

VIII konferencja naukowa „Słowa klucze 2022”

Spis abstraktów

Uwaga: nie wszystkim referatom przedstawiające słowa klucze z danego kraju towarzyszą abstrakty.)

Agnieszka Będkowska-Kopczyk

ATH, Bielsko-Biała

Słowo „dobrostan” i jego ekwiwalenty w językach słowiańskich wśród słów kluczy współczesnej kultury

Marcin Będkowski

UW

Systemowe środki perswazji w tekstach prasowych. Wstępne studium korpusowe

Zasadniczym celem wystąpienia będzie przedstawienie wyników wstępnego studium korpusowego mającego na celu identyfikację w tekstach prasowych zjawisk o charakterze argumentacyjnym. Na wystąpienie złożą się cztery części. W pierwszej scharakteryzowane zostaną wybrane zjawiska tekstowe o charakterze argumentacyjnym, w szczególności metaoperatory należące do systemowych środków perswazji (Awdiejew 2004, Sinnott-Armstrong, Fogelin 2015) i spójniki wyrażające relacje logiczne (Segura-Tinoco A., Cantador I. 2021). W drugiej części przedstawiony zostanie mikrokorpus tekstów prasowych ręcznie zaanotowany pod kątem wyróżnionych zjawisk. W części trzeciej omówione zostaną techniki uczenia maszynowego oparte na nowoczesnych architekturach sieci neuronowych pozwalających na automatyczną klasyfikację tokenów (Devlin i in. 2018) oraz wyniki ich trenowania na opracowanym mikrokorpusie i rezultaty osiągnięte w odniesieniu do zebranego korpusu tekstów. Część czwarta zawierać będzie konkluzje dwojakiego rodzaju. Po pierwsze, wnioski dotyczące wybranych z zakresu zjawisk argumentacyjnych występujących w opracowanym korpusie tekstów prasowych. Po drugie, wnioski o charakterze metodologicznym – dotyczące możliwości tworzenia zaanotowanych korpusów tekstów na podstawie stosunkowo niewielkich próbek ręcznie zaanotowanych danych.

Marta Chojnacka-Kuraś, Katarzyna Osior-Szot, Magdalena Wanot-Miśtura

UW, Fundacja Języka Polskiego

Słowa klucze prostego języka

Waldemar Czachur

UW

Wrzesień jako słowo klucz w polskiej pamięci

Agnieszka Dryjańska

UW

Francuskie słowa kluczowe i ich ekwiwalenty polskie

Iryna Dilay, Marianna Dilay

Politechnika Lwowska

Association networks of the keywords of british online news in 2022

This study addresses online news discourse from the standpoint of corpus-assisted discourse studies (Partington, 2004) and focuses on the investigation of the keywords of British news on the web in 2022 and their associative networks. The research enables drawing statistically sound conclusions concerning the aboutness of the British online news, representation of the key concepts significant in society and their interconnection, which is important for the study and interpretation of ideological meanings the target audience of news reports is exposed to (O'Halloran, 2010).

For the purpose of this study, we have created a virtual corpus of the British news of 2022 (till September 2022) in the NOW corpus (<https://www.english-corpora.org/now/>) and extracted keyword lists, including specific nouns (cost-of-living, oligarchs, shelling, invasion, artillery, monarch, etc.), verbs (debate, invade, misled, abuse, etc.), adjectives (soaring, appalling, humanitarian, etc.), adverbs (utterly, reportedly, absolutely, etc.). Then, by means of #LancsBox software (<http://corpora.lancs.ac.uk/lancsbox/>) collocation (association) networks of the key nouns were built visually representing collocational relationship between the keywords and their collocates (MI-score). The collocation networks showed us how distant collocates can affect the meaning of a word by sharing a similar conceptual space with this word through associations with other words that directly collocate with that word (Brezina, 2018). For instance, the most frequent key noun cost-of-living is associated through the collocate crisis with families, migrant, refugee, energy, pandemic, Ukraine. We have noticed that the keywords intercollocate sharing common collocates (e.g., crisis is a common collocate for key nouns cost-of-living, invasion, pandemic, etc.). Thus, we can make a conclusion about the interconnection of keywords through word associations and cross-associations.

The proposed methodology combining the automatic extraction of keywords and building their associative networks facilitates and accelerates the content analysis and critical discourse analysis of news reports.

Simona Klemenčič
Słoweńska Akademia Nauk, Lublana
Słoweńskie słowo roku

W Słowenii konkurs na słowo roku przeprowadzany jest od 2016 r. Internauci zgłaszają swoje propozycje, następnie komisja specjalistów z ZRC SAZU (Centrum Naukowo-Badawcze Słoweńskiej Akademii Nauk), portalu MMC RTV Slovenija i czasopisma „Delo” ustala z nich zestaw 10 finałowych słów, spośród których w internetowym głosowaniu wybiera się słowo roku. Przedstawiono również neologizm roku.

Od dwóch lat odbyło się głosowanie na znak roku słoweńskiego języka migowego we współpracy ze Związkiem Stowarzyszeń Głuchych i Niedosłyszących Słowenii.

