Denis Gaščić: ELEXIS visit grant

Methods for detection and evaluation of neologisms for the Croatian language

QUESTIONNAIRE BEFORE THE VISIT

How did you learn about the ELEXIS travel grants?

I was a MA student of information science at the Faculty of Humanities and Social Sciences at the University of Zagreb when Call 4 was announced. My professor Kristina Kocijan sent me a call, so I decided to apply because I am interested in natural language processing and e-lexicography.

What is your project about?

The purpose of my visit is to obtain an overview of

1. methods and tools for detecting and evaluating neologisms;
2. Estonian resources (corpora, lexical resources) and tools used for (semi-)automatic detection of neologisms;
3. Croatian resources (corpora, lexical resources) and tools that can be used or are used for (semi-)automatic detection and evaluation of neologisms.

During the visit, I also plan to define a theme and write a research proposal for my PhD thesis, probably in the field of neology.

What is your background that brought you up to this point?

I have a MA in information science, but I would say that I am a computer and language enthusiast who every day tries to learn some new things and acquire new skills.

Which hosting institution did you apply to and why?

The hosting institution, Institute for the Estonian Language, is focused on modern automated lexicography, so they can introduce me methods and tools (DWS-s and CQS-s) they use for dictionary compilation and automatic detection of neologisms. They can also teach me how to use them and implement them in the research of the Croatian language.

Where does your interest in lexicography come from and what keeps you motivated?

I was always interested in creating new words (I created a few of them, and two of them were chosen for the final competition for new Croatian words). I am also interested in the whole lifecycle of neologisms. My area of interest also includes the field of machine translation, speech recognition, spell checking, diachronic analysis and automated detection of neologisms. I hope I will be able to develop algorithms for the automatic detection of neologisms for Croatian. My primary motivation is improving and developing NLP applications for modern Croatian and their practical implementation in tools and apps.

Travel Grant	Call 4	
Period of stay	12. – 26.6.2022	
Project title	Methods for detection and evaluation of neologism for the Croatian language	
Home institution	Faculty of Humanities and Social Sciences, Zagreb (Filozofski fakultet, Zagreb)	#elexis_hr
Hosting institution	Institut of the Estonian language (Eesti Keele Instituut)	#elexis_ee

****

REPORT

The period of my visit was from the 12th to the 26th of June. During the first week, I met colleagues from the Institute of the Estonian Language and participated in the 19th Conference of Applied Linguistics, “Influence of the language: from Data to Content-Rich Knowledge”, organised by the Estonian Association of Applied Linguistics. During the second week, I focused on my research and studied bibliographic sources. finalised my project result and wrote a PhD proposal.

1. Introduction

During a project, I obtained an overview of

1. methods and tools for detection and evaluation used in modern lexicography (e.g. Sketch Engine, Neoveille, Google Ngram Viewer);
2. Estonian resources (corpora, taggers, and lexical resources) used for (semi-) automatic detection of neologisms
    - lexical databases: ekilex.ee, WordNet;
    - corpora: Estonian National Corpus 2021, incl.Web Corpus 2021 and monitor corpora Timestamped_Feeds_2014-2021;
    - tools for Estonian NLP: NLP Toolkit for Estonian Estnltk and Universal Dependencies for the Estonian language;
    - Corpus Query Systems: Sketch Engine, Korp
3. Croatian resources (corpora, taggers, tools and lexical resources) that can be used or are used for (semi-)automatic detection and evaluation of neologisms
    - corpora: National Corpus, corpora in Sketch Engine;
    - lexical databases and dictionary portals: Hrvatski jezički portal, Hrvatski jezični korpus, Mrežnik;
    - tools for Croatian NLP: Universal Dependencies for Croatian.

I also defined a topic and wrote a research proposal for my PhD thesis. The subject of my PhD thesis will be “Grammar checker for the Croatian language: theory and modelling”.

For implementing a grammar checker, there are many preconditions to be done. One of them is also automatic detection of neologism because it is essential to distinguish between real neologisms and misspelt words. This is also why I, during the research visit, focused on the automatic detection of neologisms. Furthermore, besides detecting neologism, for developing a grammar checker, it is necessary to have corpora, structured lexical data (which includes the misspelt words related to right-spelt) and the API, which connects the database with the application. However, much time is needed to explore all these preconditions, so I primarily focused on neologism.

