevant and up-to date.' (Crystal 1996: 15).

From a descriptive linguistic perspective, lexicography is the process of taking a census of wordforms observed in their usage (Quemada 1987). Nevertheless, until recently, material conditions were not available for dictionaries to attempt a synchronic nomenclature, that is, a nomenclature that reflects language use in a limited period of time. Following Crystal (op. cit.), linguistic changes are in the hands of many people, and that is why they are unpredictable and difficult to grasp. Nevertheless, corpus-based studies and the availability of electronic corpora have opened the way to study of the behavior of lexical units in their real context of use and, consequently, to develop a methodology that identifies the vocabulary that is active in a given period of time.

It is far from our idea of lexicography to suggest that words that are infrequent or completely extinct should be permanently deleted from dictionaries, since dictionaries not only help read

[1] This work was possible thanks to funding from Project APLE (Spanish Ministry of Science and Education: Ref. FFI2009-12188-C05-01, subprogram FILO) 'Updating processes of the Spanish lexicon from the press', whose project leader is Dr. M. Teresa Cabré. We would also like to thank the anonymous reviewers and several friends and colleagues who contributed with valuable comments.

current texts. That said, the proportion of the vocabulary which is in actual use, with respect to the whole repository of vocabulary of a given language, is of great importance from a lexicological standpoint. Furthermore, several concrete applications of a synchronic nomenclature are foreseeable, e.g. dictionaries for special purposes, such as desk-size dictionaries or second language learner's dictionaries. Other specific applications could be in NLP tools, for instance in terminology extraction: a lexical database that reflects the set of the vocabulary that is considered 'normal' in a given language could inform a term extractor that certain items in an analyzed text are deviations with respect to standard use. Finally, from a sociological or historic-linguistic perspective, monitoring the vocabulary units that are coming into use and falling out of use, as well as the pace of that process, is also interesting, since in this process there are traces of the *Zeitgeist.*

The aim of the methodology presented in this article is thus to study the life cycle of the vocabulary by the comparison of a lexicographical corpus -*in vitro* corpus, in Cabré's words (2000) -, comprising the last editions of six of the most important dictionaries of the Spanish language, and a diachronic textual corpus -*in vivo* corpus - comprising all the editions of a Peninsular Spanish newspaper for the last three decades. Consequently, it covers the study of 'birth' and 'death' of words. The selection of the corpus is, however, independent of this methodology, that is to say that this experiment can be replicated in different languages provided that we have usage dictionaries and a diachronic corpus. It is through the particular example of the Peninsular Spanish nomenclature that we illustrate the methodology of analysis. Therefore, it follows that we do not pretend to present a finished product such as the nomenclature of a dictionary. Such finished product would imply many other decisions that are not related to the present work. Our goal here is to propose a methodology, to present a detailed description of every step for its replication and to include an account of the results. The article is organized as follows: Section 2 offers a general background about the topic of the article. Section 3 consists of a global presentation of our approach of setting-up synchronic dictionary. Section 4 presents the results of the experiments and Section 5 includes a discussion.

2. Background

2.1. The Synchronic Nomenclature: Considerations over the Life Cycle of Vocabulary Units

It can be assumed that the main part of the nomenclature of a synchronic general dictionary is composed of entries corresponding to the lexical units in use at the moment when a dictionary is published. However, a review of the literature related to the classification of dictionaries confirms that, strictly speaking, no synchronic dictionary exists (Dubois & Dubois 1971; Haensch et al. 1982; Hartmann & James 2001). The category of synchronic dictionary has been used by Landau (1984) to denote the dictionaries that represent the vocabulary used at a particular period of time, but at that time, the notion of 'synchronic' was obviously relative. The period of time between two updates of a given dictionary usually corresponds to several years, as a consequence of the traditional process of elaborating dictionaries in which the various steps of a top-down assembly-chain usually implies an important delay of time. This therefore represents a problem, a gap in the lexicographical representation of the vocabulary in use, or maybe a challenge for research.