Jolanta Klimek-Grądzka
KUL
O „formacie” słów kilka

Przedmiotem oglądu będzie historyczne i dzisiejsze znaczenie leksemu „format”. Pretekstem do badań – opartych na ekscerpcji słownikowej (SPXVI, ESWIL, SL, SW, SJPD, USJP, WSJP PAN) i korpusowej (NKJP), oraz materiałach własnych autorki (prasa, tv, Internet) – są szerzące się, zwłaszcza w odmianie publicystycznej i politycznej polszczyzny współczesnej, połączenia typu „format normandzki” / „Format Normandzki”, „format współpracy”, „format ministrów spraw zagranicznych”, „format telefonicznej konferencji”, „format rozmów”. Celem badań jest nie tylko pokazanie wykształcania się nowego znaczenia leksemu jako kalki angielskiego „format”, ale także jego synonimów tekstowych i dwustronnych kolokacji. Podjęta także zostanie próba chronologizacji nowego „formatu”.

Nataliya Hrytsiv
Politechnika Lwowska
Key Words in Meaning Back Constructing Model: Translation Mode

The hypothesis of Meaning Back Constructing Model (hereinafter — MBCM) is that multilingual versions present the blend of entities which lead us backwards to the original

text meaning unit. Corpus-driven generated data allow juxtaposing translated prevailing entities and modeling cognitive, conventional and compositional modes of meaning of the original. Experimental, this algorithm would best fit within non-essentialist and deconstructed approaches to translation [1]. If further elaborated, MBCM is considered perspective for unit of meaning detection.

This vision is supported by the theoretical view that modern research rejects the idea of the translator as an intermediary between the original and translated messages; it also denies the understanding that stable, fixed 'messages are transferred' through the interpreter. Instead, the hypothesis of 'jointly constructed meaning' is supported.

The main idea of this cognitive semantic approach is that the meanings of expressions are mental [2]. Semantics is considered as mapping from the linguistic expressions to cognitive structures. Language itself is considered as a part of the cognitive structure, and not as a separate independent entity. Within cognitive semantics, the emphasis is on lexical meaning, not the meaning of sentences. Therefore, it is possible to test the feasibility of detecting original text meaning via reconstructing its keyness represented in translation(s). The algorithm is as follows: securing electronic versions of the English original text of 'Lincoln in the Bardo' by G. Saunders [3] and its Ukrainian [4] and German [5] translations; determining semantic parameterization of meaning-centered study; shaping corpora samples for further processing on Sketch Engine digital platform [<https://app.sketchengine.eu>]; illustrating pre-modeling stage of MBCM, which would resemble a chain of key word meanings: *спалах світлоречовина* (Ukrainian)+*materienlichtblüte* (German)→*matterlightblooming* (English). With further analysis. Keywords are extracted from the parallel corpora.

Key words: keyness, translation, cognitive semantics, meaning, parallel corpus.

- [1] The Routledge Handbook of Translation and Cognition, 2021, Christian M. I. M. Matthiessen
- [2] Cognitive semantics : meaning and cognition / edited by Jens Allwood, Peter Gärdenfors. 1999, 216 p.
- [3] G. Saunders, Lincoln in the Bardo. USA: Random House, 2017, 368 p.
- [4] Dzh. Sonders, Linkoln u bardo. Per. Andrii Masliukh. Vydavnytstvo Staroho Leva,2019, 464 s.)
- [5] G. Saunders, Lincoln im Bardo. Ub. von Frank Heibert, Luchterhand Literaturverlag in der Verlagsgruppe Random House GmbH 2018, 448 s.

Damian Kaleta

UW

Słowo roku na Węgrzech

Miroslaw Koziarski

Warszawa

The RetroDict Project: An Online Platform for Diachronic Lexicographic Research

This paper outlines the status and goals of a privately developed project tentatively named RetroDict, which aims at creating a repository of retrodigitized dictionaries within a single lexicographic framework. The project has been initiated in the hopes of facilitating access to primarily specialized historical lexis through an interface that is both human- and computer-readable.

Thus far, comprehensive diachronic metalexigraphic and lexicological studies have been severely limited. As there is an apparent lack of dedicated historical specialized dictionary search engines, researchers could either work with digitizations of general-purpose dictionaries (i.e. <https://eswil.ijp.pan.pl/>), access dispersed resources characterized by low interoperability (i.e. <https://academic.ru/>), use narrow language-specific aggregates (i.e. <https://slounik.org/>), or perform plain text searches on “dirty” OCR (i.e. <https://szukajwslownikach.uw.edu.pl/>). Moreover, most of available resources neither allow complex querying nor do they possess a standardized interface.

The RetroDict project is meant to address this problem by digitizing and integrating as much as 400 dictionaries spanning over 20 languages and covering around 80 domains of human activity: science (i.e. botany, mathematics, economy), industry, trade, technology, engineering (i.e. aeronautics, railway engineering, mining), craftsmanship (i.e. watchmaking, amber craftsmanship, honey hunting), law, sport (i.e. fencing, gymnastics), art and fashion. Both theoretical and technical aspects of the undertaking will be discussed.

From the theoretical perspective, questions such as dictionary selection criteria, past and present lexicographic access structure, and article structure will be examined. Special focus will be given to the issue of how information is presented in historical dictionaries and how can their content be abstracted away from their form.

From the technical perspective, various manual and automatic methods employed to extract and convert the data will be discussed, including optical character recognition, regular expressions and a proprietary toolchain.

Research methods include selective literature review, theoretical deliberations, as well as practical exemplification of data transformations.

