Automatic detection of neologisms in Russian newspaper corpora with Neoveille

Tatiana Iakovleva

5. Conclusion

A project focusing on a multifaceted typology of MWEs was presented. The typology mainly focuses on the description of morphological, syntactic, semantic, lexical and pragmatic idiomaticity. Also, a MWE lexical database reflecting this typology is being developed. The MWEs contained in the database are extracted from corpora of synchronic Czech and they will be used, i.a., for the improvement of parsing of Czech.

References

1. *Cvrček V.* (2014), Kvantitativní analýza kontextu (Quantitative analysis of context). Praha, Nakladatelství Lidové noviny.
2. *Čermák F. et al.* (1983–2009), Slovník české frazeologie a idiomatiky 1–4 (Dictionary of Czech Phraseology and Idiomatics, SČFI 1–4), Praha, Academia / Leda.
3. *Evert S.* (2004), The Statistics of Word Cooccurrences: Word Pairs and Collocations. PhD dissertation, IMS, University of Stuttgart. Published in 2005. Available at: http://www.collocations.de.
4. *Jelínek T.* (2016), Partial Accuracy Rates and Agreements of Parsers: Two Experiments With Ensemble Parsing of Czech. In: Brejová B.(ed.), ITAT 2016: Information Technologies–Applications and Theory Proceedings. Tatranské Matliare, Slovensko, 42–47.
5. *Pecina, P.* (2010), Lexical association measures and collocation extraction. In: Language Resources and Evaluation, 44 (1–2), 137–158.
6. *Petkevič V., Skoumalová H.* (2015a), The utilisation of valency dictionaries in creating a large Czech treebank. In: Prace Filologiczne, LXVII: 261–277.
7. *Petkevic˘ V., Rosen A., Skoumalová H.* (2015b), The grammarian is opening a treebank account. In: Prace Filologiczne, LXVII: 239–260.
8. *Skoumalová H., Rosen A., Petkevič V., Jelínek T., Vítovec P., Znamenáček J.* (2014), A grammar-licensed treebank of Czech. In: Henrich V., Hinrichs E., de Kok D., Osenova P., Przepiórkowski A. (eds.), International Workshop on Treebanks and Linguistic Theories (TLT13) (December 12–13, 2014, Tübingen, Germany). University of Tübingen, Tübingen 2014, 218–229. Available at: http://tlt13.sfs.uni-tuebingen.de/tlt13-proceedings.pdf.

Milena Hnátková,
Vladimír Petkevič,
Hana Skoumalová
Charles University, Prague (Czech Republic)
E-mail: first_name.surname@ff.cuni.cz

AUTOMATIC DETECTION OF NEOLOGISMS IN RUSSIAN NEWSPAPER CORPORA WITH NEOVEILLE

Abstract. Neoveille is a web platform that automatically detects new words and monitors word usage change in seven languages [Cartier 2016, 2017]. The platform allows to select corpora, to automatically detect neologisms, to describe them linguistically and to follow their life-cycle. This paper focuses on corpus-based automatic neologism identification in Russian and describes broad tendencies in novel word formation processes. We focus on borrowings and morpho-semantic novel items.

Keywords. Neologisms, natural language processing, corpus linguistics, Russian, word formation.

1. Presentation of the "Neoveille" platform

In the context of globalization, a growing number of studies focus on how English influences the morphological, syntactic and orthographic systems of various languages, including Russian [Galtseva 2014; Rochtchina 2012; Rybushkina 2015]. These studies mainly examined borrowings which were reported to be the largest group of neologisms in modern Russian. To the best of our knowledge, Russian neologisms that are partially or fully composed of native (as opposed to borrowed) linguistic elements, received less attention [Zhdanova and Raciburskaya 2015].