Corpus based dictionaries have appeared for several decades now (Lara et al. 1979). However, the current increase in the use of computers, which permits to gather large amounts of text in digital format, has enormously facilitated the creation of corpora and has sparked a linguistic revolution. As Calzolari puts it, ‘Carefully constructed, large written and spoken corpora are essential sources of linguistic knowledge if we hope to provide extensive and adequate descriptions of the concrete use of the language in real text.’ (Calzolari 1996: 4). As mentioned above, language is an entity in permanent flux. Part of this change is reflected in the life cycle of the vocabulary. Neology studies the creation of new lexical units. Desuetude (Algeo 1993), as we call it here, is the opposite moment of the cycle, the removal or death of lexical units. To set-up a synchronic dictionary, it is necessary to study both phenomena. And, as it can be inferred from what has been previously explained, such a study must involve a diachronic corpus-based methodology.

2.2. Neology, Desuetude and Dictionary Updating

According to Rey, neology is ‘une activité, c’est-à dire un processus, un dynamisme, quelque chose qui, à l’intérieur d’un système linguistique, d’une entité culturelle ou d’un groupe social de communiquants, produit des *unités lexicales nouvelles*’[2] (quoted in Cabré 2002a: 15). Neology has been studied by many authors and it can be considered a consolidated field. The related literature offers different criteria to detect neologisms and suggests different ways to automatically extract them from texts.

Rey considers two ways to detect neologisms: the temporal criterion, which designates a lexical unit as neologism if it has recently appeared in the usage, and the psycholinguistic criterion, which corresponds to the subjective perception of novelty of the user of the lexical unit. Cabré (2002b) lay down two different criteria: the lexicographical criterion – a lexical unit is considered to be a neologism if it does not appear in a lexicographical corpus of reference – and the cognitive criterion. Janssen (2009) offers an overview of the different kinds of automatic neology extraction methods that have been developed during the last decades. He distinguishes three main types of tools: (1) those that work by way of exclusion based on a list of known words (for instance a dictionary); (2) those that work by looking for patterns in the texts that are characteristic for neologisms; (3) those that work by using statistical analysis of texts, typically, the comparison of distributional behaviour of words between new texts and older texts. A variant of this third method would consist in dividing the reference corpus in temporal slices in order to observe the evolution of a lexical unit across the time line (Nazar & Vidal, forthcoming). As we will justify in Section 3, we will use this last kind of algorithm.

Desuetude, or the study of ‘disappearance of old words or the discontinued use of new ones’ (Algeo 1993), has long been neglected. The fact that there are still many different names for this is indicative of its epistemological instability: *lexical death* (Grzega 2002), *necrology* (Dury & Drouin, forthcoming), *obsolescence* (Dury & Picton, forthcoming), among others. This probably is a consequence of the fact that the empirical study of desuetude is technically more complex than neology, since it is easier to measure the presence of a new word than the progressive disappearance of others. One methodology has been proposed by Algeo (1993), who took the 3 565 new entries of two English dictionaries edited between 1944 and 1976 to find that as many as 58% of them were not recorded in the dictionaries a generation later. Commenting on these results, Algeo said: ‘successful coinages are the exception; unsuccessful ones the rule.’ (Algeo 1993: 281).

[2] ‘...an activity, that is to say, a process, a dynamism, anything that, in the interior of a linguistic system, of a cultural entity or a social group engaged in communication, produces new lexical units’.

Both phenomena, neology and desuetude, are necessary aspects of the process of dictionary updating. Computational linguists have long been interested in the study of dictionaries with the intention of updating their nomenclature on the basis of corpus statistics (Heid et al. 2000; Evert et al. 2004; Heid et al. 2004). The rationale behind the strategy proposed by these authors is to perform both addition and removal of entries by observing the frequency of such units in a corpus, among other heuristics to avoid elements such as proper nouns being suggested as entries. The representativeness of the corpus is, thus, a critical aspect and a problem that has been regarded as crucial in corpus linguistics. Biber (1994), for instance, has proposed methods to quantify the diversity of the vocabulary of a corpus as a measure of lexical representativeness. Other authors (Juilland & Chang Rodríguez 1964; Geyker 2004) have emphasized the importance of a careful selection of the materials to constitute the corpus. A balanced corpus must include different textual genres, such as press, but also prose, scientific and technical literature, etc. This of course depends on the purpose of the study. In our case we would not be interested in technical literature since it is not representative of normal use of the vocabulary. Other authors are increasingly using the web as a corpus (Kilgarriff 2004). In the case of our study, we did not used the web because it does not reflect the normative standard use as it is not as carefully edited as printed word, thus errors and other types of non normative use are very frequent.

