Orthographic Neologisms

Selection Criteria and Semi-Automatic Detection

Maarten Janssen
Instituto de Linguística Teórica e Computacional
Rua Conde de Redondo, 74 – 5
Lisbon, Portugal
maarten@janssenweb.net

Abstract

The two most commonly used ways of defining when a word is a neologism are the lexicographic definition and the corpus-based definition, which hold that a word is a neologism if it resp. does not occur in the dictionary, or was not previously used in reference corpora. This article argues that the lexicographic definition does not properly define words to be new, since words can be lacking from dictionaries for other reasons, and that the corpus-based definition does not provide the necessary control over which words appear in the reference corpus. What is called for is a hybrid method, called the *extended lexicographic diachronic definition* – which depends both on lexicographic absence, and manual verification in reference corpora.

Keywords: neologisms, identity criteria, semi-automatic detection

1. Introduction

Neologisms form a highly relevant linguistic category for many reasons – they are the elements that make a language living and dynamic rather than dead, they are indicative of language change, they form a serious obstacle in computational analysis and translation, and they help to show productive morphology of a language. Consequently, there is a substantial number of linguistic research units devoted to the observation and analysis of neologisms – OBNEO in Barcelona, ONP in Lisbon, ONLI in Rome, CGTN in France, APRIL in Liverpool, WortWarte in Tübingen, etc. Furthermore, many of the major lexicographic publishing houses have their own methods of observing neologisms.

There are two main goals in the linguistic observation of neologisms. On the one hand, updating existing lexicons and dictionaries with the newly arisen words. And on the other hand the analysis and description of the neologisms themselves in terms of distribution over word-classes, statistics on derivational methods, statistics on loan word origination, etc. Especially the latter type of research depends on the detection of *all* the neologisms occurring in a given corpus. However, there is no clear definition of what a neologism is – making the detection of all neologisms a far from trivial task.

One of the best attempts at a definition of a neologism is given by Rey (1975/1995), who concludes that there are no objective criteria for being a neologism. Furthermore, there is a classification of neologism definitions formulated by Cabré (1999): a psychological definition, a lexicographic definition, a diachronic definition, and a definition based on a word exhibiting systematic signs of formal or semantic instability. The fact that none of these give fully satisfactory identity criteria is accepted as an inevitable consequence of applied linguistics, and almost all observation groups take one of the four criteria of Cabré without further discussion.

This article attempts to give a more substantial analysis and comparison of the various criteria. It will give an overview of the advantages and shortcomings of the existing criteria, mainly the corpus-based and lexicographic criteria, and argue that a hybrid approach is called for. This hybrid criterion will be called the *extended lexicographic diachronic criterion*.

The analysis will be given from the perspective of the semi-automatic detection of neologisms, and hence will focus primarily on *formal* or more precisely *orthographic* neologisms (the difference will be discussed in section 7.2). For the semi-automatic detection, a flexible on-line tool called *NeoTrack* will be used, described in section 3. The detection of

neologisms with NeoTrack using the hybrid criterion goes hand in hand with the creation of a morphological database, containing all those word-forms that are considered to be non-neologisms.

2. Psychological Neologisms

A neologism is, by the very meaning of the term, a *new word*. The term 'neologism' is commonly found in dictionaries, where it is intended to be an established yet new word in the language, as for instance defined by Bußmann (1990): *"Neugebildeter sprachlicher Ausdruck ... der zumindest von einem Teil der Sprachgemeinschaft ... als bekannt empfunden wird"*[1]. But from a linguistic perspective, the more interesting notion is that of a really new word that has not yet made it to the lexicon – words in the process of lexicalisation. Often, words are even traced from their very origin, meaning that any (intentional) occurrence of a new word will do, putting neologisms on a par with occasionalisms[2].

Although a common label in dictionaries for a long time, a first serious attempt at a definition of a neologism was given in 1975 by Alain Rey, former chief editor of *Le Grand Robert.* Rey presents a discussion of what of a word is, the types of neologisms one can distinguish (formal, semantic and pragmatic), and what it means for a word to be 'new'. Concerning the latter, Rey concludes that no solid, objective criteria for newness can be given, and that hence the label *neologism* is only an indication of a subjective sentiment. His position is characterised by what Cabré (1999) calls the *psychological definition*:

- *A neologism is a word that is perceived as new by the language community*

Given the vagueness of the term, Rey argues against the use of the word *neologism* in dictionaries, stating that 'neologism' is only a *pseudo-concept.* The basic motivation for Rey is the absence of the necessary temporal stability of a language and can be summarised as follows. Diachronically, the notion of neologism bears no meaning – it is not the abstract, timeless word that is "new", but a neologism is a word that is new in a given language *at a given moment in time*. And only with respect to that time can a word count as a neologism.

The word being new should imply that the word is currently part of the language, but was not so previously. But language does not progress through well-defined stages, where the words in the new lexicon can be compared to the words in the old lexicon. Firstly, "new" is a relative notion – some words may be older than others, but there is no demarcated period for being new[3]. And secondly, there is no well-defined, stable lexicon of a language against which the newness can be tested - a language cannot be stably defined within *"its limits in the chronological, spatial, and social dimensions"* (Rey, 1995, p.75).

2.1. Neologism and Community

Without the possibility of verifying a word as new against a stable language setting, the notion of a neologism reduces to a subjective feeling of being new, as in the psychological definition above. Given the nature of language, the feeling of newness should reside with the language community rather than the lexicographer assigning the term.

But there are two objections against using the perception of the language community as definitional for neologisms. Firstly, measuring the perception of the community is a tedious and time-consuming process, and not feasible for an entire lexicon. And secondly the language community itself is not the most reliable source for perceived novelty. This last point is made clear by an example given by Rey himself:

> I have been able to verify that French speakers categorise the word 'stockfish', borrowed from Dutch in the 14th century, as a neologism and an Anglicism in the same way as the recent 'stockcar'. (Rey, 1995, p.74)

Rey does not intend to count the word *stockfish* as a neologism – the subjective perception is not meant as a correct criterion for being a neologism, but what Rey argues is that being a neologism reduces to a feeling in the absence of real criteria: merely the subjective opinion of the individual linguist or lexicographer assigning the label. In assigning the label *neologism* the linguist or lexicographer will attempt to apply a real notion of newness, but have no proper way of doing so.

