Semi-Automatic Detection of New Words in Georgian

Tamar LALUASHVILI (tamar.laluashvili.1@iliauni.edu.ge), Ilia State University, Tbilisi, Georgia
Tinatin MARGALITADZE (tinatin.margalitadze@iliauni.edu.ge), Centre for Lexicography and Language Technologies, Ilia State University, Tbilisi, Georgia

The present study is a part of the three-year project, dedicated to the comprehensive study of neologisms in the Modern Georgian language. The work is supported by Shota Rustaveli National Science Foundation of Georgia (FR-23-4304). This paper aims to present the methodology worked out at Ilia State University for the semi-automatic detection of new words in Modern Georgian.

Changes, taking place in a language are usually very slow and difficult to notice. These changes occur over long periods before they become perceptible on a synchronic level. But there are exceptions from this general tendency and the contemporary Georgian language is a very good example of this. Currently, Georgia and the Georgian language are in an interesting era from a historical point of view. After the collapse of the former Soviet Union, our country saw the emergence and rapid development of a free market economy, multiparty political system, private banking sector, the Georgian national armed forces, etc. There developed intense relations with foreign countries on diplomatic, as well as on educational, economic and merely personal levels. These processes are only intensified by the increasing availability of computer, telecommunication and mobile technologies. Consequently, all prerequisites are in place which cause considerable changes in the Georgian lexis. The detection and study of latent diachronic processes taking place in the Georgian language becomes especially interesting at this linguistic-historical moment.

The study of neologisms is particularly relevant for lexicography, which, in addition to studying the theoretical aspects of the issue, also serves purely practical purposes. It involves updating existing dictionaries, adding new words, and assigning new meanings to the already existing ones. Contemporary users evaluate the quality of dictionaries by their ability to keep pace with the latest vocabulary and meanings. Dictionaries that fail to capture modern vocabulary tend to lose their appeal and popularity. The study of neologisms is also important as it reveals the lexical creativity of a language at a certain stage of its development.

Neologisms introduced into the Georgian language have been the focus of many Georgian researchers. However, there was no established methodology for identifying new words in modern Georgian. Therefore, we decided to work out such a methodology and for this purpose, we studied existing methods of detecting new words for other languages (Cabré & Nazar 2012; Janssen 2009; Kernerman & Klosa-Kückelhaus 2021) and formulated our approach.

For this project, a special corpus was composed including textual material from Georgian-language online newspapers and magazines, news websites, social media sites, websites of various governmental and non-governmental organizations, Georgian Wikipedia and some other sites, covering the last 15 years. Our methodology for the identification of Georgian new words is based on the combination of several approaches. We have applied the lemmatization tool for the Georgian language developed at Ilia State University. The lemmatizer is available on the website of the University at the URL https://qartnlp.iliauni.edu.ge (Lobzhanidze 2021). The lemmatizer relies upon the dictionaries, integrated with it, which are mainly composed of the word lists of the 8-volume *Explanatory Dictionary of the Georgian Language* and various other Georgian normative dictionaries. The lemmatization tool is capable of lemmatizing only those lexical units, which are included in these resources and serves as a kind of exclusion source allowing us to look

for neologism candidates in unlemmatized, out-of-vocabulary (OOV) lexis. After the tokenization and lemmatization of the corpus material, the OOV lexical units are subjected to analysis and the potential neologisms are sampled. In order to reduce the number of OOV lexis, we applied to it another lemmatizer, developed by Paul Meurer (2014) for the Georgian National Corpus (GNC, Gippert & Tandashvili 2015). The GNC contains over 200 mln tokens and P. Meurer's lemmatizer can recognize much more words than the one applied by us for the first phase as an exclusion source. As a result, the number of unlemmatized vocabulary was reduced by 45 %.

Georgian word embedding software proved to be very instrumental in finding additional neologism candidates (https://wordembedding.spellchecker.ge).

As a result of the study, we have identified 1000 lexical neologisms in Modern Georgian. The study of semantic development of existing words is planned for the next phase within the framework of the grant project. We mostly selected words that belong to the common vocabulary of Georgian, including colloquial words and slang. We also selected some general terms from the fields of social media, online media, tourism. New terms from specialized fields were excluded from the study.

Based on the research carried out: (a) an online dictionary of neologisms will be composed and published; (b) the methodology and tools for the detection of neologisms will be perfected; (c) a website will be set up for the monitoring of neologisms in future which will be published on the website of Ilia State University.

References

Cabré, M.T. & Nazar, R. 2012. Towards a New Approach to the Study of Neology. *Neologica*, Classiques Garnier. N 6, p. 63-80.

Gippert, J. & Tandashvili, M. 2015. Structuring a diachronic corpus. The Georgian national corpus project. Gippert, J., Gehrke, R (eds). *Historical Corpora. Challenges and Perspectives. Corpus Linguistics and Interdisciplinary Perspectives on Language,* v. 5. Tübingen: Narr.

Janssen, M. 2009. Detección de Neologismos: una perspectiva computacional. *Debate Terminológico* 5, 68-75.

Kernerman, I. & Klosa-Kückelhaus, A. (eds.) 2021. Lexicographic Neology and Neological Lexicography. Special issue of *International Journal of Lexicography,* 34/3.

Lobzhanidze, I. 2021. *Principles of Morpho-Syntactic Annotation and Morphological Analysis of the Finite State*. Tbilisi: Ilia State University Publishing (in Georgian).

Meurer, P. 2014. The Morphosyntactic Analysis of Georgian. *Georgian National Corpus.* http://gnc.gov.ge/gnc/page