2.3. The Ideal Headword

The technical challenges that the updating of dictionaries may raise are not as complex as the theoretical challenge of defining what should be the ideal headword for a general dictionary. Furthermore, the question of 'what is a word' is still a confusing issue for many linguists (Haensch et al. 1982; Trask 2004). We may find orthographic words, phonologic words, lexemes, lemmata, grammar words, words by inflection and derivation, multiword expressions, clitics, acronyms, abridged forms, etc. The most difficult problem, however, is to decide what is not a word. Proper nouns, for instance, are objects that are not interesting from a lexicographic perspective, since they have reference instead of meaning. More generally, we say they are referring expressions, i.e. expressions used to designate the infinite variety of entities of real and imaginary worlds, and thus should be included in encyclopedias rather than in dictionaries (Rey 1988). Most dictionaries will not include units such as 'Winston Churchill' or 'Eiffel Tower' as entries. However, the distinction between words and referring expressions is still far from clear. The main difficulty in determining when a word designates a particular entity is that we lack a profound understanding of what constitutes a particular entity. Many words that were born as proper nouns easily transform themselves into common nouns (e.g. *boycott*, *leotard*) and are, thus, absorbed into the lexicon. This often occurs with the name of products (e.g., *xerox*, *kleenex*, *jacuzzi*, etc.). Coseriu (1967) also pointed out the problem of the use of plural in the case of proper nouns (as in 'we have here some Picassos'), which are used not to designate a particular entity but rather a class. Other problems arise with the consideration of multiword expressions. To what extent can they be considered entries in a dictionary? Of course we cannot offer a solution to all these problems in this short paper, but we are sure that they should be taken into account in the development of a nomenclature.

3. Our Approach

The methodology that we propose for the selection of the synchronic nomenclature consists of studying a set of existent lexicographic resources plus a diachronic textual corpus which are presumably representative of the vocabulary used in a particular linguistic community. We will address the question of the selection of the corpora in Section 3.1., along with a justification of the claim of their representativeness. For now, it should be important to recall the distinction, already made in Section 1, between the selection of the material to study and the methodology itself. If the methodology is a function, the corpora to study would be the input parameters.

The study of lexicographic resources and textual corpora is intended to determine the following sets: 1) the intersecting vocabulary of different dictionaries; 2) the set of units that can be considered neologisms of the language; and 3) the set of units which, in some cases, can be considered domain specific terminology or, in other cases, units in desuetude. The inactive vocabulary, defined as the subset of vocabulary that is not shared by most dictionaries or that has a null or decreasing frequency in the corpus, should be excluded from the nomenclature, which would only include neology alongside with the units that, being present in the dictionaries and in the corpus, are considered the core vocabulary of the language.

For the study of neology and desuetude, we will rely on frequency distribution curves of the units in the diachronic corpus, using statistical algorithms that were applied in previous research for the automatic extraction of neology in general language (Nazar & Vidal, forthcoming) and for the detection of desuetude of terminology in a diachronic corpus of a scientific literature (Nazar, forthcoming). Lack of space prevents us from including other phenomena that would be nevertheless relevant to the present study, and for that reason we refer to previous work on the subject: namely, the automatic detection of semantic neologisms (Nazar & Vidal, op. cit.) and the automatic detection of referring expressions (Nazar 2009). In the first case, that is a complementary study because a synchronic nomenclature could add new entries for existent words, and in the second case, because for our purposes referring expressions are not considered vocabulary units and should not be included in a dictionary, as commented in Section 2.3. It would have also been interesting to include in the analysis the study of neology and desuetude of affixes. To take an example from Algeo (1991), the *-gate* suffix in the context of the Watergate case, entered into the language and proved to be very productive. In the case of Spanish, we could study, for instance, the desuetude of suffixes such as *-ísimo* / *-ísima*. Another arbitrary decision has been to exclude from the analysis the set of syntagmatic vocabulary units. This decision was made for the sake of simplicity rather than on theoretical grounds. There are, indeed, no technical reasons to exclude them, since they could in principle be treated as different lexical entries. For convenience, and because their study is not necessary for the purpose of this paper, we will reserve such material for future work.