Current status of the project and preliminary results will be presented as well. Moreover, it will be posited that a multi-lexicographic framework presents a viable alternative to linguistic corpora, both in terms of cost-effectiveness and layers of linguistic information.

References:

CORBIN, Pierre (2013): „Lire les dictionnaires (3). Éléments pour une modélisation des opérations impliquées dans leur consultation : de l’indentification des articles” [w:] Gasiglia, Nathalie (red.): Interpréter les dictionnaires : pluralité d’approches, [seria:] Corbin,

- Pierre (red.): *Lexique*, t. 21, s. 32–100. Villeneuve-d'Ascq: Presses Universitaires du Septentrion, 252 s. ISBN: 978-2-7574-0437-9. ISSN: 0756-7138.
- DEBUS-GREGOR, Esther & HEID, Ulrich (2013): „Design criteria and ‘added value’ of electronic dictionaries for human use” [w:] Gouws, Rufus Hjalmar & Heid, Ulrich & Schweickard, Wolfgang & Wiegand, Herbert Ernst (red.): *Supplementary Volume Dictionaries. An International Encyclopedia of Lexicography: Recent Developments with Focus on Electronic and Computational Lexicography*, s. 1001–1013. Berlin, Boston: De Gruyter Mouton. ISBN: 9783110238129. e-ISBN: 9783110238136. DOI: <https://doi.org/10.1515/9783110238136.1001>.
- GOUWS, Rufus Hjalmar (2018a): „Dictionaries and access” [w:] Fuertes-Olivera, Pedro A. (red.): *The Routledge Handbook of Lexicography*, s. 43–58. London and New York: Routledge, 809 s. ISBN: 978-1-138-94160-1. e-ISBN: 978-1-315-10494-2.
- HEID, Ulrich (2014): „2. The impact of computational lexicography” [w:] Gouws, Rufus Hjalmar & Heid, Ulrich & Schweickard, Wolfgang & Wiegand, Herbert Ernst: *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*, s. 24–30. Berlin & Boston: De Gruyter Mouton. ISBN: 9783110238136. DOI: <https://doi.org/10.1515/9783110238136.24>.
- KAMMERER, Matthias & WIEGAND, Herbert Ernst (1998): „Über die textuelle Rahmenstruktur von Printwörterbüchern. Präzisierungen und weiterführende Überlegungen” [w:] Gouws, Rufus Hjalmar & Heid, Ulrich & Herbst, Thomas et al. (red.): *Lexicographica*, t. 14, s. 224–238. Berlin: De Gruyter. ISSN: 1865-9403. e-ISSN: 2224-0039. URL: <https://www.proquest.com/docview/1859872664>.
- PRUVOST, Jean (2000): „Rapport de colloque des dictionnaires papieraux aux dictionnaires électroniques. VIIe Journée des dictionnaires (22 mars 2000)” [w:] *International Journal of Lexicography*, t. 13, z. 3, s. 187–193. ISSN: 0950-3846. e-ISSN: 1477-4577. DOI: <https://doi.org/10.1093/ijl/13.3.187>.
- ZGUSTA, Ladislav (1971): *Manual of Lexicography*. Prague: Publishing House of the Czechoslovak Academy of Sciences; The Hague, Paris: Mouton, 360 s.
- ŻMIGRODZKI, Piotr (2018b): „3. Makro- i mikrostruktura słownika.” [rozdział w:] Żmigrodzki, Piotr & Bańko, Mirosław & Batko-Tokarz, Barbara & Bobrowski, Jakub & Czelakowska, Anna & Grochowski, Maciej & Przybylska, Renata & Waniakowa, Jadwiga & Węgrzynek, Katarzyna (red.): *Wielki słownik języka polskiego PAN: geneza, koncepcja, zasady opracowania*, s. 25–38. Kraków: Instytut Języka Polskiego PAN, 267 s. ISBN: 978-83-64007-49-1. URL: <https://rcin.org.pl/dlibra/publication/edition/68631?id=68631&from=publication>

Lutz Kuntzsch

Gesellschaft für Deutsche Sprache, Wiesbaden

50 Jahre „Wörter des Jahres“ in Deutschland. Schlüsselwörter im internationalen Vergleich

Olena Levchenko (Олена Левченко)

Politechnika Lwowska (Львівська політехніка)

Інновації в текстах українського воєнного інтернет-дискурсу

Мета цієї праці – проаналізувати відображення рецепції війни в українській мові, зокрема виявити основні тенденції okazіоналізації/неологізації лексико-фразеологічного рівня мови; метафоризації, які вербалізують концептосферу ВІЙНА; дослідити трансформації інтертекстем; актуалізацію жанрів т.зв. міського та мережевого фольклору; ресеміотизацію. Матеріалом дослідження послуговував український воєнний дискурс, зокрема дописи в соціальних мережах.

