Human Language Technologies – The Baltic Perspective
A. Utka et al. (Eds.)

doi:10.3233/FAIA200625

Development and Research in Lithuanian Language Technologies (2016-2020)

Andrius UTKA [a,1], Jurgita VAIČENONIENĖ [a], Monika BRIEDIENĖ [b] and Tomas KRILAVIČIUS [b]
[a] *Vytautas Magnus University, Lithuania*
[b] *Vytautas Magnus University and Baltic Institute of Advanced Technology, Lithuania*

Abstract. The paper presents an overview of the development and research in Lithuanian language technologies for the period 2016-2020. The most significant national and international LT related initiatives, projects, research infrastructures, language resources and tools are discussed. The paper also surveys research production in the field of language technology for the Lithuanian language. The provided analysis of scientific papers shows that machine translation and speech technologies were the most trending research topics in 2016-2019.

Keywords. Lithuanian language, language technology research, language resources and tools, research infrastructures

1. Introduction

The scientific paradigm of *language technology* (LT) is changing in the world, intelligent technologies are developing at an incredible rate, robotisation and the Internet of Things are emerging. The amount of data is increasing exponentially and will soon reach hundreds of zeta bytes [1]. Most of the technologies are designed for the wider-used languages such as English, Chinese, Spanish, Arabic, or German, and we may observe groundbreaking achievements and breakthroughs in all complex language technology areas: machine translation, speech recognition and synthesis, dialogue systems, and natural language processing, among others.

On the other hand, the European Parliament resolution on "Language Equality in the Digital Age" adopted on September 11, 2018 [2] highlights that there is a "widening technology gap between well-resourced languages and less-resourced languages" maintaining that "European lesser-used languages are at a significant disadvantage on account of an acute lack of tools, resources and research funding" [2]. It is, therefore, important for smaller and less resourced languages to assess the current situation correctly and set up adequate goals for the future.

The state of art of LT in Europe and the Baltic states has been overviewed by [3, 4]. Discussion on language resources and technologies in Lithuania (2012-2015) was presented in [5], while this paper focuses on the landscape of human language technology developments in Lithuania, in 2016-2020. We survey Lithuania's involvement in inter-

[1] Corresponding Author: Andrius Utka; Vytautas Magnus University; E-mail: andrius.utka@vdu.lt.

national and national LT related initiatives and research infrastructures, and overview the advancements in language resources and technologies, as well as key projects and research.

2. Language Technology Related Initiatives and Projects

“Guidelines on the Development of Lithuanian Language in Digital Environment and Advancements in Language Technologies (2021-2027)”, adopted by the State Commission of Lithuanian Language, provide a thorough overview of the European and Lithuanian strategic documents, funding instruments and institutions regulating LT development in the country [6]. The main goal of the guidelines is to overview and facilitate the full use of the Lithuanian language in the digital environment. Drawing on the information provided in this as well as other relevant documents, this section will briefly highlight the major initiatives, projects and developed language resources.

Over the course of the last five years, Lithuania became involved in various European LT initiatives such as European Federation of National Institutions for Language (EFNIL)[2], European Language Resource Coordination (ELRC)[3], European Open Science Cloud (EOSC)[4], and FAIR data[5]. In 2019, together with other European countries, Lithuania signed the “EU Declaration on Cooperation on Artificial Intelligence” [7] and prepared “Lithuanian Artificial Intelligence Strategy” [8], which highlights the necessity of language technologies for the development of economy and science. Increasing international cooperation and researcher mobility are also evidenced by the participation of Lithuanian higher educational institutions in language technology related COST actions (e.g., IC1408, IC1207, CA18209, IC1002, CA18209)[6].