3.1. Selection of the Corpus

As already explained, the corpus used in the study shown in Section 4, is divided in two categories: the lexicographical corpus and the textual corpus. The selection of the corpus is a complex decision. The main problem is, as already mentioned, to determine what can be considered a representative corpus of the language. In the case of the dictionaries for Spanish, we rely on the repository of vocabulary that prestigious lexicographers have gathered. We believe that six of the most well-known dictionaries can offer a substantial base for the vocabulary of the language. In the case of the experiment explained in Section 4, these dictionaries are: 1. DRAE (RAE 2001), 2. MOLINER (Moliner 2007), 3. VOX (VOX 2009), 4. SECO (Seco et al. 1999), 5. CLAVE (Maldonado et al. 2000) and 6. SALAMANCA (Gutiérrez et al. 1996).

With respect to the textual corpus, we used press archives of EL PAÍS, a renowned Spanish newspaper. This corpus is diachronic and composed of all articles since its first edition in 1976 until 2007, which are published on the web. It is divided in periods of one year, in such a way that one can monitor the relative frequency of any word form each year. Of course, being built from only one newspaper, it is a rather biased corpus with respect to register, genre and diatopic variations. However, we assume that its large size (approx. 245 million words) will make up for this bias.

3.2. Study of the Intersection between Dictionaries

The first stage of the study is to compare the dictionaries among themselves, in order to determine which proportion of the lemmata of each dictionary is shared by all or at least some of the rest of the dictionaries. This will show which dictionaries are more representative of the vocabulary in use; which dictionaries deviate more from the set of active vocabulary, and the subset of the vocabulary that can be considered the core vocabulary of the language, a subset that should be the intersection of all or at least most of the analyzed dictionaries. In order to estimate the intersecting vocabulary in our lexicographic corpus, we obtained random samples of 1000 entries per dictionary using a robot. To estimate the proportion of units shared by the studied dictionaries, we checked, for each word in each sample, if it is contained as an entry in the other dictionaries.

3.3. Comparison between Dictionaries and Textual Corpus

With the result of the previous study at hand, we will delve into the study of the vocabulary of the textual corpus. Thus, the second part of our methodology involves the study of the corpus taking as input both the units found in the dictionaries as well as the units found in the textual corpus. This will let us determine the following subsets: 1) the subset of the lemmata from the dictionaries that is not used in that corpus (i.e., the inactive vocabulary); 2) the subset of vocabulary units that are beginning to appear in the most recent part of the corpus and are still not registered in the dictionaries (i.e., neology) and 3) the subset of the vocabulary units that appear both in the dictionaries and in the corpus but are being progressively less used (i.e., desuetude). The comparison of the dictionaries with the textual corpus consisted in determining the proportion of words in the dictionaries present at least twice in the corpus. The procedure is to take samples of words from the dictionaries and to observe if these words appear in the corpus. Several samples per dictionary were taken in order to compare the percentages obtained for each sample and thus have a better assessment of what might be the real percentage.

3.4. Identification of Neology and Desuetude

As already stated in 3, from the set of words that appears in the texts of the corpus, the subset considered neological corresponds to the set of words that begin to appear in the texts from a recent period and are still not included in the dictionaries. In contrast, the subset of lexical units in desuetude consists of those units that are not present in the corpus or, if they are present, their frequency is decreasing progressively. In both cases of neology and desuetude, we first present what would be the curve of their ideal frequency distribution, and then we cast the classification of words purely as a geometric problem: we compare the observed frequency distribution of each word with each of these two models of neologisms and units in desuetude. Figures 1 and 2 show two opposite cases –ideal or extreme cases– of neology and desuetude. The horizontal axis presents the years of the newspaper archive in chronological order while the vertical axis presents the relative frequency of use of each vocabulary unit. Frequencies are expressed as relative to each year to make up for the differences in size between the years of the corpus. In Figure 1, we imagine a word that appears in the corpus in the nineties and increases its frequency in such a way that later it appears as being completely incorporated into the language. In the opposite case, Figure 2, we imagine a word being regularly used in the first years of the collection but then falling into progressive disuse. Notice that, contrary to Figure 1, in this case we do not necessarily expect that the frequency of the units in desuetude at some point will drop to zero. The curve tends to zero, but the fact that a word is effectively used in the corpus later would not contradict our expectations.