Дослідження українських воєнних текстів засвідчує появу okazіоналізмів/неологізмів різних типів, які стосуються явищ суспільного життя (рашизм, рашист, депутатнізація, могилізація, пакетизація та ін.). У досліджених текстах функціонує значна низка okazійних онімів на позначення країни-агресора та її представників: Ерефія; зомбі-кацапи, зомбі-нацики, Зомбіленд; кремлядь; лапте- (рейх, недоімперія, фашисти, орк, Шишкостан); Мордор; москалота; Москворепія; орко- (орко, мутанти, бидло); пушкініст; рабсія, рабсіяне; рашковани, рашери, роіся, росня/русня, руссіші, терорраша, узкоязики; фашня, х**лостан; шарикостан та ін. Спостерігаємо появу нових дієслів, зокрема й від власних назв: за-200хсотити/задвохсотити, затрьюхсотити, макронити, чорнобаїти, відкобзонити/закобзонити. Сьогодні значно активізувалося утворення слів-комполітів: Божеякевонокончене, Яневтік/Боневтік/БоНеВтік, мишебратья, неначасі. Звертає на себе увагу використання різноманітних графічних засобів як стилістичних прийомів (великі літери, латинські символи, закреслення): каРФаген, руZZкій мір, Семеро козлят козломордих зустрілися з «борсуками» 503 ОБМП (Telegram).

Спостережені трансформації інтертекстем (Ніч яка місячна зоряна ясная видно хоч орків стріляй... (FB)) спрямовано на осучаснення канонічних висловів. У мережі натрапляємо на численні приклади добро- і злопобажань: ППО спаси нас (FB), Земля скловатою паскуді (FB), най їх Кобзон прийме (Tw). У мережі активізувався і такий жанр, як замовляння: Сію тобі в очі, Сію проти ночі. Буде тобі, враже, Так, як Відьма скаже (L. Gogova, FB). Значно зросла кількість мемів, присвячених воєнній тематиці (ключові слова-символи, використовувані в текстовій частині: ЗСУ, Азов, Херсон, байрактар, Белгород, бурят, вишивка, джавелін, Доброго вечора, ми з України!, Кобзон, Польща, Чорнобаївка, Шевченко тощо).

Простежуємо актуалізацію низки метафоричних моделей, загалом типових для цієї концептосфери: ВІЙНА – ЦЕ ЛОКУС, (АГРЕСИВНА) ІСТОТА, ОБ'ЄКТ, ХВОРОБА тощо.

У мережєвих текстах, у рекламі зросло вживання пейоративів, матизмів (Русский_воєнный__корабль_иди_н***). Мовлення інтернаєтв (як і більшості публічних осіб) відзначається експресивністю та агресивністю.

Oksana Melnychuk

Lwów

Key words representing concept of war in relation to body language in Frances Liardet's We Must Be Brave

The article aims to research key words representing the concept of war through their relation to the most frequent words denoting body language (*eyes, hand, head*) in Frances Liardet's *We Must Be Brave* which is a book of contemporary British domestic fiction revealing wartime events. British fiction with overtaken various war scenes as a type of socially significant narrative is a "symbolic system" that supports a specific field of knowledge, that is, the field of social beliefs and procedures. The relevance of the topic highlights the fact that fiction is a social instrument, having the task to reduce social tension in society and contribute to the processes of social harmonization, safety, and stability. In this account, body language as concentrated parts of the novel's narrative are the semiotic signs conveying different movements, postures, and gestures to describe characters' actions, thoughts, and emotions reflecting various aspects of social life. Body language designations are incorporated by the author as a tool for projecting the concept of war through the semantic relations of body language cues to key words which are symbols of the "aboutness" of the researched novel and at the same time are part of the vocabulary of culture and society.

Involving the quantitative method of computer-based processing of a fictional text, the lexical correlations for the three most frequent words describing body language in the researched novel were consequently generated. Further, the method of qualitative analysis to sort out all of the generated correlations was used to separate key words embodied in thematic groups "negative emotions", "negative characteristics of state or event", and "military terms and events" that unveil the concept of war. The results of the research outline further studies of key words in other types of different social media contexts.

List of Literature

Liardet F. (2019). *We must be brave*. London: Harper Collins, 2019.

Newlin, K. (2019). *The Oxford Handbook of American Literary Realism*. New York: Oxford University Press.

Moor, B. (2018). Invisible Architecture and Social Space in North and South. *The Gaskell Journal*, 32, 17–36.

Kvas, K. (2019). *The Boundaries of Realism in World Literature*. Lanham, Boulder, New York, London: Lexington Books.

Williams, R. (1976). *Keywords: A Vocabulary of Culture and Society*. London: Fontana.

Pavel, Thomas (2015). *The Lives of the Novel: A History*. Princeton: Princeton University Press.

Vanzant, Kevin. "Problems with Narrative in the U.S. Survey and How Fiction Can Help." *The History Teacher* 52, no. 4 (2019): 677–96.

Caiani, Fabio, and Catherine Cobham. "Autumn Visions: War and the Imagery of Muḥammad Khudāyyir." *Journal of Arabic Literature* 49, no. 3 (2018): 243–70.

Ogunyemi, Chikwenye Okonjo. "The Poetics of the War Novel." *Comparative Literature Studies* 20, no. 2 (1983): 203–16.

Cobley, E. (1986). *Description in Realist Discourse: The War Novel*. *Style*, 20(3), 395–410.

Pollmann, I. (2018). *Open Bodies, Open Stories: Evolution, Narration, and Spectatorship in Post-war Film Theory*. In *Cinematic Vitalism* (pp. 207–256). Amsterdam University Press.

Cobley, Evelyn, (1993). *Representing War: Form and Ideology in First World War Narratives* (U of Toronto Press).