Although the arguments given by Rey still hold, much work has been done in lexicology to attempt to give a delimitation of solid language segments – augmented greatly by the arrival of electronic corpora. The remainder of this article will attempt to give a definition of a stable language fragment in the light of these new developments. Although it is clear that no strict delimitation can be given, any clarification of when a word should count as new will lead to a more useful definition of a neologism than the psychological definition – which will be rejected as little more than the absence of a definition.

3. Exclusive Definitions - NeoTrack

Almost all definitions of a neologism besides the psychological one depend on negative evidence – definitions of neologism based on exclusion[4]:

- *Any word not appearing in a pre-determined exclusion lexicon is a neologism*

The exclusion lexicon is intended to define the stable language fragment, against the background of which the neologisms counts as new.

A major advantage of the exclusive definition is that it is easy to automate. There are various tools for semi-automatic detection of neologisms, but this article will be based on an web-based tool called *NeoTrack.* NeoTrack is a flexible tool for finding neologisms defined by the exclusive definition, working in a modular fashion to allow easy modification of the detection method used. This section is not intended as a full explanation of the NeoTrack system, but to give a brief description of its basic operation, as a general tool for semi-automatic neologism detection.

The way NeoTrack works is illustrated in figure 1: from whichever source is considered to be most appropriate, an *exclusion list* is created[5]. The exclusion list is a file listing all those words that are considered non-neologisms. A corpus that is to be searched for neologisms is entered in its original format (currently only HTML), cleaned up, and tokenised to render a list of all token words occurring in the corpus. This list, minus the exclusion list, is exactly the list of neologisms as defined by the exclusive definition.

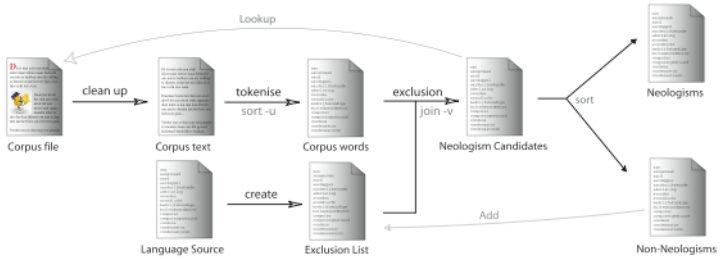

Figure 1. The general set-up of the NeoTrack application

The exclusive definition is rarely used at face value - the results of the exclusion step are taken to be only *neologism candidates*. Amongst the list of candidates there will be non-neologisms, such as typos, proper names, and possibly words that were accidentally not on the exclusion list. Therefore, the neologism candidate list has to be evaluated, sorting the neologisms proper from the false candidates. In NeoTrack, this sorting is done completely manually. In this process it is possible to add, if so desired, those words that are classified as non-neologisms to the exclusion list (see section 7.1).

Because of its set-up, NeoTrack is incapable of detecting semantic neologisms: there is no semantic analysis, so new meanings of existing words cannot be detected. NeoTrack cannot even track formal neologisms: there is no part-of-speech analysis, meaning that homographs of a different grammatical category cannot be detected (more on this in 4.1). What NeoTrack detects is what I call *orthographic neologisms* (see section 7.2). This restriction to orthographic neologisms is not unique for NeoTrack, but holds for most if not all semi-automatic neologism applications.

There is no predetermined definition of an exclusion list in NeoTrack – any list of words can function as an exclusion list - the result of the exclusive definition is highly dependent on the source that is taken to create the exclusion list. There are two common sources for this: dictionaries and corpora. The result of using either of these will be discussed in the next two sections. Because of the independence of the source of the exclusion list, NeoTrack is a general tool, allowing comparison between neologism definitions.

To give a quick impression of the basic workings of NeoTrack: NeoTrack is a lightweight application, basically implementing a web-based front end to a simple set of UNIX commands and PERL scripts. All steps in the process of the detection of neologism candidates operate independently, and can easily be replaced by alternative procedures, making it highly adjustable.

There are some possible ways in which the selection of neologism candidates can be made more restrictive, reducing the manual workload. For instance, proper names form a large part of the neologism candidates, which is why the Cenit prototype (Roche & Bowker, 1999) attempted to apply a list of tools to recognise proper names to exclude them automatically. For Portuguese, NeoTrack uses a dedicated tokeniser to split clitics from verbs, avoiding verbs with clitics from becoming false candidates. And NeoloSearch (Janicijevic & Walker,

1997) attempts to automatically filter out typographic errors by using approximate string matching with other words in the corpus.

But the default set of scripts in NeoTrack is intended to filter out neologism candidates sparsely, intentionally depending more heavily on manual filtering. The philosophy behind the emphasis on manual labour is this: it is easy to filter out false candidates, but virtually impossible to retrieve candidates that were overlooked – and any advanced algorithm is destined to accidentally throw away real candidates as well every now and then. Furthermore, by avoiding the use of algorithms such approximate string matching or lemmatization, the neologism criteria are much more transparent.

4. Lexicographic Neologisms

The most obvious source of a list of established word of a language (the exclusion list) are lexicographic resources: thesauri, vocabularies and in particular dictionaries. Dictionaries attempt to be stable, synchronic projections of the lexicon of a language. Accepting the dictionary as an authority on the matter, novelty of a word can simply be defined in terms of lexicographic absence:

- *Any word that does not appear in the dictionary is considered a neologism*

The advantage of the lexicographic criterion is that it is relatively well defined, since a dictionary provides a fixed number of non-neologisms. The definition of a neologism depends on the choice which dictionary is actually used (for there is more than one for many languages)[6], but with that choice, the list of non-neologisms is fixed.

Despite the relative strictness of the lexicographic definition, there are at least two problems with it. The first is that the definition is not as well defined as it appears at first glance – especially not from a detection perspective. The next few paragraphs will list some problems with the strict definition listed above. But more importantly, the lexicographic definition does not actually define 'new' words – it defines words that are not well established enough to be included in the dictionary. And this non-establishment can be due to novelty, but there are many other reasons for words not appearing in the dictionary, as will be argued in section 4.4.