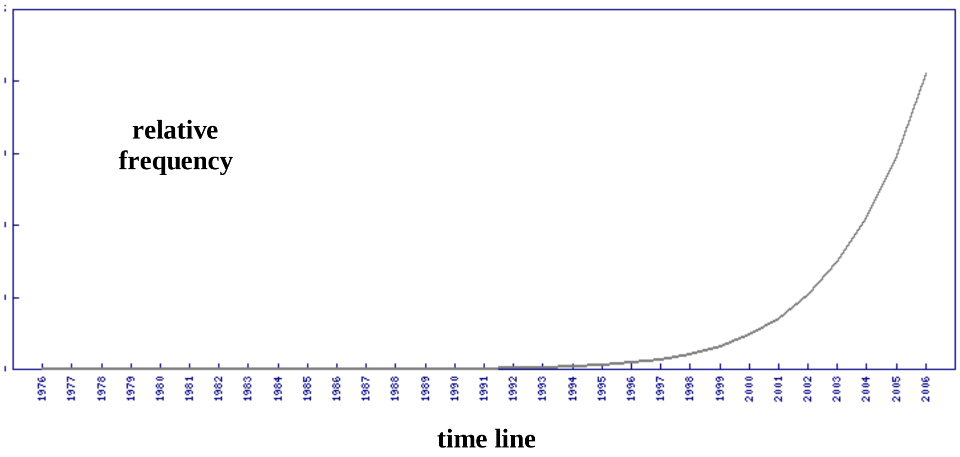

Fig. 1. The case of neology. Profile of the curve of frequencies of an ideal neologism

(1) $f(x) = x^{10}$

(2) $f(x) = x^{-1.5}$

$$\sqrt{\sum_{j=1}^{n} (X_j - Y_j)^2}$$

(3)

$$t'_{i,j} = \frac{t_{i,j}}{max(t_i)}$$

(4)

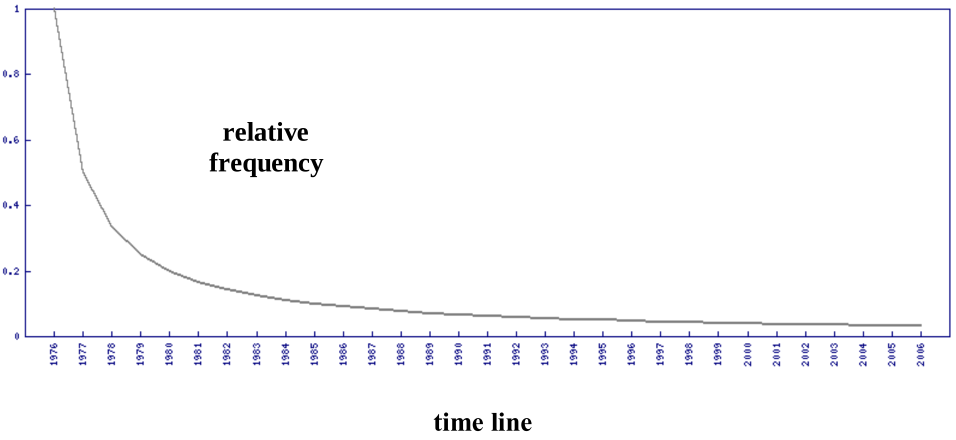

Fig. 2. The case of desuetude. Profile of the curve of frequencies of an ideal unit in desuetude

The way to extract neologisms from the corpus is, thus, to define what would be the *ideal neologism*, which is a unit that has an exponential increasing frequency with time (Formula 1). In the opposite case, from the set of words that do appear both in the dictionaries and in the texts of the corpus, we identify the subset of units in desuetude again by a curve of the *ideal unit in desuetude*, which represents the words that have a decreasing frequency over time (Formula 2). To calculate the similarity of two curves we use the Euclidean distance (Formula 3). In that way we retrieve the words which have a curve of frequency distribution similar to that of the ideal neologism or the ideal unit in desuetude. Before we compare two curves, however, we first have to normalize them (Formula 4), which is to put in the same scale the curves of words that have different frequency.