Kortmann, B. "Reflecting on the quantitative turn in linguistics", *Linguistics*, vol. 59, no. 5, 2021, pp. 1207-1226. <https://doi.org/10.1515/ling-2019-0046>

J. Pääkkönen, 2021. "Data do not speak for themselves: interpretation and model selection in unsupervised automated text analysis", in *Composition and Big Data*, A. Licastro and B. Miller Eds. Pittsburgh, Pa: University of Pittsburgh Press, pp. 245-261.

Kılınçoğlu, D. T. (2020). *Changing Reality through Fiction: A Cognitive Approach to Late Ottoman Literature*. *Journal of the Ottoman and Turkish Studies Association*, 7(1), 205–226.

Lidiia Melnyk

Schiller Universität Jena

Top2Vec-based keywords research of the war-related tweets from the Ukrainian users

In my presentation, I will focus on automated keyword extraction methods for topic modeling. My research aims to identify the leading topics of the Ukrainian and Russian-speaking tweets coming from the Ukrainian geolocation, which could help better understand overall trends on social media. Topic-modeling-based techniques of keyword retrieval and keyword allocation are tested on the multilingual Kaggle dataset consisting of 49 million tweets on the topic of the Russian invasion. The share of tweets in the Ukrainian or Russian languages is way smaller. The collection of tweets started on the 24th of February, the beginning of the Russian invasion of Ukraine, and is being automatically updated on the daily basis. The dataset is publicly available via Kaggle. With over 280 thousand tweets coming from the Ukrainian geo location in the Russian and Ukrainian languages, we can compare the tendencies of language use as well as the topics that prevailed in the specific language. Having carried out topic modeling with the help of Top2Vec, I could identify 16 key topics for the Ukrainian language and 22 - for Russian speakers. Once the topics are already identified, we can also extract the main keywords for each topic and trace the spread of the keywords across topics, time, and languages. Given the changing attitude towards the Russian language in Ukrainian society, statistical topic modeling and keyword research might help identify the different purposes and topics people are using Ukrainian and Russian languages for and trace the dynamics of language use.

Adam Pawłowski, Tomasz Walkowiak

Uniwersytet Wrocławski, Politechnika Wrocławska

Mężczyźni są z Marsa, a kobiety z Wenus? Analiza słów kluczy generowanych automatycznie z wielkich bibliografii narodowych (przykład Polski i Niemiec) z uwzględnieniem płci autora

Wielkie bibliografie powstawały przez dziesięciolecia we wszystkich państwach, które w swym rozwoju wytworzyły znaczący zasób piśmiennictwa. Nie są one tylko rejestrami metadanych, służącymi zaspakajaniu doraźnych potrzebom informacyjnych. Odzwierciedlają także wielkie cywilizacyjne trendy oraz profil kulturowy danej społeczności.

Ich wykorzystanie do celów badawczych stało się możliwe dzięki cyfryzacji i narzędziom NLP. Duże znaczenie miały też staranne i systematyczne zapisy pracowników informacji naukowej oraz powtarzalna struktura formatu MARC. Przedmiotem artykułu jest analiza wybranych pól tekstowych w wielkich bibliografiach, reprezentujących piśmiennictwo polskiego i niemieckiego obszaru kulturowego, tworzonych przez powołane do tego instytucje, czyli biblioteki narodowe. Objętości obu baz są znaczące (ok. 1,5 miliona rekordów dla bazy polskiej i ok. 5 milionów dla niemieckiej). Jako okres badawczy wybrano tzw. późną nowoczesność i współczesność (1800—2200). Korpus danych stworzono z zawartości pola tytułu, a dodatkowo uwzględniano także pole opisu rzeczowego. Zapis rekordów w formacie MARC umożliwił przypisanie każdemu minitekstowi w korpusie roku oraz płci autora. Słowa klucze generowano metodą topic modeling oraz BERT (Bidirectional Encoder Representations from Transformers). Wyniki generowano dla podkorpusów męskiego i żeńskiego, stosując dodatkowe podziały na okresy (wszystkie lata, współczesność) i makrokatunki (teksty naukowe, nienaukowe). Badanie wykazało, że zarówno w polskim, jak i niemieckim piśmiennictwie obszary tematyczne, na których koncentrują się kobiety i mężczyźni są zupełnie różne.

Piotr Pęzik

Uniwersytet Łódzki

Keyword extraction vs. Keyword generation. Evaluating text-to-text transformers

Joanna Senderska, Anna Wileczek

Słownictwo dotyczące człowieka w slangu młodzieżowym (na materiale Obserwatorium Języka i Kultury Młodzieży)

Obserwatorium Języka i Kultury Młodzieży, redagowane przez pracowników Uniwersytetu Jana Kochanowskiego, jest stosunkowo nowym projektem, w ramach którego gromadzone

są słownictwo i frazeologia używane przez młodych ludzi oraz takie jednostki znaczeniowe jak memy czy wirale, zajmujące ważne miejsce w elektronicznej komunikacji młodzieży. Celem projektu jest przede wszystkim popularyzacja wiedzy o slangu młodzieżowym, a także o sposobach uczestnictwa młodych ludzi w kulturze. Materiał zgromadzony na stronie Obserwatorium stanowi też cenne źródło badań nad językiem i kulturą współczesnej młodzieży.