The "Neoveille" platform [2016, 2017], supported by the IDEX-ANR grant, automatically detects new formal and semantic neologisms, regardless of whether they are composed of foreign lexis or of native linguistic items. Although neologism detection platforms such as NEOROM exist for Latin languages [see Humbley 2008 for review], the Néoveille platform is the first of its kind to encompass typologically different languages (Chinese, Czech, French, Greek, Russian, Polish, Portuguese) and to include Slavic languages. Moreover, it is the first platform to propose an automatic detection of semantic neology. The platform provides textual data that can be used for several purposes. Not only is it an on-line dynamic database that monitors neologisms emergence and lifecycle but also a monitor corpora search engine. The extracted data may also enrich on-line lexical resources, such as embedded reference language dictionaries. The following section describes the Néoveille platform focusing on the formal neologism detection, analysis and monitoring.

2. Stages of neologism analysis on Neoveille

2.1. Automatic detection of neologisms

Monitored Russian corpora are currently composed of around 50 newspapers representing general Russian language in journalistic discourse (https://lenta.ru/rss; NEWSru.com, http://izvestia.ru, among others). The Néoveille web platform enables linguists to manage their corpora (via adding, modifying and suppressing), validate or invalidate the automatically detected formal neologisms, describe them linguistically and then follow their lifecycle on monitor corpora.

Linguistic items as well as meta-data (newspaper title, author, theme and date) are automatically extracted via the newspapers' RSS links on a daily basis, three times a day. A specific program is used to extract the relevant text from html pages (https://pypi.python.org/pypi/jusText).

The neologism detection program follows four steps. First, it performs a morphological analysis to identify unknown words. We use the Treetagger [Schmid, 1994] with the language model designed by Sharoff et al. [2008]. This POS-tagger will mark the unknown words with a specific tag. A second step is performed by Hunspell spell-checker, aiming at checking if unknown words are typographic errors or not. Third, the neologism candidates are compared to a complementary exclusion dictionary, fed by linguistic experts. Finally, the resulting Neologism Candidates (CN) are analyzed by linguistic experts who either confirm their neologism status, or classify them as words belonging to a reference dictionary, a terminological lexical unit or to other categories of words to exclude (e.g. typographic mistakes). This excluded dictionaries enable to considerably improve the automatic detection process, as they are automatically re-used by the automatic detector.

2.2. Manual analysis of candidates for neologisms

The detected and validated database of neologisms for Russian currently contains around 460 items.

Linguists classify each neologism according to a typology designed by Pruvost and Sablayrolles [2016]. At the current stage, automatic detection on Néoveille targets three categories of neologisms in Russian: loanwords/borrowings, morpho-semantic novel words and syntactico-semantic words. The present paper focuses on the first two categories. According to the typology, morpho-semantic novel words include the following sub-categories: affixation (prefixation, suffixation or parasynthesis), inflexion and composition. In the present paper, we will not discuss inflexion and parasynthesis, as these word formation processes are represented by less than 10 occurrences.

3. Neologism Classification

3.1. Loanwords

In line with previous research on Russian neologisms, loanwords represent the largest group among neologisms (49 %). Some loanwords come from Arabic or French, e.g. *дезавуировать* (from French *désavouer*) 'renounce (one's claims)'. English is the major source of borrowing. Overall, loanwords vary in the use of script(s). Detected words are written in either (1) Cyrillic script, or (2) Roman script, or (3) as orthographic blends: (1) *сити-кар* 'city car', *аквафермер* 'aquafarmer'; *тег* 'tag'; *вейпинг* 'vaping, that is, using e-cigarettes'; *суперфуды* 'superfoods' (87 %); 2) *machine-learning; seal-watching* (9 %), *Наблюдение за тюленями — это отдельный вид туризма. Он называется seal-watching. (http://murmansk.mk.ru);* 3) *youtube-канал* 'youtube channel' (4 %).

3.2. Prefixation

Prefixation is a relatively infrequent type of morpho-semantic word formation (15 %). Although foreign prefixes are more frequent in novel word formation than native ones (26 vs. 17 respectively), the latter are more frequent in the context of competition (e.g. *лже-* 'pseudo-' vs. *псевдо-* 'pseudo-'). The most productive foreign prefixes are *экс-* 'ex-' (e.g. *экс-работник* 'former employee') and *анти-* 'anti-' (*антитеррористический* 'counterterrorist'). The most productive native prefixes are *лже-* 'pseudo-' (*лжесайт* 'pseudo website') and *не- 'non-'* (*недострой* 'unfinished construction site').