4. Results

4.1. Intersection between Dictionaries

The results of the first stage of the experiment are expressed in Table 1. There, we can see each dictionary as a row with its corresponding estimated number of entries. The columns show, for each random sample of 1000 words per dictionary, the percentage of units that also appears in the rest of the dictionaries. Thus, in the case of the first dictionary, the Salamanca, 70% of the sample also appears as an entry in all dictionaries. A high overlap may be an indication of how representative of use is the nomenclature. However, that figure is also correlated with the size of the dictionary: the bigger the nomenclature is, the more rare words it will contain.

Dictionary	Estimated size (# of entries)	Overlap with the other dictionaries (%)
Salamanca	36800	70.8
Clave	45000	53.9
Vox	53200	46.0
Seco	71400	34.9
Moliner	81500	29.2
DRAE	90000	28.0

Table 1. Comparison of the intersection of vocabulary between dictionaries from a random sample of 1000 units per dictionary

The intersection of the vocabulary from the dictionaries could serve as a basis for a synchronic nomenclature, assuming that the fact that a unit is present in many dictionaries is a strong indication that such unit is common in the language. It is not, as already mentioned, the only clue we take into account, since we also contrast the frequency of the units in the textual corpus. Thus, in this stage we would obtain a first list that would be filtered with complementary methods afterwards.

4.2 Comparison between Dictionaries and Textual Corpus

The second study is to determine, again from random samples from each dictionary, the proportion of these samples occurring in the textual corpus. The results of this second study are presented in Table 2, where we can again see that some of the dictionaries are more representative of the vocabulary of the newspaper than others. Hence, in a first random sample of 100 entries from Salamanca, we observe that 41% of it occurs in the corpus. In a second sample from the same dictionary, we observe that the percentage this time is 70%, and so on.

Dictionary	1st sample %	2nd sample %	3rd sample %	4th sample %	Average %
Salamanca	41	70	72	77	64.75
Clave	62	64	65	67	64.5
Vox	40	67	62	57	56.5
Seco	48	46	56	51	50.25
Moliner	53	42	44	45	46
DRAE	36	42	39	36	38.25

Table 2. Proportion of units that have appeared at least once in 31 years of editions of the EL PAIS newspaper in four random samples of 100 units

With this method we were able to extract vocabulary units that are completely unfamiliar for most speakers of contemporary Spanish. To name a few examples of these ‘linguistic corpses’: *almondiguilla* (meatball), *pólice* (thumb), *coyotero* (a dog trained to chase coyotes), *dorondón* (dense and cold fog), *rasgueador* (one who strums), *obtentor* (one who obtains something), *desorillar* (to remove the borders of a tissue), *acionero* (one who makes straps for horse tack), *jineteada* (horse taming), *azotalenguas* (Galium aparine, a plant of the family Rubiaceae).

4.3. Identification of Neology and Desuetude

The mechanism described in section 3.4. allows us to obtain, from the subset of units of the dictionaries that do appear in the textual corpus, lists of those words ordered by the similarity that exists between the curve of their frequency distribution in the textual corpus and the curve of the ideal neologism, in one case, and the ideal unit in desuetude, in the other. In the first case we obtain lists of typical neologisms such as *Internet*, *sms*, *spam*, *blog¸ teléfono fijo* (land line), etc. In the second case, we obtain the opposite: units that were used in the newspaper in the past, but are now less used. A few examples are *jocundidad* (joy), *gañote* (trachea), *camba* (part of the horse harness), *deterioración* (deterioration). To a lesser extent (being less similar to the ideal of desuetude) units such as *afectuosidad* (affection), *amoralidad* (amorality), *radiotelefonía* (radiotelephony) also fall into this class. In some cases, these units have been replaced by new words. In other cases, such as words referring to technologies, tools or certain concepts, they have simply fallen into disuse. For the purposes of illustration, we also sampled some units from the textual corpus, irrespective of their appearance in the dictionaries. That would let us examine units such as *trotskismo* (trotskism) and also multiword units, such as *partidos comunistas* (Communist parties) or *movimiento democrático* (democratic movement).