Domena semantyczna CZŁOWIEK jest licznie reprezentowana w zasobie leksykalnym języka młodzieży, ta odmiana języka bowiem ma wyraźnie antropocentryczny charakter, odpowiednio zresztą do ukierunkowania całej kultury młodzieżowej. Autorki referatu planują zbadanie leksyki odnoszącej się do człowieka zarówno w aspekcie jakościowym, jak i ilościowym. Oprócz słownictwa i frazeologii od dłuższego czasu obecnych w języku młodzieży na stronie Obserwatorium zarejestrowane zostały jednostki leksykalne nigdzie indziej nienotowane i nieopisywane. Umożliwia to – jak się wydaje – naszkicowanie najnowszych tendencji w zakresie slangowego słownictwa odnoszącego się do człowieka.

Medea Saghliani, Nato Shavreshiani

Tbilisi State University

Key words related to weather in Svan (according to the Loan-Words dictionary)

The scientific literature expresses different opinions about the cognateness of the Kartvelian languages (Georgian, Mingrelian - Laz and Svan) to other languages. Some scientists unite them in the group of the IberoCaucasian language family, Some scholars regard them as independent language families. Georgian is the state language, while Svan and Mingrelian – Laz are tribal (with the function of family) languages, which came from the dialects of one common root-language. The vocabulary reflects well both the old situation and the changes that take place in the society over time, and the study of vocabulary as semantic (sectoral) groups allows more complete account of the verbal richness of the language and the relevant social realia. This time we will discuss one of the most interesting and important lexical group for history-ethnology and linguistics of sectoral vocabulary – denoting the meteorological (atmospheric and celestial) phenomena according to the data of traditionally known dialects of Svan (Upper Bal, Lower Bal, Lashkhian and Lentekhian) and Cholur speech. In the present research, based on relevant studies and new approaches and interpretations, we present the structural-semantic and etymological analysis of a number of Svan meteorological lexical units based on the key words and phrases most frequently used in Svan. Synchronous, descriptive, historical-comparative and internal reconstruction methods are used as the research method, which is conditioned by the essence of the set tasks. The material taken as key words (related to bad and good weather, rain and snow, wind and storm...) is based on the "Loan-words Dictionary" prepared by us for publication and (about 80-90 units) is divided into the following subgroups: 1) Lexical units assimilated from Georgian-Mingrelian: borio/boria (Cholur) < Georgian boria{jj}/borio "cold wind (of the

North)"... 2) Lexical units assimilated through Georgian-Mingrelian: mex (Upper Svan, Lower Svan)/məx (Lentekhian, Cholur) < Georgian mekh-i (< Ossetian meɣ) "discharge of atmospheric electricity on the earth"... 3) Lexical units assimilated from Adyghe: mus{w} (Upper Svan, Lower Svan) < Adyghe ǎsə "snow" (for comparison Kabardian wǎs, Abaza sə, Abkhaz s-sə)... The analysis of the material showed that most of the borrowed lexical units are phonetically-semanticly changed, but at the same time, phonetically-semanticly unchanged lexical units are also quite tangible. In parallel with the material representing meteorological events directly assimilated from Georgian-Mingrelian, lexical units assimilated through Georgian-Mingrelian (from Ossetian, Greek, Russian) and directly from Adyghe were also confirmed. The most frequently used words are: rain, snow, wind, thunder, etc.

The research was conducted in 2019 within the framework of the Fundamental Research State Grants funded by the Shota Rustaveli National Science Foundation of Georgia _ „Phonetical-Phonological and Semantic Analysis of Loan-Words in the Svan Language” (fundamental-applied research, appended with a dictionary)“.

Shorena Shavreshiani, Lela Giglemiani

Tbilisi State University

Key words of Svan folk tales (according to the Loan-Words dictionary)

Along with Georgian and Colchian (Mingrelian-Laz), four traditionally known dialects of Svan (Upper Bal, Lower Bal, Lashkian and Lentekhian) belong to the group of South Caucasus languages, which most scholars consider to be one of the branches of the Caucasian language family, while some consider it an independent language. This group is known in the linguistic literature as the Kartvelian (Iberian) languages. Svan, despite being an unwritten language, is quite rich in lexical units of various fields, especially toponyms, the main stock of which is derived from the common Georgian root. In addition to this material, a large amount of data is found in Svan as directly derived from Georgian-Zan, so Greek, Turkish, Persian, Arabic, Ossetian, North Caucasian, Russian and other languages data imported through Georgian. This time the main goal of our report is the structural-semantic as well as etymological analysis of lexical units based on the lexical material of borrowed words (Georgian-Mingrelian, Abkhaz-Adyghe, Dagestani, Nakh...) presented in Svan tales (toponyms – Sodom-godom (Cholur) "fabulous building - meaningless, ugly, tall building, superficially built" < Georgian Sodom-gomor-i (< Hebrew sədom "fired" and Greek Σόδομα) "the two biblical cities that God punished according to the Old Testament for the sins of the city inhabitants", (antrophonyms _ baybäkdäw (Upper Bal)/baybäkdäw (Lower Bal) < Georgian bakbakdev-i..., common nouns _ paraxot (Lentekhian) < Georgian parakhod-i < Russian пароход "ship; boat", ʒaič-sayar (Lashkian)/čaiʒsayar (Cholur) a fabulous horse (exactly: a horse with white brow)" < Georgian taich-i "pedigreed horse, thoroughbred horse" < Pahlavi. tāčik "Arabic horse" and saghar-i "white-fronted horse" and so on.).