2. Description of work carried out during the research visit

On the first day of the project (June 13), I met my host Jelena Kallas, a Senior Computational Lexicographer-Project Manager at the Institute of the Estonian Language. She familiarised me with the Institute and its work.

During our meeting, we discussed primarily tools (DWSs and CQSs) used for dictionary compilation and neology detection (see part 2). The Institute of the Estonian Language uses DWS ekilex.ee and CQSs Sketch Engine and Korp. There is also a special NLP Toolkit for Estonian Estnltk, a Python library for performing common language processing tasks in Estonian. This toolkit is developed at the University of Tartu.

The next day (June 14), we met Martin Luts, a machine translation expert from the Institute. He gave me insight into new machine translation technologies, especially in using neural network algorithms and combining them with other technologies like statistical machine translation (SMT). We also discussed the importance of human feedback and correct training methods. Finally, we noticed that the question arises: "Where to store the data and does, for security reasons, text with confidential information can be translated via commercial translation apps?". I also met the Institute's NLP engineer Silver Vapper, who consulted me about optical character recognition (OCR) and bilingual lexicography. Although Estonia is a highly digitised country, OCR is needed to digitise old texts, i. e. to add their content to corpora, which is vital to see trends with words. In Croatia, on the other side, OCR

is also needed for getting modern language corpora too, because the government and public institutions still produce paper-based content.

On the third day (June 15), I met the Institute's project manager Marja Vaba. We discussed the Institute's products, especially Ekilex and Sõnaveeb. Sõnaveeb is the language portal of the Institute containing linguistic information from a growing number of dictionaries and databases. We concluded that one of the most important things is to know what the user of the language portal wants and what he/she needs, especially if he/she is not able to specify his/her needs. For that, we concluded that it is essential to research users' habits and needs, but also e.g. their educational background.

Maybe the most helpful conversation for my project was a discussion with Istok Kosem, a research assistant from the University of Ljubljana, who also visited the institute. Iztok is an e-lexicographer and NLP expert for the Slovene language. Because of the similarities between Croatian and Slovene, I could draw parallels between language technologies used for Slovene and those needed for Croatian. Istok also presented me for Sloleks (Slovenski oblikoslovni leksikon), Slovene Morphological Lexicon, based on the thesaurus database available at the CLARIN.SI repository. Sloleks is a lexicon of Slovene word forms, containing 100,802 headwords and 2,792,003 word forms with grammatical and accentual features. The innovation that Slolex offers to the user, compared to other lexicons, is that it predicts what the user wants when he is still typing. More precisely, it means that the application customises his/her search menu, so it offers not only a word but also additional information (e. g. the type of word, gender, grammatical information).

On Thursday and Friday (June 15-16), the fourth and fifth days of my visit, I participated in the 19th Annual Conference of Applied Linguistics. The programme included presentations about the most modern technologies and tools for language processing, but presentations about current unresolved issues in (corpus) lexicography.

During the second week (June 20-25) of the visit, I was focused on studying bibliographic sources (I used Elexifinder, EURALEX Proceedings, eLex Proceedings, materials of Globalex workshops on Lexicography and Neology), analysis of the corpora (mostly web corpora and monitor corpora) available for Croatian, finalizing my project result and writing a PhD proposal.

3. State-of-the-art techniques in neography

3.1. Introduction

According to McEnery, Xiao & Tono (2006), a corpus is "a collection of machine-readable authentic texts (including transcripts of spoken data) which is sampled to be representative of a particular language or language variety". Nevertheless, one of the main characteristics of language is that it is changeable. The changings of the language from the corpora perspective, in the opinion of the author of these lines, can be observed in three ways: (1) as changing by topics, (2) as changing in space or group, and (3) as a changing by time.

