When terms disappear from a specialized lexicon: A semi-automatic investigation into 'necrology'

Pascaline Dury Université Lumière Lyon 2
Patrick Drouin, Université de Montréal

1 Introduction

Neology and term formation have long been studied in the field of terminology, and are well documented (see for example for the French language Guilbert 1974; Boulanger 1989; Cabré and Yzaguirre 1995; Sablayrolles 2000, 2003 or Humbley 2006). But very few studies have focused on the other end of the life cycle of terms, i.e. on 'lexical death' (Grzega 2002), or, as we call it here on *necrology.*

Our assumption is that necrology can be as interesting to study as neology is, and that it can teach us a lot, for instance on the 'terminological turnover' of a specialized lexicon, i.e. on the number of terms which become obsolete over a given period of time, and on how quickly they do so. This, in turn, can prove to be very useful information for professionals in their everyday practice, to teach specialized languages, or to update ontologies and dictionaries.

We define necrology as the disappearance of a term, the disappearance of a part of term, a change in grammatical status, and/or the disappearance of a meaning over a given period of time. Neologisms which have failed to stabilize in a lexicon, and have thus disappeared from it are also considered as *necrologisms*. We also keep in mind that a term can disappear from a specialized lexicon over a given period of time, and is therefore considered, for the period involved, as a necrologism even if it may reappear in the same lexicon at a later time period, with a new or altered meaning.

Our objectives for this research are first to identify various linguistic clues that can help detect necrologisms semi-automatically, and explore various computer tools and techniques to extract them from a corpus. Once an initial list of necrologisms has been identified, our aim is then to list the different types of term necrology, and see which ones are the most productive for the corpus studied.

With this paper, we present the initial work which has been done on semi-automatic identification of necrology, and we describe the methodology used. We are fully aware that much remains to be done in this field, and we hope that terminologists and lexicographers will feel the urge to explore this subject much further.

2 Methodology

2.1 Corpus description

Documents that make up our corpus were selected on the basis of two sets of criteria; the following sections describe the process of document selection. The first selection relies on general criteria unrelated to diachronic aspects, while the second is at the heart of methodology for necrology processing.

General criteria

The study being done within the scope of the field of terminology, we first needed to define clearly the subject field to be explored. For various reasons ranging from the past experience of authors, and the availability of subject matter experts, we decided to work in the field of *terrestrial ecology.* We will focus here, as do most terminological studies, on written language. The corpus was manually built from texts published in well-established scientific journals, and is made of 159 articles and specialized book chapters, totalling just about 679,000 words. Although the size of this first corpus is somehow small in the light of modern corpus-based studies, it is large enough to allow for significant statistical testing, and to gather quality terminological information.

Diachronic criteria

The overall time period covered by our corpus ranges from 1950 to 2005. This initial range is then subdivided into six periods covering ten years each, with the exception of the latest period, which covers only five years. Thus, the time periods are: 1951, 1960; 1961, 1970; 1971, 1980; 1981, 1990; 1991, 2000 and 2001, 2005.

As mentioned above, the last time period covers only five years, and is therefore much shorter than the other time periods studied. This choice of a shorter time period for the most recent part of the corpus was made based on extra-linguistic as well as methodological considerations. The extra-linguistic consideration in building the corpus is that the field of ecology has gained extensive media coverage over the last ten years. Therefore, our assumption is that the language of ecology has changed more dramatically in the last period of the corpus than it has in the first subcorpora. Using such division will therefore allow

us to gather hints of such extra-linguistic influence on the content of our documents. Table 1 offers the breakdown of the corpus across the time periods.

Table 1: Number of occurrences for each time period

Time periods	**1951, 1960**	**1961, 1970**	**1971, 1980**	**1981, 1990**	**1991, 2000**	**2001, 2005**
Occurrences	77,454	78,258	76,476	76,799	217,395	152,296

From a methodological point of view, our aim is to compare the data that can be extracted from a ten-year period with the data that can be extracted from a five-year period. This falls into plans which form part of a larger study on the building of specialized diachronic corpora, and on how different time periods (five, ten, 50 or 100 years) can lead to different types of linguistic information (See also Dury and Picton 2009).