Synchronous, descriptive, historical-comparative and internal reconstruction methods are used as the research method. Words of foreign origin presented as key words in Svan folk tales of various genres have also become frequently used words in everyday life and have undergone phonetic-phonological and semantic adaptation over time. We singled out the most frequently used words: däw "giant", käž "evil spirit, demon"..., these lexical units came from Georgian in Svan according to the materials presented in mentioned work. The archaicism of analytical lexical units in many cases indicates ancient borrowing. The research was conducted in 2019 within the framework of the Fundamental Research State Grants funded by the Shota Rustaveli National Science Foundation of Georgia _ „Phonetical-Phonological and Semantic Analysis of Loan-Words in the Svan Language” (fundamental-applied research, appended with a dictionary)“

Teresa Uchman

Uniwersytet Śląski

Słowa klucze a kolor w telewizyjnych serwisach informacyjnych

Tematem planowanej prezentacji są badania przeprowadzone na podstawie obserwacji plansz/infografik zarejestrowanych w wieczornych serwisach informacyjnych stacji TVP1 oraz TVN. Prezentowane podejście badawcze wyznacza paradygmat lingwistyki dyskursu oraz lingwistyki mediów. Głównym przedmiotem zainteresowania jest kolor, który traktowany jest przez autorkę jako niejęzykowy środek tworzenia znaczeń, zasób semiotyczny rozumiany jako wielofunkcyjny i społecznie użyteczny artefakt, wykorzystywany do wzajemnej komunikacji i kulturowej realizacji. Celem wyznaczonego zadania jest określenie wartości denotacyjnych i ciągów konotacyjnych barwy w przekazach informacyjnych w dyskursie telewizyjnym. Metodą wykorzystaną w badaniach jest analiza koloru oraz analiza słów kluczy zawartych w mediotekście. Wspomniane już słowa klucze stanowią podstawę i fundament profilowania na poziomie tekstu – to one nadają barwie referencję oraz określają kontekst znaczeniowy. Analiza materiału włącza także profilowanie kulturowe omawianych jednostek. Obraz wynikający z profilowania oraz stworzone na tej podstawie zestawienia mają za zadanie odpowiedzieć na stawiane podczas badań pytanie, czy analiza słów kluczy pozwala na wyróżnienie i określenie skonwencjonalizowanych praktyk realizacyjnych koloru w telewizyjnych serwisach informacyjnych?

Marlies Whitehouse

ZHAW, Zürich

Wort des Jahres in der Schweiz

Anna Wileczek, Marek Łaziński

UW, UJK w Kielcach

Młodzieżowe Słowo Roku 2022. Analiza finałowej dwudziestki

Na kilka dni przed ogłoszeniem wyników plebiscytu na młodzieżowe słowo roku przyjrzymy się słowom z finałowej dwudziestki. Zwrócimy uwagę na emocje i oceny niesione przez zgłaszane słowa, reprezentację tzw. manosfery, wreszcie na źródła zgłoszeń w kulturze medialnej memów i w świecie gier.

Przedstawimy też wybrane spośród najbardziej kluczowych statystycznie słów prasy 23 słowa i wyrażenia, spośród których można wybrać słowo roku (nie młodzieżowe) w plebiscycie UW (sloworoku.uw.edu.pl, głosowanie do 31.12.2022).

Tęm dla polskich słów kluczowych roku 2022 będą ogłoszone już słowa roku w innych krajach: szwajcarskie, austriackie, brytyjskie i amerykańskie oraz niemieckie słowa młodzieżowe Jugendwort.

Natalia Zawadzka-Paluckta, Witold Kieraś

IPI PAN

Słowa klucze polskiego dyskursu politycznego na przestrzeni ostatnich stu lat: analiza korpusu exposé premierów

Artykuł przedstawia wyniki analizy słów kluczy w korpusie exposé polskich premierów od 1919 do 2019 roku. Jej celem jest opisanie zmian — odzwierciedlających również zmiany pozajęzykowe — jakim dyskurs polskiej polityki podlegał na przestrzeni ostatniego stulecia. Wedle najlepszej wiedzy autorów, stanowi ona pierwszą próbę określenia słów kluczowych języka polskiej polityki przy wykorzystaniu metodologii ilościowych. Najbardziej charakterystyczne dla zbioru słowa zostały wydobyte metodą TF-IDF, a następnie przeanalizowane za pomocą analizy głównych składowych i hierarchicznej analizy skupień. W drugim etapie badania, zidentyfikowane jednostki kluczowe zostały podzielone na kategorie tematyczne i przeanalizowane pod kątem frekwencji i funkcji dyskursywnych. Badanie pozwoliło na wyodrębnienie okresów w najnowszej historii polskiego dyskursu politycznego o wyróżniających się cechach leksykalnych, ale także na wyznaczenie elementów stałych, charakterystycznych gatunkowo. Ponadto, wskazało na pozajęzykowe — społeczne, kulturowe, polityczne — przemiany, które znalazły odzwierciedlenie w analizowanym dyskursie. Zastosowanie metod kwantytatywnych pozwoliło na uzyskanie szerszego, bardziej zobiektywizowanego i precyzyjniejszego obrazu dyskursu polskiej polityki.