4.1. Inflection and Lemmatisation

Since dictionaries only list citation form of lemmas, it is obvious that *appear in the dictionary* is intended to mean *whose citation form appears in the dictionary*. Therefore, all inflectional forms of the lemmas in the dictionary have to taken into account in the exclusion step.

Although inflection is rather well defined, there at least two types of problems with the inflectional forms implied by dictionaries. The first concerns *defective* inflections: words for which not all common inflectional forms exist. In English, these are mainly the pluralia and singularia tantum. The word *consciousness* can only appear in singular, so any occurrence of the word *consciousnesses* would be neological or wrong. In principle, good dictionaries indicate this defectiveness. In the case of *consciousness*, most dictionaries mark the word as a mass noun, hence implying it does not have a plural. However, the word *vodka* is also listed as a mass noun but can be used in the plural (*I had two vodkas today*)[7].

For a minimally inflectional language like English, defective inflections are surveyable. But for more inflectional languages they pose a substantially bigger problem. In Portuguese for instance, many verbs are defective, meaning that the verb cannot occur in one or more inflections. This can be for semantic reasons, as in the case of *nevar* (to snow), which is impersonal and can therefore only appear in the 3rd person singular. But there are also verbs that cannot be used in specific forms for grammatical or pragmatic reasons, in which case different sets of inflections are defective. For instance, the word *ressequir* (to dry thoroughly) does not appear in the present conjunctive, or in the present indicative except for the 1st and 2nd person plural, nor does it appear in the imperative (except for the positive 2nd person plural). The class of forms in which a verb can and cannot appear is open: the verb *chover* (to rain) although impersonal can be used in the 3rd person plural in the common metaphoric expression *"choveram as críticas ao filme"*[8]

The case of defective verbs is made even more problematic by the existence of *semi-defectives*: words for which it is not even clear which forms do and do not exist. An example is the verb *explodir* (to detonate). *Explodir* is considered defective in the same way as *ressequir* by some authors, but fully inflected according to others. Semi-defectives also appear because of normative issues: although the Portuguese form *subsumes* is an inflectional form of *subsumir*, it is a not an unmarked form but considered popular (or old-fashioned) – its preferred form is the irregular *subsomes*. These subtleties of defective verbs are not (completely) specified as such in normal dictionaries – the dictionary is hence an insufficient source to discover in which inflectional forms a lemma can appear.

A second type of problem with inflection in dictionaries concerns the existence of not fully productive inflectional forms. For instance, for adjectives in Portuguese neither the synthetic superlative nor the diminutive are fully productive. Yet both forms are commonly considered inflectional forms rather than derivations. To give an example: the word-form *docinho* (sweet) is not listed in the Porto Editora dictionary, since it is the diminutive inflection of *doce*. But not all adjectives have a diminutive form – it is largely colloquial and mainly exists for colloquial words. For instance, the word *confundível* (confusing) does not have a diminutive – however, there is no indication in the dictionary that the word-form **confundivelzinho* does not exist but the form *docinho* does. And in Dutch, intensifying adjectives do not have degree forms: **steenkoudst* (most stone cold - Booij, 2003: 253)

These problems affect the semi-automatic detection of neologisms using the lexicographic definition: many neologism detection applications, such as SEXTAN (Vivaldi, 2000), use a lemmatiser in their software: the word-forms occurring in the corpus are first reduced to their citation forms before they are checked against a dictionary. This method can never deal with the subtle issues of defective lemmas - apart from the more obvious problems like the inherent error rate of the lemmatiser and homographic citation forms with different inflections such as *redar* in Portuguese or *band* in Dutch[9]. The only way to precisely specify the full list of inflections implied by the dictionary is by storing all the correct inflected forms explicitly – in a full-form lexicon or morphological database. In the remainder of this article, *does not appear in the dictionary* should be read as "does not appear in the morphological database derived from the dictionary". However, it should be kept in mind that the morphological database hence contains more information than the dictionary itself.

4.2. Erroneous and False Candidates

A much more serious problem of the lexicographic definition is the fact that it is not properly a definition of a neologism. Appearing in a dictionary might be considered a necessary condition for a word being a neologism - although Sableyrolles (to appear) argues differently - but it is hardly a sufficient condition for being one.

There are two main reasons why a word that does not appear in a dictionary does not have to be a neologism. Firstly, dictionaries are by definition incomplete. This problem will be discussed in the next paragraph. And secondly, not every string occurring in a text is commonly considered a possible neologism (this is not specifically a problem of the lexicographic definition, but a problem that concerns every neology definition).

The most obvious reason for "new" words not counting as neological is typographic errors. Any corpus, even thoroughly checked newspapers, contains typos: "*50% of the tens of thousands of unique, new 'words' per year from (Dutch) newspapers, are typing errors."* (Oppentocht & Schutz, 2003: 224)[10]. Although there are problems with the recognition of typos (for instance in the case of *subsumes* in the previous paragraph), typos are commonly considered non-neological. And proper names are also not commonly considered neologisms, although this is less clearly the case for names that do not appear as proper names but as exemplifications of concepts, as for instance in *Xerox* as the general indication of a photocopier (see also 5.2).

Typos and names are non-neologisms because they are not (fully) considered to be words. But there are also words that are not of the proper type to count as neologisms. As an example: in newspapers, one can sometimes find passages which are not in the same language as the rest of the newspaper, this can be an entire section, as in the case of the English financial section of the German newspaper *Die Welt*, or paragraph, such as an advertisement by a foreign company, or even sentence level direct quotations as in the following Portuguese quotation: *Ao fim e ao cabo, o que conta é que, como se diz, business as usual...* The phrase *business as usual* as a whole might be considered neological, but the individual word *business,* and even more so *as* and *usual* are not commonly considered neologisms in Portuguese.

4.3. Incompleteness of Dictionaries

Dictionaries are almost by definition incomplete repositories of words, and because of this incompleteness, appearing in the dictionary is not enough to be a neologism – there are various reasons for a word not to appear in the dictionary that do not make it a neologism: it might be too infrequent, too predictable, too specialised, too old, obsolete, taboo, or simply accidentally overlooked by the lexicographers. A dictionary is not an objective repository of all words: *"The apparent objectivity of dictionaries rests on an extensive series of subjective editorial decisions."* (Curzan, 2000, p. 96).