3.3. Suffixation

Suffixation is almost twice as productive as prefixation (28 %). A little more than a half of suffixed words are formed with foreign roots, mostly adjectives, e.g. *тюбинговый* <tubing+adjectival suffix -ov> 'related to tubing'. In contrast, words formed with native roots are mostly nouns, e.g. *маршрутчик*'mini-bas driver'.

Results show that formation via foreign suffixes is rare (N=5), e.g. *скутерист* <scooter+ist> 'scooter driver', *зацепер* 'train surfer'.

3.4. Compounds

Compounds represent the largest group of morpho-semantic word creation (31 %). We broadly divided compounds in three groups:

- synthetic <adjective + noun> compounds with gender and number agreement (40%), *инновационная еда* 'innovative (Singular Feminine) food (Singular Feminine)'.
- analytical compounds with no number or gender agreement (35%), the components being either linked with a hyphen (a), or presented as a single word (b), (a) *кафе-кальянная* <café(noun)-hookah> 'hookah bar/lounge'; *директор-распорядитель* 'managing director' (b) *автоподход* <auto(mobile) access> 'space allowing a certain place to be accessed by car'; *электровелосипед* 'e-bike, that is, a bicycle with an integrated electric motor'.
- <noun + noun> combinations denoting new objects or concepts (24%), *автомобили smart особо малого класса* <smart cars of a particularly small class/size>, *либерализация визового режима* 'viza regulation liberalisation'; *технология слежения за глазами* 'eye-tracking'.

4. Conclusions

In this research, we analyzed novel words, automatically detected on the basis of 2016-2017 online newspaper corpora. Half of the neologisms are loanwords. The other half is mainly composed of compounds, formed either of native components only, or a mixture of native and foreign components. Finally, suffixation represents the largest group of word formation via affixation.

References

1. *Galtseva, A.* (2014), Neologizmy XXI veka [Neologisms of the XXI century]. In: Kontsept [Concept], Special Issue 13, pp. 1–8.
2. *Rybushkina, S. V.* (2015), Assimil'atsija inojazytchnyx neologizmov v sovremennom russkom jazyke pod vlijaniem ekstraligvistitcheskix faktorov [Assimilation of foreign neologisms in the modern Russian language under extralinguistic influence]. In: Vestnik Tomskogo gosudarstevennogo universiteta [Tomsk State university Newsletter], no. 392, pp. 34–38.
3. *Zhdanova E. A., Raciburskaya L. V.* (2015), Sovremennaya ukrainskaya deistvitel'nost' v novoobrazovaniyax rossiyskix massmedia [The actual Ukrain reality in new word-buildings of Russian massmedia]. In: Vesti Nizhegoroskogo universiteta im. Lobatchevskogo 2 [News from the Nizhegorodsky University], pp. 397–401.
4. *Cartier E.* (2016), Neoveille, système de repérage et de suivi des néologismes en sept langues [Neoveille, a system of neologism identification and tracking in seven languages]. In: Neologica 10, pp. 101–131.

5. *Cartier E.* (2017), Neoveille, a Web Platform for Neologism Tracking. In: European Chapter of Association for Computational Linguistics 2017, April 2017.
6. *Humbley J.* (2008), Les dictionnaires de néologismes, leur évolution depuis 1945: une perspective européenne [Dictionaries of neologisms, their evolution since 1945: a European perspectve]. In: Sablayrolles (ed.), Neologie et terminologie dans les dictionnaires, Paris, Honoré Chamion.
7. *Pruvost, J., Sablayrolles, J.-F. (eds.)* (2016), Les néologismes, Paris, Presses Universitaires de France.
8. *Sharoff S., Kopotev M., Erjavec T., Feldman A.* and *Divjak D.* (2008), Designing and evaluating Russian tagsets. In: Proc. LREC 2008, Marrakech.
9. *Schmid H.* (1994), Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.

Tatiana Iakovleva-Vigné
Université Paris Diderot (France)
E-**mail:** ***tiakovle@eila.univ-paris-diderot.fr***