Using Neoloog to detect and describe neologisms in online dictionaries

Vivien Waszink, Instituut voor de Nederlandse Taal (INT), Leiden, The Netherlands

Abstract

Every year, thousands of neologisms, or new words, are coined. Most neologisms are composite forms or derivations. The more widely adopted or firmly rooted neologisms are often described in dictionaries, especially the opaque ones, for example in the Algemeen Nederlands Woordenboek (ANW), an online dictionary of present-day Dutch. Lexicographers are always on the lookout for new words themselves, but is it also possible to detect potential neologisms automatically or semi-automatically by a computer program?

Yes, it is. In this paper I will show how potential neologisms in Dutch can be detected with the aid of the computer tool Neoloog. Neoloog (2015) was designed especially for this purpose at the Instituut voor de Nederlandse Taal (INT) by computer programmers Jan Niestadt, Rob van Strien en Mathieu Fannee for intern use as an aid for lexicographers working on the ANW and the Neologismenwoordenboek, a dictionary of neologisms.

Firstly, I will describe the corpus of Dutch newspapers which is used in Neoloog, consisting of the Dutch newspaper NRC and the Flemish-Dutch newspaper De Standaard. This corpus is updated every month. Secondly, I will go into the design and the functions of the tool Neoloog and I will present the possibilities for lexicographers to mark words as neologisms. The ‘new words’ used in the above-mentioned newspapers in the corpus are presented in lists in Neoloog and it is possible to view all the listed words in their contexts as well. This lists of potential neologisms always have to be checked manually by lexicographers, because not every word that is used for the first time in a Dutch newspaper is a neologism. Finally, I will show new features of Neoloog which make it possible to add new lemmas to the ANW and the Neologismenwoordenboek automatically. In Neoloog it is also possible to add word definitions and grammatical and linguistic information to the words which are marked as neologisms.