Words that are clearly not neologistic despite the fact that they do not appear in dictionaries are transparent compounds: the Dutch word *verjaardag* (birthday) can appear in a large number of compounds. Of these, only 3 are listed in the van Dale dictionary: *-feest* (party), *–geschenk* (present), and *–kalender* (calendar). The word *verjaardagstaart* is not listed – meaning that according to the lexicographic definition, it should count as a neologism. But it

is in no sense of the word new: it has existing for a long while, it is used relatively frequently, it is not dialectical or specialised, it is simply not in the dictionary because its meaning is transparent from its components: *verjaardag* (birthday) and *taart* (cake).

The fact that there is no real way to argue that the word *verjaardagstaart* is a neologism implies that the lexicographic definition does not really define neologisms. It simply defines the notion of "non-dictionary word", which includes typos, archaisms, terminology, semantically transparent words, etc. as well as neologisms. Where neologisms are concerned, the lexicographic definition is a partial definition at best. From the semi-automatic detection perspective, this means that the lexicographic definition does provide no criteria at all for the distinction between the real neologisms and the false neologism candidates. And he fact that the lexicographic definition is underspecified is a fundamental one: the lexicographic definition lacks any direct notion of diachronicity, and since the crucial factor of neologisms is their being 'new', it is a problem which cannot be overcome without substantially altering the definition.

5. Diachronic Neologisms: Corpus Approach

What is lacking from the lexicographic definition is a direct relation between the definition of a neologism and a notion of newness. The diachronic criterion is a *sine qua non* in the definition of a neologism:

- *Any word-form that appears in a recent general language text, and was not previously part of that language is a neologism.*

The problem of the diachronic definition is that it does not, by itself, define what words are part of the language, and hence which words are new. As shown in the previous section, the dictionary is not a sufficient source for establishing which words are part of the language. Dictionaries attempt to be faithful abstractions of a language, but can never be complete repositories. Only the language itself can be a full representation of itself:

> Il est … impossible de définir un fait et un concept linguistiques (comme le néologisme et la néologie) en recourant aux outils extra-linguistiques qui est le dictionnaire[11]. (Sableyrolles, to appear)

Without the use of extra-linguistic tools such as dictionaries, only the language itself can determine which words are part of the language. Or in linguistic terms, it seems that the only possible definition of a neologism is in terms of reference corpora:

- *Any word-form, which appears in a recent general language text, and does not appear in an established reference corpus of that language, is a neologism.*

In the corpus-based approach, the reference corpus operates as an exclusion corpus, and the exclusion list is the list of all the words occurring in the exclusion corpus. There are various ways of employing the corpus-based approach: when observing neologisms in a given newspaper, one can either use an external, established corpus (like Bank of English) as an exclusion corpus, or one can use the accumulated back-issues of the newspaper itself as the exclusion corpus.

But despite its obvious appeal, the corpus-based definition runs into a number of problems, both practical and fundamental. These problems can be divided into two groups: the corpus-based approach considers words neologistic that are not, and it misses out on possible neologisms.

5.1. False Candidates

The corpus-based definition, especially when rigorously applied, yields words that are not strictly speaking neologisms. Apart from the problem of filtering out typos and names, no corpus of any size contains all the words of a language. A word not occurring in the exclusion corpus is often a property of the corpus rather than of the language. To take an example: if the exclusion corpus does not contain any recipes, then the occurrence of a recipe in the new text is bound to contain a lot of "new" words, given the amount of words specific to food – such as *t-bone stake* – and the specific terminology for preparing food – such as the Dutch *blancheren* (simmer). The occurrence of these words can in no interesting way be called neological.

Corpus-based neology applications commonly circumvent this problem by manually filtering out the false candidates from the neologism candidate list. Although this method allows filtering out all the false candidates, it has the drawback that the corpus-based definition does not by itself provide any basis to decide which candidates are true candidates and which are not. The manual filtering consists of the addition of non-corpus-based criteria in a corpus-based framework.

5.2. False Rejections

Every exclusion corpus, no matter how carefully selected, will contain typographic errors, names, and possibly also foreign language quotations. In the corpus-based approach, these occurrences lead to the unsolicited rejection of neologism candidates – any neologism that happens to be homographous to a typo or name will not appear on the neologism candidate list. The missing out on neologism candidates is, from the semi-automatic detection perspective, more serious than the occurrence of false candidates: false candidates can be manually removed, missed candidates are commonly unrecoverable.

Low Frequency Words, Citations and Typos

Even in thorroughly corrected corpora like newspaper, typographic errors do occur. As mentioned in 4.2, they are even very numerous in thorroughly verified sources like Dutch newspapers – making up 50% of all the neologism candidates. Typographic errors in verified sources are more likely to be well-formed yet incorrect words rather than arbitrary combinations of letters, since obvious non-words are easier to spot. And since they are mostly well-formed words, they could be homographous to neologisms. For instance, the word *cannable* listed as a 1998 neologism by the APRIL project, could well have appeared earlier as a misspelling of *cannibal*.

The occurrence of foreign words in a corpus may seem like a marginal problem, but in practice it is often not: the *CETEMPublico* corpus is a 1,5 million word corpus, composed of only Portuguese texts taken from the *Público* newspaper. In this corpus, there are 5.000 token-words with a frequency over 20 which do not occur in the *Porto Editora* dictionary. Amongst the (alphabetically) first 150 of these, there are 9 English words (*about, above, access, after, against, also, am, american, and*), 4 French (*aime, ami, amour, ancien*), and 1 Italian word (*alla*). The most frequent of these is *and*, which occurs 7695 times. And of these foreign words, none appears as a proper part of Portuguese, but only as part of foreign names and foreign quotations.

For most corpus-based purposes, typographic errors are filtered out by means of a reliability threshold: only words occurring more frequently than a threshold number are considered proper words, anything below that number is simply rejected. But for neologism research, this method does not work: it would by definition classify all low-frequency words as eternally neological. A reliability threshold cannot solve this, since misspelled version of common words, such as *thye*, will easily outnumber low frequency words like *dapocaginous*.

