Databases and Information Systems X
A. Lupeikiene et al. (Eds.)

doi:10.3233/978-1-61499-941-6-18

Some Highlights of Human Language Technology in Baltic Countries

Inguna SKADIŅA[a,b,c,1]
[a] *Faculty of Computing, University of Latvia*
[b] *Institute of Mathematics and Computer Science, University of Latvia*
[c] *Tilde*

Abstract. Today, when we are surrounded by different smart devices, life of our languages is very much influenced by technologies that support them in digital environment. Language technology solutions are particularly important for languages that are small in size. This paper aims to analyze representation of languages of Baltic countries – Estonian, Latvian and Lithuanian – in digital environment. We analyze technological challenges for these languages and most important achievements (recently created language resources and tools) that help to narrow technological gap with widely used languages, facilitate use of the natural language in interaction between computer and human, and minimize threat of digital extinction. Special attention is paid to the natural language understanding task, machine translation and speech technologies.

Keywords. Language resources, natural language processing, languages of Baltic countries, machine translation, speech technologies, human-computer interaction

1. Introduction

Today we are surrounded by many smart devices that communicate with us or have interface in natural language. Many users prefer such communication in their native language. However, technologies that are widespread for widely used languages (e.g. English), such as smart home solutions or mobile applications with a speech interface, in many cases are not available for languages with a smaller number of speakers. One of the main reasons for this technological gap is a lack of natural language processing tools for the particular language. Moreover, there is a threat of digital extinction for languages that are weakly supported in digital means.

Languages of Baltic countries – Estonian, Latvian and Lithuanian – are among languages with rather small number of speakers. The number of speakers is one of the factors that influence variety and size of the language resources (monolingual, bilingual and multilingual texts, dictionaries, transcribed audio and video materials, etc.) that are available for particular language. Languages of Baltic countries are often mentioned among under-resourced or low-resourced languages. Although there is no precise definition what under-resourced language means, usually it is understood as a language that is insufficiently (in size and quality) represented in a digital form. Insufficient amount or even lack of language resources in a digital form in a turn influence the

[1] Corresponding Author, Faculty of Computing, University of Latvia, Raiņa bulv. 19, Riga, Latvia, E-mail: inguna.skadina@lu.lv

development of the language technology solutions. This is especially important today, when deep learning approach becomes dominant in language technologies.

At the beginning of this decade, the META-NET Network of Excellence forging the Multilingual Europe Technology Alliance[2] conducted a study on 30 European languages and the level of support these languages receive through language technologies. This survey was published in a book series, describing national language technology landscapes. The Whitepapers contain general facts about each language and describe recent developments in the language technology and core application areas of language and speech technology. The language technology landscape of Baltic languages was described in three volumes: “Estonian Language in Digital Age” [1], “Latvian Language in Digital Age” [3] and “Lithuanian Language in Digital age” [4].

The Whitepapers also present a cross-language comparison ranking the respective language within four key areas: machine translation, speech processing, text analysis, and resources. Table 1 summarizes ranking of language technology support for languages of Baltic countries presented in the Whitepapers. The support for Latvian and Lithuanian language in all four key areas is assessed as weak. For Estonian, a support for speech and text resources and speech processing is assessed as fragmentary. However, support for machine translation and text analysis support is assessed as weak.

Table 1. Language technology support for languages of Baltic countries presented in META-NET Whitepapers.

Language	Speech and text resources	Text analysis	Machine translation	Speech processing
Estonian	fragmentary	weak support	weak support	fragmentary
Latvian	weak support	weak support	weak support	weak support
Lithuanian	weak support	weak support	weak support	weak support

This paper extends analysis of recent achievements (mostly after the publication of Whitepapers) in human language technologies in three Baltic states presented in [1] and highlights some most important and some most recent achievements that allow to narrow the so-called technological gap in language technology field. The analysis shows that although there is still a gap between well represented English language and less-resourced languages of Europe, the language technologies in the Baltic countries have made a big step further to overcome this gap and thus making communication between users and computers easier and more attractive.

2. Language Policy and Major Activities

The necessity of language technology support for official languages in digital means has been recognized by national governments and is reflected in language policy documents of particular country. Moreover, multilingualism and its support through technologies is among priorities of the European Union (EU).