5. Discussion

One of the first results of this study is the evidence of the remarkable mismatching that exists between the nomenclatures of dictionaries. Some dictionaries are more representative of use than others, as some of them explicitly state in their introductions. Some (as Seco) state that their aim is to reflect current use while others (as DRAE) have a more conservative view and deliberately want to keep units that today are considered obsolete. In any case, one would need further testing before claiming that one dictionary is more representative of language use than others. Such a test could consist in taking samples of the vocabulary used in the newspapers and observing to what extent the units in these samples are registered in each dictionary. We obtained an estimation of the intersecting vocabulary of these dictionaries, which could be a useful aid for the identification of the core vocabulary of the language. With respect to the analysis of the textual corpus, we have demonstrated the usefulness of a corpus-based approach, as this represents the *in vivo* moment of the language and a special kind of bottom-up dictionary making.

As a conclusion, we propose the use of these techniques for the discovery of neologisms and units in desuetude that probably cannot be retrieved in any other way. With respect to the possible application of our work, at this point we cannot say that it would be possible to develop a high quality nomenclature by pure automatic means, since it is obvious that intensive manual labor would still be necessary to clean up the selection that our algorithms yield. However, we are moderately enthusiastic and believe that the time and effort that can be saved with this procedure has no proportion to the almost null cost of its development and use. A computer program could suggest raw selection material for the lexicographer to accept or reject candidates as entries. We do not rule out the possibility of a further refinement of the results combining different sources of information, such as those in the cited previous work.

References

Algeo, J. (1991). *Fifty years among the New Words: a dictionary of neologisms*. New York: Cambridge University Press.

Algeo, J. (1993). 'Desuetude among New English Words'. In *International Journal of Lexicography* 1993 (6). 281-293.

Biber, D. (1994). 'Representativeness in corpus design'. In: A. Zampolli; N. Calzolari; M. Palmer (eds). *Current Issues in Computational Linguistics: In Honour of Don Walker*. Pisa: Giardini/Dordrecht: Kluwer. 377–407.

Bloomfield, L. (1962). *The Menomini language*. New Haven: Yale University Press.

Cabré, T. (2000). 'La neologia com a mesura de la vitalitat interna de les llengües'. In M. T. Cabré; J. Freixa; E. Solé (eds). *La neologia en el tombant de segle: I Simposi sobre Neologia (18 de desembre de 1998), I Seminari de Neologia (17 de febrer de 2000)*. Barcelona: IULA. 85-108.

Cabré, T. (2002a). 'La neologia efímera', In M. T. Cabré; J. Freixa; E. Solé (eds). *Lèxic i neologia*. Barcelona: IULA. 13-28.

Cabré, T. (2002b). 'La neologia, avui: el naixement d'una disciplina', In M. T. Cabré; J. Freixa; E. Solé (eds). *Lèxic i neologia. IULA Documenta*. Barcelona. 29-42.

Calzolari, N. (1996). 'Lexicon and corpus: a multi-faceted interaction'. In M. Gellerstam; J. Jarborg; S.-G. Malmgren; K. Noren; L. Rogstrom; C. R. Papmehl (eds). *Euralex '96 Proceedings*. Goteborg: Goteborg University. 3-16.

Coseriu, E. (1967) *Teoría del lenguaje y lingüística general*. Madrid: Gredos.

Crystal, D. (1996). 'Reflecting Linguistic Change'. In *The Teacher Trainer* 10. 15-16.

Dubois, J.; Dubois, C. (1971). *Introduction à la lexicographie: le dictionnaire*. Paris: Larousse.

Dury, P.; Picton, A. (forthcoming). 'Terminologie et diachronie: vers une reconciliation théorique et méthodologique ?'. In *Revue Française de Linguistique Appliquée (RFLA), vol. XIV, 2009 (2). La terminologie: orientations actuelles.*

Dury, Pascaline; Patrick Drouin (forthcoming). 'When terms disappear from a specialized lexicon: a semi-automatic investigation into necrology'. In *Proceedings of the XVII European Symposium on Languages for Specific Purposes.* Aarhus: Aarhus School of Business.