Names, Abbreviations, and Acronyms

More common even than typographic errors are names and abbreviations. There are two different problems with names. Firstly, various common English words, such as *Xerox, nylon,* and *sandwich* started their life as proper names. In a pure corpus-based approach, such words would always be missed as neologism.

But the second problem is that on a worldwide scale, there is such a vast amount of names for companies, cities, organisations, and persons that a relevant percentage of all combinations of letters - and hence a large percentage of all potential neologisms is the name of something. The tendency to use acronyms for organisations only increases this chance. This means that in an uncontrolled corpus-based approach, the use of a large corpus, necessary to assure the inclusion of low frequency words, will give a very restrictive notion of neology.

The full scale of this problem becomes clear when using the Internet as the exclusion corpus – in the following way: a word could be considered a neologism if a Google-search for the word, restricted to English pages, gives the result *"No pages were found containing..."* This definition is extremely restrictive because of the enormous number of names, types, and abbreviation on the Internet. To give an indication: upon trying all the five-letters combinations starting with the letter *a*, only 52.937 of the 456.976 permutations did not appear on the Internet.

Establishing a reliability threshold is also difficult – a vast amount of the combinations occur more than sporadically – for instance, the dictionary word *shilly-shally* occurs only 1460 times on the internet, and more uncommon words like *nanocephalous* (having an extremely small head) and *natalitial* (relating to a birthday) even fewer – resp. 721 and 365 times. On the other hand, many seemingly arbitrary combinations and non-English words such as *aadhi,* and *acido* occur much frequently than that: *Aadhi* occurs 2180 times, mostly because it is part of the name *Ek Din Aadhi Raat* (a major Indian movie director), *asway* 3.760 times mostly as the pig Latin form of *was*, and *acido* even 21.700 times, mostly on Spanish pages on English sites, but also for instance as part of film titles like *"Ácido Sulfúrico"*. There are 14.455 combinations that occur more than 1460 times – only a small number of which consists of actual words – and almost 8% of all possible 5-letter combinations starting with *a* occurs more than 365 times.

The problem with the straightforward corpus-based approach is that it does not allow control over the treatment of the corpus. It is not desirable to have *all* the strings in the exclusion corpus appear on the exclusion list – only those strings that are properly words of the language. But the controlled creation of an exclusion list from a corpus does no longer count as fully corpus-based – in a sense, building a controlled exclusion list from corpora is more lexicographic than corpus-based in nature. Many dictionaries even explicitly consist of such controlled lexicon from established corpora.

6. Extended Lexicographic Diachronic Neologisms

If we consider dictionaries to be correct, but incomplete repositories of the words of a language, then the diachronic criterion will be a strengthening of the lexicographic definition – i.e. all diachronic neologisms will be lexicographic neologisms, but not the other way around. According to the extended lexicographic diachronic definition, only those words that did not appear in the dictionary *because they were too recent* should be considered neologisms:

- *Any word that does not occur in the morphological database derived from the dictionary because of its recentness is a neologism.*

The idea behind the extended lexicographic definition is that a dictionary leaves out many words because of restrictions in size, rather than considerations about whether the word belongs to the language or not. The definition then depends on a reverse definition of the selection criteria used by the reference dictionary.

Other than with the strictly dictionary-based criterion, the extended lexicographic criterion does have a source of information to depend on when the dictionary itself fails to mention the word – the same source the dictionary itself was derived from, i.e. the corpus. To discover whether novelty was the reason for the lexicographer to leave the word out, one has to consult the corpus to reconstruct the lexicographer's motivations.

6.1. Hybrid Approach

The extended lexicographic approach is in a sense not strictly lexicographic. In fact, one could argue that it is closer to a corpus based approach, given the fact that the exclusion list is, at least as far as the extension on the dictionary is concerned, directly compiled from the corpus sources.

One could even argue that the use of the dictionary is only a pragmatic choice: the same approach can be used starting from scratch. When starting with an empty morphological database, the methodology would start out by considering all words neologism candidates, all of which have to be sorted out manually as either words or non-words by means of corpus verification. In that way, the morphological database would be a direct representation of all the proper words encountered in the corpus. An important reason for the use of a dictionary is the practical infeasibility of this fully corpus-driven approach. From that perspective, the term *controlled corpus-based definition* might be the better designation of the approach.

On the other hand, one could also argue that the extended lexicographic approach is a true lexicographic approach, where the morphological database *is* the lexicographic source. The morphological database is created in the same fashion as dictionaries are, using the same criteria for the inclusion or rejection of words, except for the common limitations dictionaries have concerning their size: there is no reason to leave out words for lack of space, which also lessens the necessity to strictly distinguish specialised language from general language.

The extended lexicographic approach combines these two aspects, and is hence a hybrid between the corpus-based definition and the lexicographic definition. These two approaches are largely overlapping, except for the intervention of the lexicographer in the case of the lexicographic approach. What the extended lexicographic diachronic criterion relies on is not any particular lexicographic product, but the lexicographic method – and not automatic corpus exclusion but controlled corpus use.

6.2. Corpus Verification

The judgement whether a given neologism candidate is a proper neologism, or a gap in the morphological database should be based on corpus verification. But other than the common practice in corpus-based methods, this verification should be executed manually: not only should the neologism candidate occur frequently enough[12] in a pre-defined exclusion corpus, but it should occur in the exclusion corpus as a correct word. That is to say, it should be manually verified whether the occurrence is not a typo, a foreign language quotation, or any of the other marginal occurrences discussed in 5.2.

Because of this manual verification, it is even possible to use the Internet as a reference corpus. For regular corpus based research, the Internet is not a proper source, since it does not provide a balanced, stable and (relatively) error free corpus. However, for the purpose of merely verifying the existence of the word, it can be used with care. It is necessary to verify whether the source of the occurrence is a general language article rather than (partly) in dialect or technical language

When verifying the occurrence of the neologism candidate in the exclusion corpus, a threshold period should be taken into account: an arbitrary period during which new words count as neologistic, after this threshold period the word will be considered established in the lexicon. Although the actual period chosen is arbitrary (3 years seems to be the commonly acceptable period), the notion of a boundary is implied by a binary notion of neology. Because of this threshold, no texts in the exclusion corpus should be younger than 3 years[13].