In the context of the most significant projects and their results, “The Lithuanian Information Society Development Programme 2014-2020” [9], funded by the EU Structural Funds, needs to be emphasized. The programme has launched 5 large-scale projects that develop language technology solutions and services in different areas: “Lithuanian Language Speech Services (LIEPA 2)”, “The Information System for Syntactical and Semantic Analysis of the Lithuanian Language (SEMANTIKA 2)”, “Integrated Information Systems of the Lithuanian Language and Language Resources (Raštija 2)”, “Modernization and Development of Machine Translation Systems and Localization Services”, and “Lithuanian Language Resources (E.kalba)”. The following language technology resources have been created while implementing the projects: syntactically annotated corpus (ALKSNIS)[7], morphologically annotated corpus (*gold standard*) (MATAS)[8], 1,000 hour speech corpus, Internet corpus (BIT), as well as a number of modernized corpora and lexicons in the E.kalba project. The developed tools and services include: neural MT system (Lithuanian, German, Polish, French, and Russian language pairs), automatic transcription of speech files, speech recognition of computer

[2] http://www.efnil.org/

[3] http://www.lr-coordination.eu/

[4] https://www.eosc-portal.eu/

[5] https://www.go-fair.org/fair-principles/

[6] https://www.cost.eu/

[7] https://clarin.vdu.lt/xmlui/handle/20.500.11821/21

[8] https://clarin.vdu.lt/xmlui/handle/20.500.11821/33

commands, Lithuanian language synthesizer, advanced internet search, automatic text summarisation, and hate speech detection.

The opportunities provided by EU Structural Funds, specifically, instruments encouraging companies to invest in research and experimental development (R&D) of innovative products (e.g., the instruments “Inočekiai”, “Intelektas”, “Eksperimentas”) are also exploited. A number of language technology projects were implemented by the Baltic Institute of Advanced Technologies, Vilnius University, and Vytautas Magnus University in cooperation with JSC Amberlo, JCS Tilde Information Technology, and other companies.

On a smaller scale, the Research Council of Lithuania, responsible for monitoring national science development and research funding, has financed six language technology projects for the period 2016-2020, which resulted in new language resources, tools and scientific publications. We should mention here the project PASTOVU[9] in which tools and methodology for the extraction of Lithuanian multi-word expressions were created and the project “Modern Spoken Lithuanian”[10] for modernizing a searchable spoken language corpus.

Important for language services and technologies are private business initiatives such as open frameworks and tools that can be tested and adapted for less resourced languages. A crucial impact on the development of LT is made by world business leaders (e.g., *Google*, *Microsoft*, *Facebook*, *Amazon*, *IBM*, etc.), as well as by data collection initiatives such as *Mozilla Common Voice* and *Glosbe*. The largest language technology companies established the *LT-Innovate* language technology industry association, where they can share ideas and develop strategic solutions. Lithuanian business initiatives can be exemplified by the Lithuanized *SpaCy*[11] library developed by *JSC TokenMill*, and Lithuanian speech recognition, speech synthesis[12], and machine translation[13] demo online services created by *JSC Tilde*.

3. Language Technology Infrastructures

As one of the measures which could contribute to decreasing the LT breech among wider and lesser used languages is the role of EU-funded research networks such as FLaReNet, CLARIN, HBP and META-NET [2]. Lithuania began joining international research infrastructures (RIs) in 2013. In the report by the Research Council of Lithuania, the Research and Higher Education Monitoring and Analysis Centre (MOSTA), and the Ministry of Education, Science and Sport, it is agreed that Lithuania’s participation in RIs has to be based on strategic goals of the state and long-term sustainability of RI [10]. In other words, RI should be relevant at the national level, ensure scientific excellence and effective governance, have sufficient numbers of users, long-term financing and technological development. Participation in RIs for research progress and breakthrough in human language technologies also corresponds with the strategic aims of such legislation as [6], [9], [10], [11], [12], [13], [14]. In line with this, Lithuania has so far joined two

[9] http://mwe.lt/en_US/

[10] http://sakytinistekstynas.vdu.lt/

[11] https://spacy.io/models/lt

[12] https://www.tilde.lt/snekos-technologijos

[13] https://translate.tilde.com/en

international language technology related initiatives: Common Language Resources and Technology Infrastructure (CLARIN ERIC) and the European Language Grid (ELG).

Lithuania became a full member of the European Research Infrastructure for Language Resources and Technology CLARIN ERIC in 2014. At present, CLARIN-LT is a consortium of five institutions, which maintains a repository[14] and provides open access (under the academic, public and restricted licenses) to specialized and well-annotated language resources used by language researchers, teachers and students of various Lithuanian higher educational institutions. Besides, CLARIN-LT centre is actively involved in knowledge sharing activities. In 2020, CLARIN-LT became a member of the CLARIN Knowledge Centre for Systems and Frameworks for Morphologically Rich Languages (SAFMORIL) coordinated by the University of Helsinki[15] by committing to share the knowledge in corpus linguistics and natural language processing methods for Lithuanian.