Firstly, when two people are talking about sports, they do not use the same vocabulary as two people talking about politics. Even the way they construct sentences can be different. In addition, the language on the same territory is different from the language on the other territory. At the same time, the language of the young is different from the language of the older people. There also can be some subcultures that use their part of language, which we call slang. In linguistics, those parts of language are known as dialects. According to the Cambridge Dictionary, “dialect is a form of a language that people speak in a particular part of a country, containing some different words and grammar”.

For those reasons, the defining feature of a corpus is representativeness. Leech (2011) defines a corpus as representative "if the findings based on its contents can be generalized to the said language variety".

Finally, a language is changeable over time. Some words are coming, while the others, at the same time, disappear. The words which come up are called neologisms. As for anything in humanities and social sciences, there is no one definition for neologisms. Plag explains them as “those derivatives that were newly coined in a given period” (Plag, 2002).

3.2. Neologisms and their detection

According to Cartier (2017), for tracking neologism in corpora, what is needed is "to have at hand large diachronic electronic corpora". Nevertheless, there are more problems with neology. Firstly, linguistics still observes neology as a non-primary field, as Cartier (2017) pointed out. On the other hand, Klosa and Lüngen (2018) claim that “a central issue in

lexicography is to find new lexemes and to identify new meanings for existing lexemes". Nevertheless, it is not easy to observe neologism automatically because computers do not understand meanings. So, the first way to trace neologisms was to induce unknown words. That system is called "exclusion dictionary architecture" (EDA) and includes two parts: monitor corpora and a list of known words with a list of misspelt words (Ibid.). The problem is that neologism is not only a new word, sometimes the word, which already exists, can get a new meaning. For example, the word "mouse", with the advent of computers, got meaning "the part of the computer for navigation on-screen". Therefore, another type of existing neology tracking system is Semantic Neology Approaches. It is based on the principle of computational understanding of contexts. Behind it is the "idea that meaning change is linked to domain change: every text and thus the constituent existing lexical units are assigned one or more topic; if a lexical unit emerges in a new domain, a change in meaning should have occurred (Gerard et al., 2014)". Based on that, Cartier created a Neovelle web platform for neologism tracking, which architecture has five components: a corpora manager, an advanced search engine on the corpora, advanced data analytics, a linguistic description component for neologisms and formal and semantic neologisms tracking with state-of-the-art techniques (Cartier, 2017).

In a Sketch Engine, there is a Trends function, which gives users (e.g. lexicographers) the opportunity to observe not only the appearance of new words but also the disappearance of old ones. It is based on mathematical methods, which are calculated by Python script developed by Ondřej Herman (2013). Heman was also comparing a lot of mathematical methods for corpus changing monitoring. He concluded that: "the method with the highest cost-to-benefit ratio for implementation seems to be the Theil-Sen slope estimator, along with the Spearman's ρ or Mann-Kendall tests to investigate a possible trend present in the word usage data" (Ibid.). It is, he described, "calculated by the attached code along with the implementation of the other regression methods. The code interfaces with Sketch Engine and can read data from Google n-grams datasets" (Ibid.).

Because Google has a large amount of text obtained by digitising books and tracking web content in English, it has enabled its users to track changes in language. The Google n-gram dataset is "a publicly available corpus with co-occurrence statistics of a large volume of web text" (Koplenig, 2017). An N-gram is, as Mazumder, Sourav and Baru (2022) pointed out, "a contiguous sequence of n words or tokens in a text document in computational linguistics and probability". It is a probabilistic language model that "can be classified into categories depending on the unit that incorporated them" (Ibid.).

In 2018, a group of authors in the context of the Horizon 2020 project ELEXIS surveyed lexicographic practices and lexicographers' needs across Europe. The results have shown that only 4,7 % of lexicographers across Europe use automatic extraction of neologisms. The research also shows that "the majority of the respondents compile their dictionaries manually (57.9%)" (Kallas et al., 2019.).