With these consideration in mind, the extended lexicographic diachronic criterion boils down to the following: any word not occurring in the morphological database is a neologism candidate, and neologism candidates can be either non-words, proper neologisms or gaps in the database. Whether a word is a gap or a proper neologism is determined by manual verification in an exclusion corpus – gaps are those words that do occur in the exclusion corpus sufficiently frequently as a proper word. There are several arbitrary parameters in this criterion: the source used for the initial content of the morphological database, the constitution of the exclusion corpus, the threshold occurrence frequency, and the threshold neologism period. But with these parameters properly defined, the extended lexicographic diachronic criterion not only gives a well defined criterion for being a neologism, but a criterion that seriously defines a notion of newness.

7. NeoTrack Revisited

Although NeoTrack is largely a multi-standard neologism detection program, it has optimised for the extended lexicographic diachronic definition – and is currently used for the detection of neologisms in European Portuguese by the *Observatório de Neología de Português* (ONP). The morphological database used to store the full form lexicon is called *MorDebe*. This chapter will explain how the set-up of NeoTrack using the extended lexicographic diachronic criterion works, and how it interacts with the MorDebe database.

7.1.Neologism Candidate Verification

Before the execution of the exclusion step (see figure 1), the exclusion list is updated, NeoTrack creates a new exclusion list, based on the latest version of the MorDebe database. The list of neologism candidates produced by the exclusion step is added to a database of neologism candidates, and ready for manual processing. In the manual processing step, all the neologism candidates are presented one by one, including the context of their original occurrence in the corpus file. The interface for this candidate verification step is show in figure 2.

Figure 2. NeoTrack – Neologism Verification Window

The window in figure 2 consists of several parts: on the top is the orthography of the candidate, along with the corpus it occurs in, and at the bottom is the context in which the neologism occurs. If more then one occurrence of the neologism candidate is present in the text, all occurrences are shown.

On the right-hand side, those words that are not words of the proper type can be discarded. In the discarding process, it is possible to indicate the reason for their inappropriateness: *non-words* are those strings that are not real words in the text, such as code, or partial words separated by line breaks. *Typos* are those words that are occurring in the text, but orthographically incorrect. *Citation* is for words occurring in foreign language quotation, or parts of foreign names such as the word *mama* in the title of the Spanish movie *Y tu mama también*. Neologism candidates are kept in a database after being processed, along with an indication of what was done with it: whether it was added to the neologism database, to MorDebe, or discarded – along with the reason.

The judgment of whether a word is actually a proper word is done by personal opinion alone. But the classification of a word as a neologism or a proper non-neologism is done by reference to external sources. To facilitate this, on the far bottom is a number of links to directly search for the neologism candidate in a number of selected corpora. Since the verification is manual, *Google* is included as one of the possible verification corpora – with the provisions mentioned earlier.

In case the word belongs to the category of proper non-neologisms, the word can be directly added to the MorDebe database. To do this, the user has only to indicate the lemma and the word-class, and MorDebe will create the full set of inflectional forms using an internal inflecting algorithm. Before the word-forms are added to the database, all inflectional forms are presented for manual verification, firstly to check whether the suggested inflection is actually correct (and not unpredictably irregular), and secondly to check whether the full set of inflections should be added (i.e. whether the lemma is not defective).

And finally, the word can be added to the neologism candidate by using the fill-in form on the left-hand side – the orthography and the context are automatically provided, but can be adjusted manually if necessary. The other fields have to be manually filled in. Along with the neologisms, MorDebe automatically stores the corpus of origin, and the author of the neologism record.

7.2. NeoTrack and MorDebe

The extended lexicographic diachronic criterion requires a full-form lexicon, in which not just the lexemes, but the entire inflectional paradigms are listed. Although this full-form lexicon could in principle just be a list of words, NeoTrack uses a more structured morphological database set-up called *MorDebe.* MorDebe is a simple database structure with two tables, one for the lemmas, and one for the word-forms, which are the inflectional forms of the lemmas. Each word-form represents a single inflectional form of the lemma, which means that if two inflectional forms are homographous, they are still listed as separate entries.

Since the extended criterion is based on dictionaries, the MorDebe database is initially filled with the full-form version of one or more dictionaries. For Portuguese, MorDebe was originally filled with the list of lexemes from the *Porto Editora* dictionary, inflected semi-automatically using a combination of manual verification and computer techniques such as bootstrapping and cross-verification. The resulting database is used in a number of different ways, and still under constant revision to filter out any possible remaining errors.

Semi-automatic detection necessarily deals with formal neologisms rather than semantic neologisms, since semantic interpretation cannot be done automatically. But existing semi-automatic systems are even incapable of detecting formal neologisms: they can only used string-based neologism detection. This means that if the word *rubber* would be used as a verb rather than a noun, a system like NeoTrack is unable to detect it. This means semi-automatic detection concerns string-based neologism, which is an even more restricted notion.

But NeoTrack does not apply strict string-based detection: the use of *rubber* as a verb can be detected if it used in one of its inflectional forms *rubbering* or *rubbered*. When the string *rubbering* is detected as a neologism in NeoTrack, what is added to MorDebe is the inflectional paradigm for *rubber,* and not the individual string. This means that what is detected with NeoTrack is partially lexeme based, in the sense that a lexeme can be detected by any inflectional form which is not homographous to an existing inflectional form. This mixed form which lies between formal neologism detection and string-based neologism detection is what I refer to as *orthographic neologisms.*

The main reason for working with a structured morphological database rather than a simple full-form lexicon list is the inherent value of a large, verified morphological database. Although MorDebe does not give any semantic information, it does provide information on how the existing words of the language should be written – which is one of the most frequent ways in which dictionaries are used: *"over 80% of the consultations of a monolingual dictionary concern checking the existence and/or the spelling of a word."* (Oppentocht & Schultz, 2003: 224). This function of dictionaries is more easily fulfilled by a morphological database than by a paper dictionary, since a morphological database has less restrictions in

size, fully specifies all word-forms rather than rely on the users linguistic abilities, and with the constant use in neologism detection is more up-to-date than any paper dictionary can ever be.