2.2 Corpus processing

The current section describes all the processing stages necessary to obtain the results described in Section 3. All tools used to handle our corpora are freely available or accessible on the Internet.

Part-of-speech tagging

The first step in preparing the documents is handled by TreeTagger (Schmid 2004), a part-of-speech (POS) tagger. The tool goes through the document, and assigns a part-of-speech and a lemma to words based on probabilities learned from a training corpus. For our study, we have used the standard dictionary provided with the tool, and therefore have not retrained the tool for our own purposes. The part-of-speech tagging is used in order to distinguish between identical words at surface level, thus allowing us to make a crucial distinction between words like *chair*, which could either be a noun or a verb. The other component of the tags assigned by TreeTagger is the lemma, which is used to group various inflected forms of the words in order to compute reliable, representative word frequencies.

Corpus specificity evaluation

Corpus specificity is evaluated using a measure proposed by Lafon (1980), and called *calcul des spécificités*. It enables the comparison between the frequency of a word in a subcorpus (here a precise time period), and the frequency of the same word in the rest of the corpus (our reference corpora). This technique identifies three sets of words based on their frequency according to a standard normal distribution: positive, negative or neutral specificities. This distribution is

evaluated based on the observations made in the reference corpus, and is used to compute a theoretical frequency. Simply put, the specificity scores represent the distance between the theoretical frequency and the frequency found in the subcorpus. Specificity was computed on the part-of-speech lemma pair.

The positive specificities have a frequency which is higher than could be expected in the subcorpus. The negative specificities have a frequency which is lower than expected, while items from the last group have a frequency in the normal range. In our study we are mainly interested in negative specificities when compared to the previous time periods. The rationale behind this decision is that we believe a significantly low frequency in a particular time period compared to the previous ones could mean that a word is slowly disappearing from the corpus.

Collocation analysis

For all items in the list of negative specificities, we extracted a list of statistically significant collocations from our corpora. The extraction was performed using Text-NSP (Pedersen and Banerjee 2003), a set of tools dedicated to the extraction of n-grams (string entities of length *n*, e.g. character or word sequences) from a corpus. The package provides various means for the statistical analysis of n-bigram occurrences. For the current study, we are solely interested in bigrams, and the output of the tool will be limited to pairs of lexical items. The n-grams are generated using the word and part-of-speech pair while the lemma is left aside. This decision is based on the observation made by the subject matter expert who observed that, in some cases, the plural and singular forms of a term have different behavior over time. For example, the plural noun *terrains*, used in the earlier time periods of our corpus, disappears and is solely used in its singular form (*terrain*) in the later periods.

We extracted bigrams based on a window of three words with a minimal frequency of five, and an empirical threshold score of six using the log-likelihood test. This last test is one of the several measures proposed by NSP-Text to evaluate the strength of association between two words.

2.3 Term validation

All potential necrologisms that were extracted from the corpus were presented to an expert of the field of ecology in order to be validated. One of the major obstacles of terminological studies is to be able to find subject matter experts; such a difficulty is even more important when trying to describe the changes in the lexicon over the years. For some phenomena such as lexical disappearance, the corpus and specialized resources might be sufficient to validate the status of

a potential necrologism; but in the case of semantic variations over the years, the input of a subject matter expert is required.

3 Corpus exploration

The first objective of this project was to list various linguistic clues which would help detect cases of term necrology, and which are compatible with computer analysis to enable a semi-automatic exploration of the corpus. We were also interested in evaluating which clue(s) was (were) the most productive for the corpus compiled. The clues explored are:

3.1 Frequency

By frequency we understand the decrease in frequency of a term over time or its complete disappearance from the corpus. The frequency is observed through the specificity of the words. The technique used here consists in comparing each one of the five subcorpora with the others in order to find the lexical specificities of each one of them, and to detect potential necrologisms (i.e. lexical items which are abnormally absent from at least one subcorpus), as shown in Figures 1 and 2:

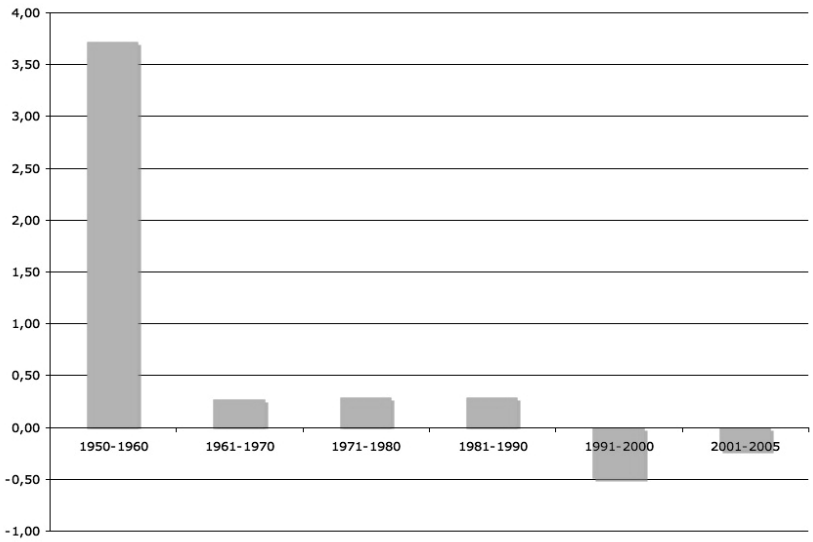

Figure 1: Frequency of occurrences for disclimax *between 1950 and 2005*

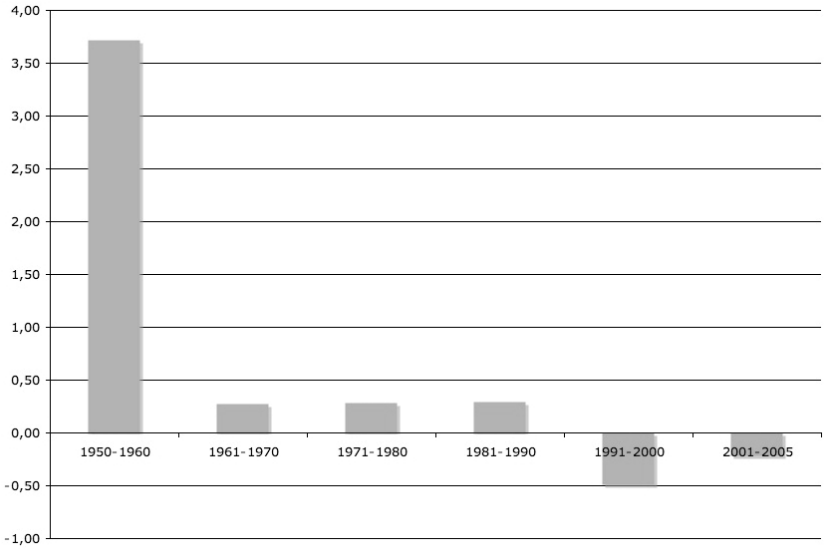

Figure 2: Frequency of occurrences for superorganism *between 1950 and 2005*

The total number of lexical items (restricted to nouns) in the whole corpus is 12,486. Out of these 12,486 lexical items, 127 were identified as having both a significantly lower frequency in one subcorpus, and a decreasing frequency in the last time period (2001, 2005). Out of these 127 items, 3 were confirmed as actual necrologisms by our expert: *biogeocoenosis*, *decreaser* and *subassociations*.

99 lexical items were identified as having both a significantly lower frequency in one subcorpus, and a decreasing frequency in the last two time periods (1991, 2000; 2001, 2005). Out of these 99 items, six were identified as having disappeared from the lexicon by our expert: *disclimax*, *ectocrine*, *genoecodeme*, *niche pre-emption*, *protocooperation* and *superorganism*.

One lexical item (*protozoon*) was identified as having both a significantly lower frequency in one subcorpus, and a decreasing frequency in the last three time periods (1980, 1990; 1991, 2000; 2001, 2005), but a further analysis discards it as a necrologism.

No lexical item was identified as having both a significantly lower frequency in one subcorpus, and a decreasing frequency in the last four time periods (1970, 1980 2001, 2005).