3.3. Resources for Croatian: corpora and lexicon

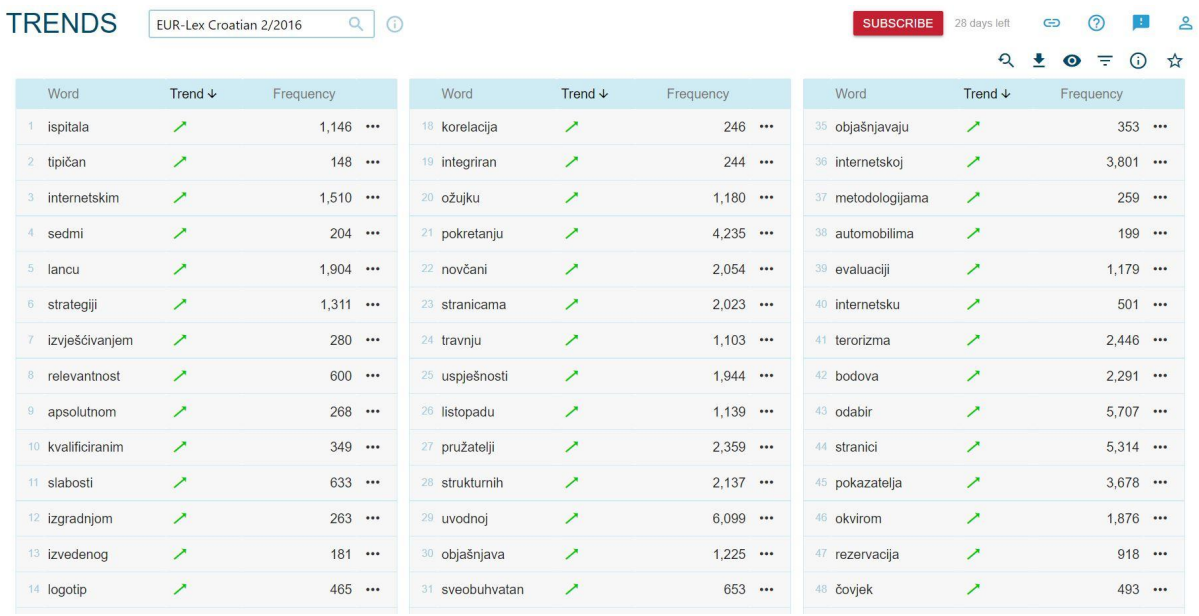
For the Croatian language, unfortunately, there are not so many corpora available. The Croatian National Corpus (Hrvatski jezični korpus), the largest one, is collected at the initiative of Prof Marko Tadić from The Faculty of Humanities and Social Sciences, and has 2,559,160 words and 2,130,095 lemmas. The last time it was updated was 11/02/2021, which is one and a half years before writing these lines. It is also based on NoSketchEngine, a free version of Sketch Engine that does not provide all the features. In addition, there is also Croatian Web (hrWaC 2.2, RFTagger) corpus in SketchEngine, which has 1,405,794,913 tokens and 1,211,328,660 words. It was crawled in 2011 and 2013, so it does not provide users with the real state of the language.

From the dictionary perspective, only the Hrvatski jezički portal (The Croatian Language Porta, HJP) is available. The Croatian Language Portal is the first and so far the only dictionary database of the Croatian language distributed on the Internet, which has been available free of charge since June 2006. The project received initial support from the Ministry of Science, Education and Sports in 2004 and has since been funded by the owner's funds. The Croatian Language Portal is the only such scientific reference work in Croatia. This dictionary requires continuous, detailed and painstaking work of several experts in the field of linguistics and other social sciences and humanities scientists to be updated following current knowledge and constant enrichment base. Unfortunately, for example, it does not contain coronavirus terms, which means that it has also not been updated for more than two years, as these words remain in the language.

Based on the Croatian Web Repository, there is also a Croatian Web-Dictionary – Mrežnik project. Authors say that "Croatia still belongs to the ever-smaller number of countries with no free online national language dictionary founded on modern e-lexicography, nor has systematic scientific research been carried out in this area" (Mrežnik). So, "the basic goal of this project is to change this in both of the aforementioned aspects.". The project is still in the working phase.

At the 3rd Globalex Workshop on Lexicography and Neology in 2021, Mihaljević, Hudeček and Lewisis (2021) presented a paper "Corona-related neologisms: A challenge for Croatian standardology and lexicography". Their research was also based on manually collected corona-terms because there was no automatic, even semi-automatic, system for tracing neologisms.