The MorDebe database resulting from the neologism detection efforts is seen not just as a by-product, but as an important goal of the research – the MorDebe database is currently being prepared for on-line consultation, and will be provided as on open source reference lexicon for academic research.

7.3. Inflection and Derivation

The relation between neologism and derivation is not entirely straightforward. The word *disturbable*, although a correct English word in a certain sense of the word, is considered neological because it is not an actual word of the English language, even though it is a *potential* word. Guilbert (1975) refers to this type of neologism as *néologie de langue*, since the neologism does only relate to actual language use (*langue*), and not to the language capacity (*langage*)[14]. Regular derivations are semantically transparent, and because of their transparency, regular derivations are commonly only listed in dictionaries if they attract unpredictable senses, or are frequently used but only in a restricted meaning.

Although regular derivations are transparent they are not fully predictable. There can be alternative ways of forming the derivation, where the existence of one blocks the alternatives: *refuse – refusal, confuse – confusion, perturb – perturbation, disturb – disturbance, confer – conferral, refer – reference*. And for not every lexeme all derivations are possible. Therefore, derivations are not automatically added to MorDebe – only those derivations that are actually encountered in corpora are added.

For inflections, on the other hand, all forms are included in MorDebe. Even though many uncommon words will never be used in the 2nd person singular, these forms are nonetheless taken to exist, accept if their existence is explicitly overridden (in the case of defective verbs). This means that the distinction between inflection and derivation is crucial for the definition of neologisms. However, this distinction is much disputed – several author claim there is no strict separation (Bybee, 1985), propose a category of inherent inflection which is halfway between inflection and derivation (Booij, *to appear*), or propose a whole system of intermediate classes called *transpositional morphology* (Bauer, 2004). In MorDebe, all marginal cases are seen as derivational (Janssen, 2005), leaving only a very selective set of forms as inflectional.

From the perspective of neologism detection, taking a minimal stance to inflection is beneficial – allowing the detection of neologisms that would otherwise be missed. Consider the Portuguese word *igualíssimo* (most equal), which is the superlative of the adjective *igual*, and superlatives commonly considered inflectional. However, *igualíssimo* is considered a neologism by the ONP: it is a word that was previously semantically blocked, since *equal* is not a gradable adjective. There is no prior corpus evidence for the term, but is recently used in Portuguese newspapers – a meaning shift similar to the relatively recent *whiter* in English. Although the same effect would have been reached by listing *igual* as a defective adjective lacking a superlative, but modelling *igualíssimo* as a derivation makes their treatment in the database more transparent.

7.4. Error and Variation

Before the judgment whether a neologism candidate is a neologism or a database-gap, the typographic errors are filtered out. In many cases, it is clear whether a word is correct or a typo. However, there is no strict separation between incorrect spellings and spelling variations. Any neologism project should clearly define which sources and methods are followed to distinguish the typographic errors from spelling variations in these borderline cases. Threre are at least three large classes of borderline cases.

Derivations

The correct deverbal noun for *perturb* is *perturbation*. However, one can quite commonly en counter the form **perturbance*, in accordance with the more frequent *disturb*. Even the more erratic form **perturption* can be found occasionally. The word *perturbance* is not an accidental typographic error - it is even included in certain dictionaries. But it is a matter of opinion whether it should be treated as an error, or a lexicalised form.

Loanwords

There often is a lot of spelling variation for new loanwords – both before and after their incorporation in the dictionary. The Portuguese word *guiché* (window; wicket) can also be written as *guichê* or in its non-adapted form *guichet* and all three forms are included in the *Academia* dictionary, the word *mítingue* as an adapted spelling of *meeting* is hardly ever found, but is still the preferred spelling according to the *Academia*. Because of this variance, different adaptations should be considered correct words, but within limits: the once loanword *líder* should now be considered wrong when written as *leader*. But for less frequent words this

is much less clear: the correct spelling of the English (or Dutch) loanword *afrikaner* in Portuguese is *africânder* or *africâner* – but the English spelling *afrikaner* is still found in newspapers, even though it is no longer included in dictionaries.

Compounds

Compounds can often be written as one word, as a multiword expression, or as a hyphenated construction. Dictionaries do not always agree upon the correct spelling of such compounds, and also do not always follow the most commonly used form. There is no clear way of distinguishing typos from spelling variations in these cases, hence any neologism observatory should define a set of guidelines to follow for these cases.

8. Conclusion

The notion of a neologism has often been considered as indefinable, leaving the notion of a neologism as a subjective and arbitrary label. Although there are several ways in which the notion is arbitrary and conventional, it is nevertheless possible to establish a relatively well-defined criterion for when a word should be counted as a neologism. However, the two most commonly used criteria have serious limitations: the dictionary-based criterion does not properly define a notion of newness, and the corpus-based criterion does not provide any kind of control. I hope to have shown in this article that these limitations can be overcome by the extended lexicographic diachronic criterion.

The extended lexicographic diachronic criterion is a hybrid between a corpus based and a dictionary based criterion and comes down to the following: initially, a morphological database is created by creating a full-form version of one or more lexicographic sources. Any word not occurring in the morphological database is a neologism candidate, and neologism candidates can be errors, proper neologisms or gaps in the database. Whether a word is a gap or a proper neologism is determined by manual verification in an exclusion corpus – gaps are those words that do occur in the exclusion corpus sufficiently frequently as a proper word.

There are several arbitrary parameters in this criterion: the source used for the initial content of the morphological database, the constitution of the exclusion corpus, the threshold occurrence frequency, and the threshold neologism period. But with these parameters properly defined, the extended lexicographic diachronic criterion not only gives a well defined criterion for being a neologism, but a criterion that seriously defines a notion of newness.

Despite the obvious limitation of the strict dictionary-based criterion, there is a reason why it is very frequently used: neologism research is often initiated and/or funded by dictionary publishers - "*One of the most widespread uses of large corpora of contemporary language is to identify changes in vocabulary*". (Aston & Burnard, 1995, p. 51). But from an academic perspective, it is much more productive to make use of the NeoTrack system, since it not only yields a database of neologisms, but a large-volume, high-quality, and up-to-date morphological database at the same time. For Portuguese, the MorDebe database is already the largest repository of words – with close to 1,5 million word-forms listed.