Bibliografia:

Eder M. i Górski R. L. 2019: *Typologia tekstów oparta na miarach kwantytatywnych: Studium korpusowe o zróżnicowaniu polszczyzny*, „Język Polski” 3, s. 5-17.

Kieraś W., Kobyliński Ł. i Ogrodniczuk M. 2018: *Korpusomat — a tool for creating searchable morphosyntactically tagged corpora*. „Computational Methods in Science and Technology” 24(1), s. 21-27.

Milizia D. 2010: *Keywords and phrases in political speeches*, [w:] M. Bondi i M. Scott (red.): *Keyness in texts*, John Benjamins Publishing Company, Amsterdam/ Philadelphia, s. 127-146.

Ogrodniczuk M. 2018: *Polish Parliamentary Corpus*, [w:] D. Fišer, M. Eskevich i F. de Jong (red.), *Proceedings of the LREC 2018 Workshop ParlaCLARIN: Creating and Using Parliamentary Corpora*, European Language Resources Association, Paryż, s. 15-19.

Rayson P. 2008: *From key words to key semantic domains*, „International Journal of Corpus Linguistics” 13(4), s. 519-549.

Siewierska-Chmaj A. 2005: *Język polskiej polityki: Politologiczno-semantyczna analiza expose premierów Polski w latach 1919-2004*, Wydawnictwo Wyższej Szkoły Informatyki i Zarządzania, Rzeszów.

Wierzbicka A. 1997: *Understanding cultures through their key words*, Oxford University Press, Oxford.

Natalia Zawadzka-Palucktau, Aleksandra Tomaszewska, Joanna Wołoszyn

IPI PAN

Po Narodowym Korpusie Języka Polskiego — zmiany w słownictwie polskim ostatniej dekady na przykładzie słów kluczowych

Lingwistyka korpusowa od lat zdobywa coraz większe znaczenie w językoznawstwie. Kluczowe dla powstania i rozwoju tej dziedziny są, oczywiście, korpusy, czyli duże zbiory autentycznego tekstu reprezentatywnego dla danego języka czy jego odmiany w określonym czasie. W niniejszym badaniu chcemy położyć nacisk właśnie na aspekt diachroniczny badań korpusowych i pokazać najbardziej znaczące trendy leksykalne, jakie nastąpiły w polskim dyskursie publicznym w ciągu ostatniej dekady. Wykorzystując metodę słów kluczowych, porównaliśmy a) zbiór tekstów z czterech czasopism publikowanych między 2011 i 2020 rokiem, zawierający 206 milionów segmentów, i b) podzbiór korpusu danych internetowych, na który składają się statyczne strony internetowe monitorowane regularnie od listopada 2021 roku i który obecnie liczy około 990 milionów segmentów, z c) podzbiorem prasy ze zrównoważonego Narodowego Korpusu Języka Polskiego (150 milionów segmentów, teksty do 2010 roku). Do wygenerowania list słów kluczowych posłużyliśmy się miarą ilorazu szans. Następnie, pogrupowałyśmy według kategorii semantycznych 60 rzeczowników pospolitych o najwyższych wartościach miary kluczowości dla każdego z nowszych zbiorów (w porównaniu z NKJP). To pozwoliło na wyodrębnienie dominujących trendów, które odróżniają zbiory tekstów z ostatniej dekady od NKJP. Wśród nich znajdują się zarówno zjawiska językowe identyfikowalne intuicyjnie, związane między innymi z postępem technologicznym i wydarzeniami o światowej skali, które w dużej mierze definiowały ostatnie lata (przede wszystkim pandemia COVID-19), jak również te bardziej subtelne,

„niewidoczne gołym okiem”, ale wskazujące na istotne zmiany pozajęzykowe o charakterze społecznym, ekonomicznym, politycznym itd. Wyniki te stanowią punkt wyjścia do dalszych, bardziej pogłębionych badań nad najistotniejszymi trendami w polskim dyskursie publicznym na przestrzeni ostatniego dziesięciolecia, a także zwracają uwagę na wartość badań diachronicznych, a tym samym na konieczność aktualizowania korpusów polszczyzny ogólnej, stanowiących niezwykle cenny zasób (nie tylko) naukowy, między innymi jako korpusy referencyjne w badaniach nad słowami kluczowymi.

Bibliografia:

- Brezina, V. (2018). Statistical choices in corpus-based discourse analysis. W C. Taylor i A. Marchi (Red.). *Corpus approaches to discourse. A critical review* (259-280). Routledge.
- Gabrielatos, C. (2018). Keyness analysis: Nature, metrics, and techniques. W C. Taylor i A. Marchi (Red.). *Corpus approaches to discourse. A critical review* (225-258). Routledge.
- Kilgarriff, A. (2005). Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory*, 1-2, s. 263-276.
- Pojanapunya, P. i Todd, R. W. (2018). Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis. *Corpus Linguistics and Linguistic Theory*, 14(1), s. 133-167.
- Przepiórkowski, A., Górski, R. L., Łaziński, M. i Pęzik, P. (2010). Recent developments in the National Corpus of Polish. *NLP, Corpus Linguistics, Corpus Based Grammar Research*, s. 302-309.
- Rayson P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*, 13(4), s. 519-549.