Sketch Engine has a few more corpora for the Croatian language. Only one enables the Trends function. That is a EUR-Lex Croatian 2/2016, EUR-Lex multilingual corpus of all the official languages of the European Union, which contains (only) 17,819,540 sentences and 156,309,317 words. Unfortunately, the quality of lemmatization and morphological analysis is not good enough (more detailed evaluation is needed), there are a lot of mistakes. It might be good to evaluate the lemmatizer used by Sketch Engine and possibly use another one developed especially for Croatian, not for Slovene. As a result, when we search trends by lemmas, we get almost the same result as searching by words (compare figures 1 and 2).

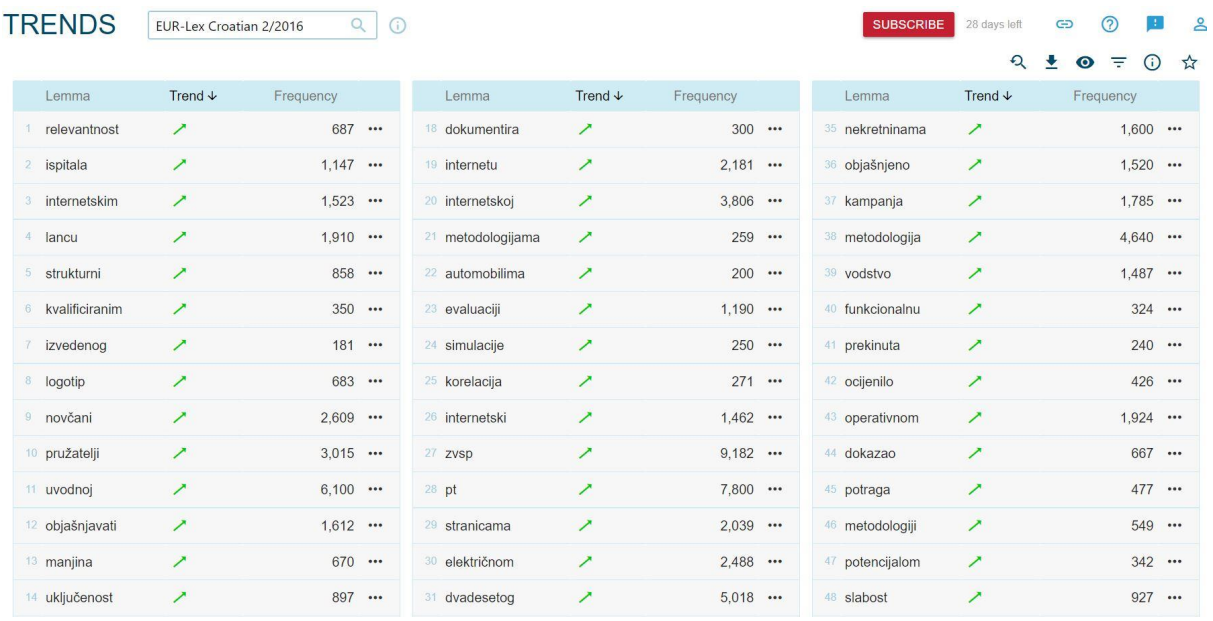


	Word	Trend	Frequency
1	ispitala	↗	1,146
2	tipičan	↗	148
3	internetskim	↗	1,510
4	sedmi	↗	204
5	lancu	↗	1,904
6	strategiji	↗	1,311
7	izvješćivanjem	↗	280
8	relevantnost	↗	600
9	apsolutnom	↗	268
10	kvalificiranim	↗	349
11	slabosti	↗	633
12	izgradnjom	↗	263
13	izvedenog	↗	181
14	logotip	↗	465

	Word	Trend	Frequency
18	korelacija	↗	246
19	integriran	↗	244
20	ožujku	↗	1,180
21	pokretanju	↗	4,235
22	novčani	↗	2,054
23	stranicama	↗	2,023
24	travnju	↗	1,103
25	uspješnosti	↗	1,944
26	listopadu	↗	1,139
27	pružatelji	↗	2,359
28	strukturnih	↗	2,137
29	uvodnoj	↗	6,099
30	objašnjava	↗	1,225
31	sveobuhvatan	↗	653