Notes

[1] A newly created linguistic expression which is considered known by at least part of the linguistic community. (my trans.)

[2] This more occasional notion of a neologism might more correctly be called a *neologistic occurrence*: the occurrence of a word in a language which is new. But for clarity, we will stick to the term *neologism* – despite its political associations.

[3] Rey does much less object to historic labels like "since 1744", which do require a notion of words existing before that time, but does not have the arbitrary demarcation of a period for newness.

[4] The APRIL system uses a definition based merely on frequency – given the correlation between rare words and neologisms. But since this is only an indirect attempt at defining a neologism, we will ignore that option here.

[5] The term *exclusion list* will be used for the simple list of terms – the term *exclusion corpus* will be reserved for the body of text from which such a list may be derived.

[6] Sablayrolles (to appear) gives a good overview of the problems of selecting dictionaries for the exclusion corpus in neologism research.

[7] This because of the regular polysemy between mass nouns and count nouns. This matter is further complicated by the fact that some dictionaries list *vodka* as polysemous, and others as homonymous.

[8] "They rained criticisms on the film", *i.e.* criticism on the film was abundant. Example provided by Margarita Correia. These examples are different from creative expressions such as *I thundered* when uttered by Zeus.

[9] These homographs directly affect neology: the occurrence of *contravou* in Portuguese would be neological despite the fact that its citation form *contrair* (to contract) does exist – but not as a compound of the irregular verb *ir* (to go).

[10] Although this number may be debatable: Lemnitzer (2003) only reports a 3,87% of typos in the WortWarte results – which must be due to different ways of counting typos. Our own counts lie around 10%.

[11] It is impossible to define a fact and a concept (like neologism and neology) with recurrence to the extra-linguistic tools that dictionaries are (my trans.)

[12] The problem with a frequency threshold is that it mixes up the notion of neologism and low-frequency word again – a true verification would need to take the expected frequency of the word into consideration – which is however hard to do in practice.

[13] Theoretically, if all texts are taken into account, words can only be neologistic if their first occurrence happens to be in the text under investigation.

[14] The difference between possible and actualised words has been argued for from many different perspectives, including the psychological model by Meijs (1985).

References

Aston, Guy and Lou Burnard. 1996. *The BNC Handbook: Exploring the British National Corpus with SARA.* Oxford University Computing Services.

Bauer, Laurie. 2004. The Function of Word-Formation and the Inflection-Derivation Distinction. *In:* Aertsen, Hannay & Lyall (eds.) *Words and their Places. A Festschrift for J. Lachlan Mackenzie.* Amsterdam: Vrije Universiteit.

Booij, Geert. 2003. The Codification of Phonological, Morphological, and Syntactic Information. *In:* P. van Sterkenburg (ed.) *A Practical Guide to Lexicography*. Amsterdam: John Benjamins Publishing.

Booij, Geert. *To appear.* Inflection and Derivation. *In:* Brown (ed.) *Encyclopedia of Language and Linguistics*. Oxford: Elsevier.

Bußmann, Hadumod. 1990. *Lexicon der Sprachwissenschaft.* Stuttgart: Kröner.

Bybee, Joan L. 1985. *Morphology: a study of the relation between meaning and form*. Amsterdam: Benjamins.

Cabré, M. Teresa. 1999. *Terminology: Theory, methods, and applications.* Amsterdam: John Benjamins Publishing Company. *Translation of: La Terminología: Teoría, metodología, aplicaciones*. Barcelona: Editorial Antártida, 1992.

Curzan, Anne. 2000. The Compass of the Vocabulary. *In:* Mugglestone (ed.) *Lexicography and the OED. Pioneers in the Untrodden Forest*. Oxford: Oxford Universit Press.

Guilbert, Louis. 1975. *La Créativité Lexicale.* Paris: Librairie Larrousse.

Janssen, Maarten. 2005. Between Derivation and Inflection: paradigmatic lexical functions in morphological databases. *In: Proceeding of the 2nd Internation Conference on Meaning-Text Theory (MTT2005).* Moscow, Russia.

Janicijevic, Tatjana & Derek Walker. 1997. NeoloSearch: Automatic detection of neologisms in French Internet documents. *In*: *Proceeding of Joint International Conference ACH-ALLC'97*, Queen's University, Kingston, Ontario, Canada.

Lemnitzer, Lothar. 2003. Ist das nicht doch alles das Gleiche? Regeln und Distanzmaβe zur Berücksichtigung orthographischer Idiosynkrasien bei der Abbildung von Zeichenketten auf lexicalische Einheiten. *In:* Cyrus, Feddes & Schumacher (eds.) *Sprache zwischen Theorie und Technologie. Festschrift für Wolf Paprott zum 60. Geburtstag.* Wiesbaden: DUV.

Meijs, Willem J. 1985. Morphological meaning and the structure of the mental lexicon. *In*: T. Weyters (Ed.), *Meaning and the lexicon*. Dordrecht: Foris Publications.

Oppentocht, Lineke & Rik Schutz. 2003. Developments in Electronic Dictionary Design. *In:* P. van Sterkenburg (ed.) *A Practical Guide to Lexicography.* Amsterdam: John Benjamins Publishing.

Rey, Alain. 1995. The Concept of Neologism and the Evolution of Terminologies in Individual Languages. In: Sager (ed.) *Essays on Terminology*. Amsterdam: John Benjamins Publishing. Translation of: *L'aménagement de la Néologie*. Office de la language française du Québec. January 1975, p. 9-28.

Roche, Sorcha & Lynne Bowker. 1999. Cenit: Système de Détection Semi-Automatique des Néologismes. *Terminologies Nouvelles,* vol. 20.

Sablayrolles, Jean-François. *to appear*. Neologie et Dictionnaire(s) comme Corpus d'Exclusion. *Actes de la Journée des dictionnaires de l'Université de Cergy*. Cergy-Pontoise, France.

Vivaldi, Jordi. 2000. SEXTAN: prototip d'un sistema d'extracció de neologisms. *In*: Cabré, Freixa, &. Solé (eds.) *La Neologia en el tombant de segle*, Barcelona: IULA. pp. 165-173.