26 lexical items were identified as having both a significantly lower frequency in one subcorpus, and a decreasing frequency in the last five time periods (1960, 1970, 2001, 2005). Out of these 26 items, eight were confirmed as actual necrologisms by our expert: *associes*, *biociation(s)*, *bioces*, *life zone(s)*, *mesobiota*, *terrains*. Table 2 provides a summary of the results obtained using the frequency clue only:

Table 2: Number of potential and confirmed necrologisms detected using the frequency clue

Time periods	**1961, 1970**	**1971, 1980**	**1981, 1990**	**1991, 2000**	**2001, 2005**
Number of candidates, or potential necrologisms	127	99	1	0	26
Number of confirmed necrologisms	3	6	0	0	8

3.2 Grammar

The extraction of contexts showing changes in the part of speech of a term over time may contribute to detect what we call grammatical necrology. For example, the term *dominant* was used only as a noun in the corpus up to 1990 and from then on has appeared only as an adjective, as shown in the corpus extracts below:

(1) In land communities, plants usually are major dominants because not only are they producers but they provide shelter for the great bulk of the organisms in the community. [1950, 1960]

(2) Changes in environmental conditions may thus produce appreciable changes in composition and activity of subordinates with little change in dominants. [1981, 1990]

(3) A shift from annual weed species to dominant perennial species with increasing distance from water. [1991, 2000]

(4) In early succession, chewing insects, mainly Coleoptera and Diptera were dominant. [1991, 2000]

The term *permeant* is another similar case, since it appears only in the noun form in the two first subcorpora, but then only as an adjective in the succeeding time periods. In this case however, it seems that the term has also changed its meaning over time as it appears clearly in the extracts below. We found no trace

of the term *permeant* used to describe animals in the later part of the corpus, where the adjective appears only related to chemical elements.

The grammatical necrology seems therefore, in this particular case, to go hand in hand with the semantic necrology.

(5) …the term here used to include both plants and animals which cling to or are attached to the major plants; and the permeants, which include the highly motile animals such as birds, reptiles, mammals, and active flying insects. The latter organisms move freely (or "permeate") from one strata to another and from one community to another. [1950, 1960]

(6) The "permeants" require still other techniques, such as direct observation, trapping in the case of mammals, and the mapping of territories in the case of birds. [1961, 1970]

(7) In particular, the role played by different mitogen activated protein kinases (MAPKs) and by the production of eicosanoids were investigated utilising specific cell permeant, pharmacological enzyme inhibitors. [1991, 2000]

(8) The influence of the permeant salinity on the intrinsic permeability of the material has been highlighted, as well as the swelling capacity of this mixture. The ionic species distribution depends on the kind of infiltration water and the duration of the test. [1991, 2000]

3.3 Morphology

As well as grammatical changes, context extraction can also help detect morphological necrology, i.e. the disappearance of a term affix over time, as is the case with *subdominant* (which has disappeared from the corpus whereas the term *dominant* is still used in the latest time period), and *ecodeme*, *genodeme*, *genoecodeme*, which were all defined as different types of *demes* (or ecotypes), and which have also disappeared from the corpus. As it is the case with the term *permeant* presented above, the disappearance of an affix sometimes also means the disappearance of the term meaning.

(9) The subordinate species exist because they are able to occupy the niche or portions of it that the dominants cannot effectively occupy. The subdominants tend to be more specialized in their environmental requirements and more narrow in their physiological tolerances. [1961, 1970]

(10) Any prefix only states one characteristic of such a group of individuals: an 'ecodeme' denoting a deme occurring in a specified kind of habitat; a 'genodeme', a deme differing from others genotypically; and, accordingly, a genoecodeme an ecodeme differing from others genotypically. [1961, 1970]

Four other terms extracted from the corpus and confirmed as disappeared by the expert can be considered as morphological necrologisms: *subformation*, *subregion*, *biociation*, and *superorganism*.

3.4 Distribution

A drastic change in the collocates associated with a lexical item can indicate a change in meaning over time as shown in Figures 3 and 4 with the term *province*. In the early part of the corpus (1950, 1960), *province* was used as a synonym of the terms *zone* and *ecological area*, as in *animal province*, *floristic province*, *biotic province*, whereas in the latest part of the corpus (from 1981 onwards), it seems to be used rather as a synonym of *geographical region,* as in *the Quebec province*. The use of the distribution clue helped us to detect two other semantic necrologisms, *district* and *range*. As the term *province*, *district* and *range* were used to designate an ecological area, as in *forest (district)*, *zoological (district)*, *wooded (district)*, *(district) formation*, *segregate (range)*, *alien (range)* in the first part of the corpus (until 1971). This meaning seems to have disappeared from the later part of it. *Province*, *district* and *range* are the only terms which were found thanks to the use of the distribution clue.

