M. Grochocka

EXTRACTING NEOLOGISMS FROM A CORPUS USING *NEODET*

Abstract. *NeoDet* is an application which has been developed in order to automatise the process of neologism extraction from a corpus. It serves three main functions. Firstly, it is used to compile a study corpus of journalistic texts out of which new lexical items could be extracted. Next, it semi-automatically extracts formal neologisms from the study corpus following a procedure based on the exclusion principle. Finally, with its help it is possible to manage the database of the extracted neologisms.

1. NeoDet – neologism detector

NeoDet is a web-based application which has been created with the aim of making the process of neologism detection as automatic and objective as possible. The researcher commissioned its development to Wiktor Latanowicz, an IT specialist. The application is language independent and works on a corpus not marked morphologically. Generally speaking, it works on the exclusion list principle, just as most such tools do, e.g. *Cenit*, *NeoTrack*, and *BuscaNeo*. In other words, it operates according to the lexicographic definition and the corpus-based definition of a neologism: a word is regarded as a neologism (to be more precise, a neologism candidate) if it occurs neither in the existing dictionaries nor in a reference corpus.

The exclusion list principle is simple. First, *NeoDet* creates a list of all the word forms occurring in the study corpus, i.e. a corpus from which neologisms are to be extracted, and then checks if the words appear in the exclusion sources consisting of words deemed as established in the language. The study corpus is a corpus of journalistic texts comprising most widely read British daily broadsheets and tabloids. The exclusion sources include: *The British National Corpus* (BNC), general online dictionaries (British and American), dictionaries of slang, as well as word lists of proper names and geographical names. A combination of exclusion sources is used

in order to make the filtration process more efficient, yielding much less noisy output. The words from the study corpus are looked up in the BNC and the online dictionaries in the same way as if they had been typed in the dictionaries' search engines by a dictionary user. For example, if the plural form of a noun or the past participle of a verb is looked up, the dictionaries recognize the form and redirect the user to the basic form of the word. Hence, it can be said that the dictionaries are equivalent to a morphological database, not to mention the fact that a huge part of the filtration of inflected forms is done by the BNC. It needs to be added that occurring in just one of the exclusion sources disqualifies a word from being a neologism candidate. Only words which are not found in any of the sources are classified as neologism candidates to be subjected to further analysis.

Another important point that must be made is that this method works well exclusively for extracting formal neologisms and borrowings. As stressed by Janssen, any semi-automatic method of neologism detection is deficient when confronted with the task of extracting semantic neologisms, for the identification of which human intervention is indispensable. By the same token, pragmatic neologisms cannot be detected, either. What can be successfully fished out is formal neologisms, or to be more precise, orthographic neologisms. As the name suggests, in order to extract a list of neologisms, the system relies solely on the orthography and, in consequence, it does not recognize word classes. Nonetheless, it needs to be noted that since different word classes inflect in a different way, a neologism can be detected by its inflected forms with the proviso that they do not happen to be homographic with already existing inflectional forms.

2. The functions of *NeoDet*

NeoDet consists of four major components connected with the process of analysing neologism candidates, i.e. *Neologisms*, *Candidates*, *Exceptions*, and *Words*. The *Neologisms* section contains a list of all the words which have been marked as neologisms, whereas the *Candidates* section provides a list of all the items which have not

been found in the exclusion sources, including noisy output. All noisy output in the form of non-neologisms absent from the exclusion sources is put under the *Exceptions* option and classified depending on the type of noise it represents into one of the following categories: *Number*, *Typo* (a typographic error), *Proper name*, *Paralanguage*, *Non-word*, *Citation* (an item is part of a citation in a foreign language), *e-mail/www*, *Wordplay*, *Term*, *Slang*, *WF* (short for «word formation» and referring to semantically transparent composite words), and *Other* (if an item cannot be subsumed under any of the available categories). The *Exceptions* section also comprises those items from the noisy output which, though absent from the exclusion sources, are established words, and as such should be added to the exclusion sources. The *Words* section provides access to a list of all the words appearing in the database which have been found in at least one of the exclusion sources, and hence are dictionary words.