	Word	Trend	Frequency
35	objašnjavaju	↗	353
36	internetskoj	↗	3,801
37	metodologijama	↗	259
38	automobilima	↗	199
39	evaluaciji	↗	1,179
40	internetsku	↗	501
41	terorizma	↗	2,446
42	bodova	↗	2,291
43	odabir	↗	5,707
44	stranici	↗	5,314
45	pokazatelja	↗	3,678
46	okvirom	↗	1,876
47	rezervacija	↗	918
48	čovjek	↗	493

Figure 1. The Trends search by words



	Lemma	Trend ↓	Frequency
1	relevantnost	↗	687
2	ispitala	↗	1,147
3	internetskim	↗	1,523
4	lancu	↗	1,910
5	strukturni	↗	858
6	kvalificiranim	↗	350
7	izvedenog	↗	181
8	logotip	↗	683
9	novčani	↗	2,609
10	pružatelji	↗	3,015
11	uvodnoj	↗	6,100
12	objašnjavati	↗	1,612
13	manjina	↗	670
14	uključenost	↗	897

	Lemma	Trend ↓	Frequency
18	dokumentira	↗	300
19	internetu	↗	2,181
20	internetskoj	↗	3,806
21	metodologijama	↗	259
22	automobilima	↗	200
23	evaluaciji	↗	1,190
24	simulacije	↗	250
25	korelacija	↗	271
26	internetski	↗	1,462
27	zvsp	↗	9,182
28	pt	↗	7,800
29	stranicama	↗	2,039
30	električnom	↗	2,488
31	dvadesetog	↗	5,018

	Lemma	Trend ↓	Frequency
35	nekretninama	↗	1,600
36	objašnjeno	↗	1,520
37	kampanja	↗	1,785
38	metodologija	↗	4,640
39	vodstvo	↗	1,487
40	funkcionalnu	↗	324
41	prekinuta	↗	240
42	ocijenilo	↗	426
43	operativnom	↗	1,924
44	dokazao	↗	667
45	potraga	↗	477
46	metodologiji	↗	549
47	potencijalom	↗	342
48	slabost	↗	927

Figure 2. Trends search by lemmas

The word “ispitala” (Croat. “examined”) has been shown as the most frequent lemma. It is an impersonal participle form of the verb “ispitati”, but it is lemmatised as a masculine noun. In addition, there are also tokenisation mistakes. The problem is that for tokenisation Sketch Engine uses the MULTEXT-East Slovenian part-of-speech tagset. Although Croatian and Slovenian are similar languages, it is still reasonable to use taggers developed especially for Croatian, even quite frequent adjectives and numerals have wrong lemmas and POS. On the other hand, the word "ispitala" is in the Croatian Web (hrWaC 2.2, ReLDI) corpus tokenised well - as the verb participle singular feminine. For the lemmatization of the Croatian Web (hrWaC 2.2, ReLDI) the MULTEXT-East Croatian part-of-speech tagset is used. That tagset is a product of the MULTEXT-East project (Multilingual Text Tools and Corpora for Eastern and Central European Languages), whose main task it was to develop standardised language resources (Erjavec et al., 2017). The Croatian specifications were compiled soon after the MULTEXT-East project ended in 1997, using the project's Final report as the template (Ibid.). One of the objectives of MULTEXT-East has been to make its resources freely available for research (Ibid.). So, after my visit, I plan to contact the Sketch Engine team and propose to use a different parser for Croatian.