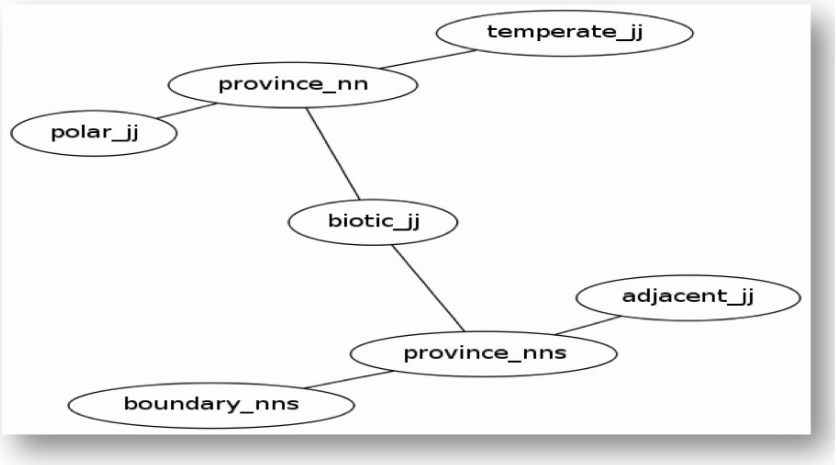

Figure 3: Most frequent collocates associated with province *in the first sub-corpus (1950, 1960)*

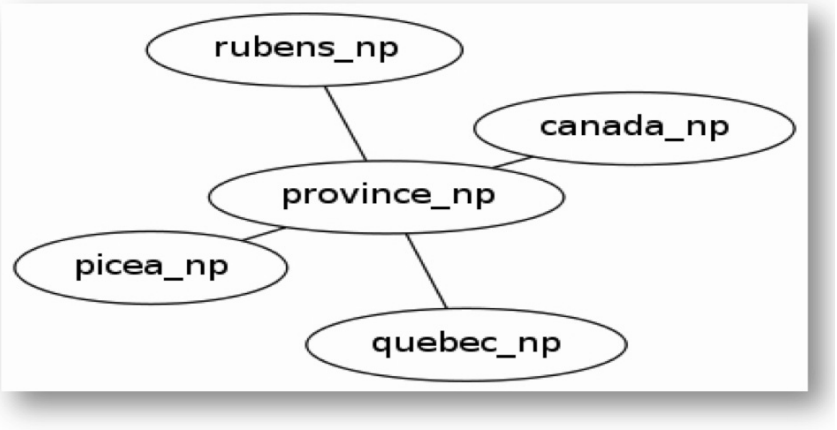

Figure 4: Most frequent collocates associated with province *from 1981 onwards*

3.5 Punctuation marks

Some punctuation marks (parentheses, inverted commas) as well as typography (italics for instance) can point to passages in texts where authors express themselves about the words they use. Although these marks are obviously also used in a wide range of phenomena as new words, citations, words considered as inappropriate by the author, etc., they are quite productive when used with other clues, like the linguistic marker clue for instance.

(11) In other words, the life form spectrum of the vegetation (also known until recently as the community) and the life form spectrum of the flora are not necessarily the same. [1951, 1960]

(12) Seral plant communities in the Clementsian scheme were called "associes" to distinguish them from climax communities which were "associations." [1950, 1960]

(13) His word "mores", which has an awkward and alien sound, disappeared, but the idea he was seeking to express by it later found an outlet in Elton's "niche" and became a central ecological concept of great intellectual reward. [1971, 1980]

3.6 Markers

The use of linguistic markers (like *formerly known as*, *previously named*, *the term....turned out*, etc.) by ecologists when they describe or define their own terminology may help detect obsolete terms or necrologisms in the making, as shown in examples 14 to 16. As mentioned above, this clue can be used in association with the analysis of punctuation marks.