The maturity of the infrastructure has been acknowledged by the Research Council of Lithuania which recommends including CLARIN-LT into the renewed "Lithuanian Research Infrastructures Roadmap for 2020-2023". In the nearest future, CLARIN-LT aims to achieve the Service Providing Centre (CLARIN B centre) certification, especially important for the full integration into the international infrastructure. CLARIN B centres guarantee sustainable storage of language tools, resources and open access to other CLARIN services for various research communities which would increase the visibility and uptake of Lithuanian contribution to LT development on national and international levels. Lithuania's membership fee for CLARIN ERIC paid by the Ministry of Education, Science and Sport is ensured until 2021, thus, the continuity of CLARIN-LT activities is largely dependent on the strategic decisions of the ministry.

In 2019, the Institute of Lithuanian Language signed a subcontract with the EU-funded project European Language Grid (2019-2021) becoming one of the National Competence Centres of the network[16]. By establishing a scalable cloud platform, ELG aims to become the leading platform for Language Technology in Europe offering both commercial and non commercial LT communities to store, use and promote their services. Currently, there are 32 National Competence Centres responsible for implementing the successful operation of ELG and pursuing the main goals. Being in their active stage of establishment, the Lithuanian National Competence Centre provides information on the national level about the ELG consortium, the European Language Grid cloud platform and organizes knowledge sharing events.

In addition to joining international initiatives, three national (Raštija LT[17], LKS-SAIS[18], E.kalba[19]) infrastructures and a few smaller LT portals were launched or modernized in 2016-2020. Services offered by international and national research and LT related infrastructures are increasingly being integrated in university studies, distance learning, and development of new technologies.

Moreover, the role of RIs in the present AI hype cannot be underestimated as in order to efficiently employ new machine learning methods, massive amounts of accessible and

[14] https://clarin.vdu.lt/xmlui/?locale-attribute=en

[15] https://www.kielipankki.fi/safmoril/

[16] https://www.european-language-grid.eu/

[17] raštija.lt

[18] semantika.lt

[19] ekalba.lt

quality language data are needed. Further development and modernization of RIs which may ensure better language data storage and sharing conditions are especially important for lesser used languages that strive to be visible in the digital space.

4. Language Technology Research

This section surveys different research topics in the field of language technology for Lithuanian. We have collected and analyzed papers and studies on language technology research for 2016-2019. We do not include publications that were published in 2020, as the data would only present partial information.

Figure 1 illustrates the trends, i.e. more and less popular topics in language technology research for Lithuanian as well as the progression of the topics by year. The information was retrieved from the major subscribed databases (such as arXiv, Google Scholar, IEEE Xplore, Mendeley, Scopus, Semantic Scholar, SpringerLink, Web of Science, VDU DSpace/CRIS). The search was organized in two ways: verification by entering the surnames of well-known Lithuanian language technology specialists and by using the basic terms (e.g., Lithuanian/ Lithuanian language + language technologies/ corpora/ speech/ lexical/ morphological/ multiword/ authorship/ chatbot/ wordnet/ embeddings/ media/ NER/ NLP/ NLG/ NLU/ classification/ clusterization) with AND operator). In total, 91 publications were retrieved, which were then grouped thematically according to their content and keywords. The full list of the bibliography is accessible from the CLARIN-LT repository[20]. It should be noted that the collected data includes publications not only by Lithuanian researchers, but all work on Lithuanian language technology for the discussed period.

In terms of categorization, we avoided overgeneralizing labels so that the specificity of the LT field would be reflected. Naturally, the presented taxonomy of language technology research topics is not absolute and many other classification schemes are possible. Also, the same paper is ascribed to several categories in some cases (for example, corpora and morphological analysis), thus Figure 1 does not reflect the exact number of papers in each category, but rather the thematic scope of research distributed across all papers.

Judging by the number of scientific papers for 2016-2019 in Figure 1, the most researched topic is machine translation (12 papers), where one of the languages is Lithuanian. A possible reason for this trend is that despite a considerable progress in machine translation achieved as a result of neural machine translation developments, expectations for MT quality continue to increase and numerous experimentation attempts are done testing different NMT frameworks for many languages (including Lithuanian). As a result, presently, MT research for the Lithuanian language is mostly conducted by international groups.