NeoDet also contains two sections referring to the study corpus, these being the *Add article* and *Articles* sections, as well as a search engine (*Search*) and the *Statistics* section. The *Article* section is where all the articles can be accessed individually and the search can be filtered by the title, the newspaper, the newspaper section, the date of publication, and the date the article was added.

The *Search* option is a search engine which makes it possible to put queries about any string of letters that may occur in the study corpus, which is very useful when carrying out morphological analyses. The option *Search as prefix/infix/suffix* enables one to seek items beginning with, incorporating, or ending with a particular string of letters. To give an example, if the three-letter string «vuv» is queried «as infix», apart from providing all the occurrences of the word itself, the results will also contain, e.g. *anti-vuvuzela*, *vuvu-lover*, *vuvuzela*, and *vuvuzela-free*. Additionally, one can look not only for single strings, but also for whole phrases.

In the *Statistics* section information is provided on the total number of types and tokens occurring in the corpus, as well as the number of tokens that have been classified as neologisms. From this section one can also learn how many articles have been uploaded and

what the figures are regarding the number of words per newspaper and per section. In addition, the average article length for each newspaper is calculated. All the figures are also presented in the form of bar charts.

It is important to note that the study corpus can be subdivided into smaller subcorpora on the basis of the date of publication of the articles, giving the opportunity to analyze data only from a limited, strictly defined period of time. This possibility is provided by the *First appearance between ... and ...* option. Such an option may prove useful for drawing extralinguistic conclusions from the data, such as matching sudden, frequent occurrences of a given neologism with a particular event. Finally, the *Minimum occurrences* and *Minimum articles* options make it possible to make more restricted searches by defining the minimal number of tokens and/or articles in which a given lexical item occurs. These functions are applicable to both neologism candidates and items already marked as neologisms.

3. Neologism management

The *NeoDet* application enables the researcher to manage the database of neologisms by indicating the citation form of each new word, its possible typographic variations and syntactic category. Furthermore, the type of a neologism can be defined. For instance, *ultraportable* can be marked both as an example of prefixation and suffixation. In addition, the meaning of a neologism can be provided in the *Definition* box and comments can be made in the *Notes* box.

Sometimes the meaning of a new word is explicitly explained in the article/s it comes from, which is actually to be expected, especially when dealing with a very recent neologism. However, it is not always the case and sometimes the meaning has to be elicited from the context. As the context in which a given neologism appears in the study corpus may prove insufficient to arrive at a satisfactory definition, links are provided to the *Google* and *Wikipedia* search engines, as well as to the *WebCorp* tool, a linguistic search engine that can be used to get access to the World Wide Web as a corpus. Thanks

to that, it is possible to establish the meaning of a neologism by means of looking at more ways in which it is used on the Internet.

4. Problems

A few problems have been identified which may influence the effectiveness of *NeoDet* as a semi-automatic neologism detection tool. Firstly, the online exclusion sources sometimes fail to respond to the queries made by *NeoDet* due to technical reasons. Another problem is the fact that items which do not have headword status but are offered as examples in the online dictionaries belonging to the exclusion sources are not discarded as dictionary words, but appear on the neologism candidates list. Moreover, an item may sometimes occur in the neologism candidate list, even though it was in fact present in the exclusion sources, due to an alternative spelling not provided in the exclusion sources (e.g. *micro-blog* vs. *microblog*). Another spelling-related problem concerns examples such as *G & T* (gin and tonic). The abbreviation is identified by *NeoDet* as a potential neologism due to the fact that it is spelt with spaces in between characters, whereas in the exclusion sources it is either spelt as one word, or the ampersand is replaced by *and*.

5. Conclusion

The *NeoDet* application fulfils its roles quite effectively. Not only can it be used to compile the study corpus, but it can also perform the task of neologism detection quite successfully by means of checking the word forms from the study corpus against a number of exclusion sources. Additionally, it helps the researcher to analyse and manage the obtained database of formal neologisms. All its merits notwithstanding, *NeoDet* is not without limitations, therefore, there is still room for improvement when it comes to the IT side of the application, but not only. Extending the exclusion sources, e.g. by including an encyclopaedia, could be yet another step towards the enhancement of the efficacy of the software, as it would definitely

lead to producing less noisy output and, in consequence, there would be fewer invalid neologism candidates to verify.