(14) The term superorganism turned out to be unnecessary, not even allowing an always welcome simplification of ecological jargon. [1991, 2000]

(15) The deme-terminology formerly used by ecologists offers no appropriate term for the afore-mentioned concept, the term 'deme', denoting any group of individuals of a specified taxon, being quite as vague as the concept of population. Any prefix only states one characteristic of such a group. [1961, 1970]

(16) Associations, formerly known as sociations, however defined, were also likely to afford living places to particular species of animals which had adapted to them and whose presence must have some

impact on the vegetation, and various attempts were made to include the names of characteristic animal species in the association descriptions. [1971, 1980]

3.7 Synonymic variation

«La démarche terminologique ne serait pas complète en néologie si l'on n'examinait pas toutes les dénominations utilisées pour le même concept. En effet, si l'une finit généralement par s'imposer, c'est seulement après un certain foisonnement néologique» (Humbley 1994: 709).

According to John Humbley, synonymic variation often accompanies the creation of a new concept and the term that describes it. Our assumption here is that such a synonymic variation may also occur when terms disappear from a lexicon, as shown in example 17:

(17) The term ecosystem was first proposed by Tansley in 1935, but the concept is by no means so recent. Microcosm (Forbes, 1887), holocoen (Friederichs, 1930), biosystem (Thienemann, 1939), and bioinert body (Vemadsky, 1944) are terms which have been used to express similar ideas. [1951, 1960]

4 The various types of necrology identified

The following types of term necrology were identified thanks to the data extracted from the corpus:

- The semantic necrology (the disappearance of a meaning over a period of time), like *province*, *district* and *range*,
- The lexical necrology (the disappearance of a lexical unit from the corpus, corresponding to knowledge evolution), like *superorganism*, *biociation*, *bioces*, etc.,
- The grammatical necrology (the disappearance of a part of speech), like *dominant* and *permeant* (the noun form has disappeared), terrain (the plural form has disappeared),
- The morphological necrology (the disappearance of an affix) like *biogeocoenosis*, *subdominant*, *bioecodeme*, etc.

5 Conclusion

As mentioned in the introduction, our aim with this first work on necrology was to explore various clues in order to extract them semi-automatically. Both the

performance and the availability of natural langue processing tools to support the extraction vary from one clue to another.

The frequency clue is the most productive of the clues explored with the highest number of necrologisms found when extracting information from the subcorpus going furthest back in time (in our case 1950, 1960). Such results could be anticipated since the terms present in the oldest part of the corpus are more likely to disappear from usage over the years if there is indeed some process of necrology over time.

Our assumption that the language of ecology has changed dramatically over the last decade because of strong media attention also proves to be true. In fact, 21 necrologisms (out of 38) are found in the last two time periods of the corpus with 3 necrologisms being extracted from the last five year-period. Such an observation shows that, when building a diachronic corpus using documents taken from a rapidly evolving field of knowledge (like ecology, or computer sciences, etc.), short-time periods (less than ten years) are relevant for study.

As stated, frequency is the most productive of all clues, but none of the linguistic clues explored above are satisfactory, if used alone, to find all the necrologisms. For instance, the semantic necrologisms (*province*, *range*, *district*) could not be extracted using the frequency clue alone. The same could be said of the grammatical necrologisms (*dominant*, *permeant*) since they could not be detected using the lexical marker clue on its own. They are therefore more productive when used together. Picton (2008), when studying the evolution of knowledge through terminology, comes to a similar conclusion and suggests combining clues in order to increase performance.

Prior studies (Condamines, Rebeyrolle and Soubeilles 2004) have shown that the resorting to subject matter experts in order to review or to validate potential neologisms may sometimes prove to be a difficult task, since they are not always aware that the terminology they use at work is evolving. From our experience with the expert involved in our study, identifying necrologisms and confirming that a term has disappeared -or not- from a lexicon is easier.

In the near future, we plan to enlarge our global corpus by adding new documents for the existing time periods and by going back further in time for the periods studied. During the analysis of the results, we could see that the absence of a term from our corpus did not mean that it was a necrologism, but that it sometimes was just not included in the papers we gathered for the study. Such problems arise with just about any corpus-based studies but enlarging the corpus would probably minimise the number of cases where this happens.