Other popular research topics are: traditional corpus-based research (9), word embeddings (7), analysis of multiword expressions (7), media monitoring (7), stylometry (6), morphological analysis (6), and authorship classification problems (6), followed by speech synthesis (5), speech recognition (5), and language technology overviews (5). However, if we combine speech-related topics of speech synthesis, speech recognition,

[20] http://hdl.handle.net/20.500.11821/38

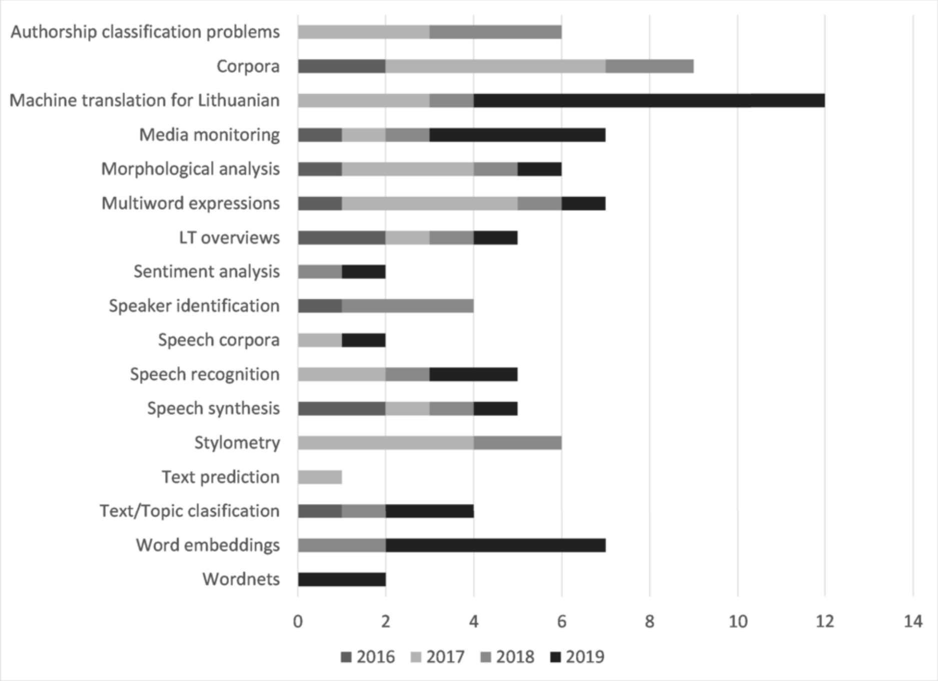

Figure 1. Research related to language technology for Lithuanian (2016-2019)

speaker identification, and speech corpora into one group of speech technologies, then this group becomes the most popular topic in this period with 18 publications.

It has also been observed that the rise of deep learning techniques created an increasing demand for specially designed data. Thus, beside traditional data sets such as treebanks, news corpora, speech corpora, spoken language corpora, or lexical databases, considerable research has been carried out on developing and implementing different word embeddings for Lithuanian as reflected in scientific publications on the topic.

It should be noted that even though several research groups and companies are involved in developing practical applications for the Lithuanian language in such worldwide trending areas as natural language generation, natural language understanding, automatic summarization, and chatbots, very few scientific research papers have been published on these topics.

In terms of methodology, we found out that the number of publications which apply deep learning and other advanced machine learning techniques has surpassed the number of publications on traditional, symbolic and rule-based approaches (53 % vs 47 %). The increasing use of machine learning methods can be highlighted as one of the achievements of the analyzed period.

As to the limitations of this survey, although we tried to include all relevant LT papers, naturally, not all publications might have been identified. Still, we hope that this overview provides a general insight on the changing trends of LT research for Lithuanian, determined by societal changes, financed projects and their aims, private sector initiatives and other factors.

5. Conclusions

The paper has shown that the last five years in Lithuania were productive in LT-related policies, infrastructure development, projects and cooperation initiatives both nationally and internationally.

Developments contributing to the integration of Lithuanian as less resourced language in the digital environment are as follows: inclusion of language technology problems into strategic documents and legislation, developing national and international language technology infrastructures, promoting the international visibility of language resources or technologies adapted to or created for the Lithuanian language, implementing Lithuanian language services in large scale projects, offering knowledge sharing services to interested parties, developing open access possibilities of language technologies and resources, and pursuing advanced research goals.