Evert, S; Heid, U.; Säuberlich, B.; Debus-Gregor, E.; Scholze-Stubenrecht, W. (2004). 'Supporting corpus-based dictionary updating'. In *Proceedings of the 11th Euralex International Congress*. Lorient. 255-264.

Geyken, A. (2009). 'Automatische Wortschatzerschließung großer Textkorpora am Beispiel des DWDS'. In *Linguistik online* 39, 2009 (3). Berlin.

Grzega, J. (2002). 'Some Aspects of Modern Diachronic Onomasiology'. In *Linguistics* 40. 1021-1045.

Gutiérrez Cuadrado, J.; Pascual Rodríguez, J. (1996). *Diccionario Santillana Universidad de Salamanca*. Madrid: Santillana.

Haensch, G.; Wolf, L.; Ettinger, S.; Werner, R. (1982). *La lexicografía*. Madrid: Gredos.

Hartmann, R. & James, G. (2001). *Dictionary of Lexicography*. London: Routledge.

Heid, U; Evert, S.; Docherty, V.; Worsch, W.; Wermke, M. (2000). 'Computational tools for semi-automatic corpus-based updating of dictionaries'. In *Euralex 2000*. Stuttgart. 183-196.

Heid, U.; Evert, S.; Säuberlich, B.; Debus-Gregor, E.; Scholze-Stubenrecht, W. (2004). 'Tools for upgrading printed dictionaries by means of corpus-based lexical acquisition'. In *Proceedings of LREC-2004*. Lisbon. 419-423.

Janssen, M. (2009). 'Detección de Neologismos: una perspectiva computacional. In *Revista Debate Terminológico* 5.

Juilland, A.; Chang-Rodríguez, E. (1964). *Frequency Dictionary of Spanish Words*. The Hague: Mouton.

Kilgarriff, A. (2004). 'The Sketch Engine'. In *Proceedings of the 11th Euralex International Congress*. Lorient. 105-116.

Landau, S. (1984). *Dictionaries. The Art and Craft of Lexicography*. Cambridge University Press.

Lara, L.; Ham Chande, R.; García Hidalgo, Mª I. (1979). 'Investigaciones lingüísticas en lexicografía'. In *México. El Colegio de México*.

Maldonado González, C; Hernández Hernández, H.; Almarza Acedo, N. (2000). *CLAVE diccionario de uso del español actual*. Madrid: EDICIONES SM.

Moliner, M. (2007). *Diccionario de uso del español*. Madrid: Gredos.

Nazar, R. (2009). 'Diferencias cuantitativas entre referencia y sentido'. In *Applied Linguistics Now: Understanding Language and Mind. La Lingüística Aplicada Hoy: Comprendiendo el Lenguaje y la Mente. Actas del XXVI Congreso de la Asociación Española de Lingüística Aplicada*. Almería: Universidad de Almería.

Nazar, R; Vidal, V. (forthcoming). 'Aproximación cuantitativa a la neología'. In *Actas del I Congreso Internacional de Neología en las lenguas románicas, CINEO 2008*. Barcelona.

Nazar, R; (forthcoming). 'Evolución de la terminología lingüística en las Actas de Congresos de AESLA entre 1983 y 2006'. In Actas *del XXVIII Congreso Internacional de AESLA*. Vigo.

Quemada B. (1987). 'Notes sur lexicographie et dictionnairique'. In *Cahiers de lexicologie* 51. 229-242.

Real Academia de la Lengua Española. (2001). *Diccionario de la Lengua Española*. Madrid: Real Academia.

Rey, A. (1988). *Encyclopédies et dictionnaires*. Paris.

Seco Reymundo, M; Andrés Puente, O; Ramos González., G. (1999). *Diccionario Manuel Seco del Español*. Madrid: Santillana.

Trask, R. L. (2004). *What is a word? Working Papers in Linguistics and English Language*. University of Sussex.

VOX (2009). *Diccionario General de la Lengua Española*. Barcelona: VOX Ed.