We expect that the linguistic clues described in this paper are going to be exploited with other languages (French, Danish) and other subject areas such as

medicine, computer sciences and tourism. Other linguistic clues to detect term necrology are also going to be explored in order to identify other types of necrology. For instance, complex terms sometimes disappear and are replaced by acronyms or general terms disappear and are replaced by a series of hyponyms, etc. We need to establish a set of clues and methods to handle such cases.

Although NLP tools have been used for the current study, we need to integrate them as much as possible so as to minimize the manual validation of large amount of data. Statistical techniques are going to be further integrated in order to help us identify necrologisms in the making. On a similar note, linguistic studies on motivation need to be explored in order to see if *ill-* and *un-* motivated terms are more at risk of becoming necrologisms.

We would also like to explore the optimal size of the time slices to be used for necrology detection, if such a thing exists. For example, would we reach better or worse results if we used larger of smaller time periods to build our subcorpora. Our intuition is that this might be directly correlated to the level of activity in the field being studied and that subject areas where concepts evolve more quickly might benefit from smaller time windows.

References

Boulanger, Jean Claude. 1989. L'évolution du concept de 'néologie' de la linguistique aux industries de la langue. In C. de Schaetzen (ed.). *Actes du colloque sur l'histoire de la terminologie*, *Terminologie diachronique,* 193–211. Paris-Bruxelles: Conseil international de la langue française.

Cabré, Maria Térésa and Lluis de Yzaguirre. 1995. Stratégie pour la détection semi-automatique des néologismes de presse. *Traduction, Terminologie, Rédaction (TTR)* 8 (2): 89–100.

Condamines Anne, Josette Rebeyrolle and Annie Soubeille. 2004. Variation autour de la terminologie dans le temps: Une méthode linguistique pour mesurer l'évolution de la connaissance en corpus. *Proceedings of the Euralex International Congress*, 547–557. Lorient: Université de Lorient.

Drouin, Patrick, Nathan Ménard and Annie Paquin. 2006. Extraction semi-automatique de néologismes dans la terminologie du terrorisme. *Actes des 8èmes Journées internationales d'Analyse statistique des Données textuelles (JADT 2006)*, 389–400. Besançon: Presses Universitaires de Franche-Comté.

Dury, Pascaline. 2000. Les variations sémantiques en terminologie: Étude diachronique et comparative appliquée à l'écologie. *Dyalang, Dynamiques sociolangagières* 275: 17–32. Rouen: Presses Universitaires de Rouen.

Dury, Pascaline and Aurélie Picton. 2009. Terminologie et diachronie: Vers une réconciliation théorique et méthodologique? *Revue Française de Linguistique Appliquée (RFLA)* Vol. XIV, 2009-2, *La terminologie: Orientations actuelles*, 31–41.

Grzega, Joachim. 2002. Some aspects of modern diachronic onomasiology. *Linguistics* 40 (5): 1021–1045.

Guilbert, Louis. 1974. *La néologie lexicale*. Paris: Didier-Larousse, Langages.

Humbley, John. 2006. La néologie: Interface entre ancien et nouveau. In R. Greenstein (ed.). *Langues et cultures: Une histoire d'interface*, 91–103. Paris: Publications de la Sorbonne.

Lafon, Pierre. 1980. Sur la variabilité de la fréquence des formes dans un corpus. *MOTS* 1 : 128–165.

Pedersen Ted and Satanjeev Banerjee. 2003. The design, implementation, and use of the ngram statistics package. *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, 371–381. February 2003.

Picton, Aurélie. 2008. Combining clues to explore knowledge evolution. *Proceedings of Terminology and Knowledge Engineering (TKE)*, 19–21. Copenhagen: Copenhagen Business School.

Sablayrolles, Jean-François. 2000. *La néologie en français contemporain. Examen du concept et analyse de productions néologiques récentes*. Paris: Honoré Champion.

Sablayrolles, Jean-François. 2003. *L'innovation lexicale*. Paris: Honoré Champion.

Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. *Proceedings of the International Conference on New Methods in Language Processing*, 44–49. Manchester.