The presented review of scientific publications has shown that machine translation and speech-related research were the most researched topics in 2016-2019. In the Lithuanian language technology research, similarly to other European countries, we can see increasing number of publications that apply innovative machine learning methods for language analysis.

References

[1] European Commission. White paper on artificial intelligence–a European approach to excellence and trust. 2020. Available from: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

[2] European Parliament. European Parliament resolution of 11 September 2018 on language equality in the digital age. 2018. Available from: https://www.europarl.europa.eu/doceo/document/TA-8-2018-0332_EN.html.

[3] Skandina I. Some Highlights of Human Language Technology in Baltic Countries. In: Databases and Information Systems X: Selected Papers from the Thirteenth International Baltic Conference, DB&IS 2018. vol. 315. IOS Press; 2019. p. 18.

[4] Rehm G, Marheinecke K, Hegele S, Piperidis S, Bontcheva K, Hajič J, et al. The European Language Technology Landscape in 2020: Language-Centric and Human-Centric AI for Cross-Cultural Communication in Multilingual Europe. arXiv preprint arXiv:200313833. 2020.

[5] Utka A, Amilevičius D, Krilavičius T, Vitkutė-Adžgauskienė D. Overview of the Development of Language Resources and Technologies in Lithuania (2012-2015). In: Human Language Technologies-the Baltic perspective: Proceedings of the 7th International Conference, Baltic HLT 2016, Riga; 2016. p. 12–19.

[6] Lietuvos Respublikos Seimas. Seimo nutarimo "Dėl Lietuvių kalbos plėtros skaitmeninėje terpėje ir kalbos technologijų pažangos 2021–2027 m. gairių patvirtinimo" projektas + gairės; 2020. Available from: https://e-seimas.lrs.lt/portal/legalAct/lt/TAP/ee468a02945611eaa51db668f0092944?positionInSearchResults=0&searchModelUUID=2a0d20bc-b1f2-4573-a67a-78f9e820afdb.

[7] European Commission. EU Declaration on Cooperation on Artificial Intelligence. 2018. Available from: https://ec.europa.eu/jrc/communities/en/node/1286/document/eu-declaration-cooperation-artificial-intelligence.

[8] Ministry of the Economy and Innovation of the Republic of Lithuania. Lithuanian Artificial Intelligence Strategy. 2018. Available from: http://kurklt.lt/wp-content/uploads/2018/09/StrategyIndesignpdf.pdf.

[9] Government of the Republic of Lithuania. Information Society Development Programme for 2014-2020 'Digital Agenda for the Republic of Lithuania'. 2014. Available from: https://eimin.lrv.lt/uploads/eimin/documents/files/30310_LRV%20nutarimas(en).pdf.

[10] LMT. Lietuvos MTI kelrodis; 2015. Available from: https://www.lmt.lt/lt/mokslo-politika/moksliniu-tyrimu-infrastrukturos/lietuvos-mti-kelrodis/2358.

[11] Ministry of Finance of the Republic of Lithuania. Operational Programme for the European Union Funds' Investments in 2014-2020; 2014. Available from: https://ec.europa.eu/regional_policy/en/atlas/programmes/2014-2020/lithuania/2014lt16maop001.

[12] Research Council of Lithuania. Directions for the Lithuanian Studies Research Development 2012-2020; 2012. Available from: https://www.lmt.lt/en/doclib/ujv8xc7kauxwnnp3e6r7dw5dxwqdy6hq.

[13] Ministry of Education and Science of the Republic of Lithuania. State Lithuanian Studies and Dissemination Programme for 2016-2024; 2015. Available from: https://www.lmt.lt/en/research-commissioned-by-the-state/state-lithuanian-studies-and-dissemination-programme-for-2016-2024/803.

[14] Parliament of the Republic of Lithuania. Guidelines for the State Language Policy 2018-2022; 2018. Available from: https://e-seimas.lrs.lt/portal/legalAct/lt/TAD/cd0584707b6e11e89188e16a6495e98c?positionInSearchResults=0&searchModelUUID=2fa062c8-0d9b-4b80-9a44-8938b12fe0a4.