4. Conclusion remarks

Detection of neologism is a field which needs further research. A lot of lexicographers nowadays still detect neologisms manually, which takes time that they could devote to other tasks. As Kilgarriff et al. (2015) describe, “lexicographers read texts which are likely to contain neologisms – newspapers, magazines, recent novels – and mark up candidate new words, or new terms, or new meanings of existing words. It is a high-precision, low recall approach, since the readers will rarely be wrong in their judgments, but cannot read everything, so there are many neologisms that will be missed” (Ibid.). Only automated methods for corpus linguistics can provide a systematic analysis of large amounts of text, offering neologism candidates to lexicographers. There is a need to set up an infrastructure for neologism detection to supply lexicographers working with neologisms with candidates for inclusion in dictionaries. The problem with the detection of neologism is how to recognize not only new words but also new meanings of words which already exist. That is a reason why it is vital to also develop semantically annotated corpora, develop algorithms for sense clustering and share expertise in this field. This issue has been dealt with in the ELEXIS project, see e.g. deliverable on the topics of semantically annotated corpora, and Word Sense Disambiguation (WSD) algorithm for sense clustering, developed by Federico Martelli and Roberto Navigli (the results are available also at GitHub). They conclude that there are also two more directions of text analysis to have been explored: domain-labelling of texts and diachronic distribution of senses (Martelli et al., 2019).

On the other hand, there is no predisposition to implement (semi-)automatic detection of neologisms in the Croatian language nowadays. However, to enable such a system, a few essential things must have been done. Firstly, it is necessary to have a big, timestamped corpora so that language changes can be followed regularly, for example, monthly. It is also essential to develop monitor corpora and create Web corpora. A good example for Croatia can be Slovenia. For example, the newest version of Slovenian corpus Trendi (version 2022-05) contains 565.308.991 lemmas from 1.436.548 words. The Trendi 2022-05 corpus is available in three CLARIN.SI concordances: KonText, NoSketchEngine and the old version of the NoSketchEngine interface.

Mentioning the web corpora, there are two main problems with them. The first is cleaning, which means "removing those sections of a document that are textual but not linguistically informative" (Pomikálek, 2011), such as advertisements, headers, etc. The second problem is removing duplicate text (Ibid.) so the system is representative.

Furthermore, developing an advanced search engine and NLP tools for the corpora is essential. In addition, it is necessary to give human expert feedback to the system. Ultimately, it is crucial to follow the state-of-the-art technology and regularly analyse which methods and tools are used for other languages, especially Slavic, which could be implemented for the Croatian language. Without these predispositions (lemmatized corpora, dictionaries, thesaurus databases), there is also no predisposition for developing a (functional) grammar checker.

In summary, it is essential, as Tiberius et al. (2020) pointed out, to create "robust documentation, guidelines and collections, best practices in order to promote clearly defined workflows for producing, describing and annotating lexicographic resources (both synchronic and diachronic) in accordance with international standards and interoperability formats" (Ibid.).

The main problem with research of the Croatian language is, in the opinion of the author of these lines, that language processing is not recognized as an essential field, resulting in the non-investment of public money in language technologies. This field is also not recognized in the private sector. In my opinion, it is time to change the language policy in Croatia and to start investing in the development of language technologies.

Lastly, I would like to repeat that text analysis is one of the fields that still have to be discovered, especially when we talk about detecting neologisms. Although the lack of tools for automatic detection of neologisms is a problem today, at the same time, it could also be an opportunity for researchers like me who are interested in developing new things. Because of all that has already been said, I can find myself doing a PhD thesis in natural language processes.

5. Final words

The research visit grant at the Institute for the Estonian language played a significant role in my understanding of neology, automation detection of new words, but also in natural language processing generally. It was also useful and enjoyable to participate in the 19th Annual Conference of Applied Linguistics to get a professional overview of state-of-the-art methods and tools.

Furthermore, the themes and authors I have discovered during my visit to the Institute and the conference inspired me to continue with new research in the field. At the

same time, it inspired me to implement similar tools for the Croatian language because I consider how intensive and efficient their role is, especially in practical use by native speakers but also by language learners. .

Resources developed and used at Institute are practical, and they support both lexicographers and other language-orientated scientists like sociolinguists to get a better understanding of language, language processes and their social impacts. In addition, I would like to stress that NLP tools developed by the University of Tartu and the Technology University of Tallinn are open-sourced, which means they could be reused and implemented for other languages, like Croatian.

Finally, I was honoured to have been invited to the Institute of the Estonian Language to meet NLP engineers, lexicographers and other people who work at the Institute, especially my host Jelena Kallas. Of course, the knowledge I got from them will help me in my future work, but I think it is more important that their work inspired me.

References

Publications

1. Arppe, A. (2000, December). Developing a grammar checker for Swedish. In *Proceedings of the 12th Nordic Conference of Computational Linguistics* (NODALIDA 1999) (pp. 13-27).

2. Cartier, E. (2017, April). Neoveille, a web platform for neologism tracking. In Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics (pp. 95-98).

3. Erjavec, T., Krstev, C., Petkevic, V., Simov, K., Tadić, M., & Vitas, D. (2003, April). The MULTEXT-East Morphosyntactic Specification for Slavic Languages. In *Proceedings of the 2003 EACL Workshop on Morphological Processing of Slavic Languages* (pp. 25-32).

4. Herman, O. (2013). Automatic methods for detection of word usage in time: Bachelor's thesis. Masaryk: Masaryk University, Faculty of Informatics.

5. Kallas, J., Koeva, S., Langemets, M., Tiberius, C., & Kosem, I. (2019, October). Lexicographic practices in Europe: Results of the ELEXIS Survey on user needs. In Electronic Lexicography in the 21st Century. In *Proceedings of the eLex 2019 Conference*, Sintra, Portugal (pp. 1-3).

6. Kilgarriff, A., Herman, O., Bušta, J., Kovář, V., & Jakubíček, M. (2015, August). DIACRAN: a framework for diachronic analysis. In *Proceedings of Corpus Linguistics* (pp. 65-70).

7. Klosa, A., & Lüngen, H. (2018, August). New German words: Detection and description. In *Proceedings of the XVIII EURALEX International Congress Lexicography in Global Contexts 17-21 July 2018*, Ljubljana (pp. 559-569). Znanstvena založba Filozofske fakultete Univerze v Ljubljani/Ljubljana University Press, Faculty of Arts.

8. Koplenig, A. (2017). The impact of lacking metadata for the measurement of cultural and linguistic change using the Google Ngram data sets—Reconstructing the composition of the German corpus in times of WWII. Digital Scholarship in the Humanities, 32(1), 169-188.

9. Leech, G. (2011). Corpora and theories of linguistic performance. In *J. Svartvik (Ed.), Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82 Stockholm,* 4-8 August 1991 (pp. 105-126). Berlin, New York: De Gruyter Mouton.
10. Liou, H. C. (1991). Development of an English grammar checker a progress report. CALICO Journal, 57-70.
11. Martelli, F.; Navigli, R; Spadoni, P.; Stilo, G.; Velardi, P. (2019). Lexical-semantic analytics for NLP: sense clustering. ELEXIS - European Lexicographic Infrastructure.
12. McEnery, T., Xiao, R., & Tono, Y. (2006). Corpus-based language studies: An advanced resource book. Taylor & Francis.
13. Mihaljević, Hudeček, Lewisis. (2021). Corona-related neologisms: A challenge for Croatian standardology and lexicography. At: Globalex Workshop on Lexicography and Neology, 2021. Virtual. (Presentation).
14. Mikkelsen, I. L. S., Wiechetek, L., & Pirinen, F. A. (2022, May). Reusing a Multi-lingual Setup to Bootstrap a Grammar Checker for a Very Low Resource Language without Data. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages* (pp. 149-158).
15. Plag, I. (2002). Word-formation in English. Cambridge: Cambridge University Press.
16. Pomikálek, J. (2011). Removing boilerplate and duplicate content from web corpora. Disertacnı práce, Masarykova univerzita, Fakulta informatiky.
17. Tiberius, C., Costa, R., Erjavec, T., Krek, S., McCrae, J., Roche, C., & Tasovac, T. (2020). Best practices for lexicography – intermediate report. ELEXIS - European Lexicographic Infrastructure.

Other resources

1. CLARIN.SI
2. Ekilex
3. Estnlt
4. Estnltk
5. Google Ngram Viewer
6. Hrvatski jezički portal

7. Hrvatski jezični korpus
8. Korp
9. Mrežnik
10. Neoveille
11. Sketch Engine
12. Sloleks
13. Sõnaveeb
14. Universal Dependencies for Croatian language
15. Universal Dependencies for the Estonian language
16. WordNet