[2] http://www.meta-net.eu

2.1. International Initiatives

Languages and linguistic diversity always have been among priorities of the European Union multilingualism policy. On September 11, 2018 the European Parliament adopted resolution on language equality in digital age[3]. This resolution points on “a widening technology gap between well-resourced and less-resourced languages” and regrets “that more than 20 European languages are in danger of digital extinction”. The resolution calls the Commission “to make as a priority of language technology those Member States which are small in size and have their own language”.

Language equality and support of multilingualism through language technology has been also among topics of the EU research programmes. For instance, in recent project call of the EU research and innovation programme Horizon 2020[4] among other priorities an importance of the language technologies has been highlighted though the targeted topic of multilingual next generation internet. Activities under this topic aim to support technology-enabled multilingualism for an inclusive Digital Single Market and facilitate development of technologies that enable every European citizen to access content and engage in communication without a language barrier. Seven new projects from this call will start in 2019.

2.2. Infrastructural Developments

The fundamental support for languages in a digital environment is provided through research infrastructures, such as CLARIN and ELEXIS.

CLARIN[5] – the European research infrastructure for language resources and technology – started “from the vision that all digital language resources and tools from all over Europe and beyond are accessible through a single sign-on online environment for the support of researchers”. Today CLARIN consortium provides easy and sustainable access to digital language data in many languages [5]. All three Baltic states are members of CLARIN ERIC: Estonia was among countries that established CLARIN ERIC in 2012, Lithuania joined CLARIN ERIC in 2015 and Latvia joined in 2016. Estonia and Lithuania have established national CLARIN resource centers, while in Latvia the center is currently under construction[6]. However, researchers of Latvia already now can benefit from CLARIN tools and resources through the single sign-on through Latvian academic identity federation LAIFE.

ELEXIS[7] – the European Lexicographic Infrastructure – is a Horizon 2020 project that aims to create a sustainable infrastructure for efficient access to high quality lexical data and helps to bridge the gap between different scholarly communities working on lexicographic resources [6]. The Institute of the Estonian Language is the only project partner from Baltic countries. However, universities and research centers that are not project partners, can also benefit from the project already now. For instance, one of the

[3] http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P8-TA-2018-0332+0+DOC+XML+V0//EN&language=EN

[4] ICT-29-2018

[5] https://www.clarin.eu/

[6] CLARIN Estonia: https://keeleressursid.ee/en/clarin, CLARIN-LT: http://clarin-lt.lt/, CLARIN-LV: http://www.clarin.lv

[7] https://elex.is/

project partners, company Lexical Computing, through ELEXIS project provides free access to the Sketch Engine tool[8] to all academic institutions located in the EU.

An approach driven by practical needs has been taken by ELRC[9] – European language resource coordination – initiative [7]. The ELRC network supports collection and maintenance of the language resources in official languages of the EU with the aim to help to improve the quality, coverage and performance of the automated translation solutions of Connecting European Facility (CEF) digital services. In this initiative each country is represented by one technological representative and one representative from the public services administration. ELRC maintains repository[10] for documenting, storing, browsing and accessing language resources that are considered useful for feeding the CEF automated translation platform.

2.3. National Initiatives

For low-resourced languages support from the government is crucial for their sustainable long-term survival and life in digital means, realized through research and language technology development activities.

Such support has been provided in Estonia through National Programme for Estonian Language Technology (NPELT) since 2006. NPELT aims to provide language technology means that enable successful operation of the Estonian language in the digital world. The programme funds different language technology research and development activities - from the compilation of resources to the creation of application prototypes. Continuous language technology support from NPELT for more than 10 years has resulted in many language resources and tools and makes position of the Estonian language technology stable between the languages with the same number of speakers. Recently started NPELT programme for period 2018-2027 provides sustainable means for LRT development and will strengthen Estonian language position in digital environment. The outcome of NPELT programmes – language resources and tools - are freely available to everybody through the website of Center of Estonian Language Resources[11] (established in 2008).

Similarly to Estonia, in Lithuania since 2012 research and development in a field of human language technologies is funded through national programs for the Lithuanian language support in information society. In 2013 the State Commission of the Lithuanian language issued "Guidelines for Lithuanian Language Technologies development 2014 2020" where machine translation, speech analysis, dialogue systems, automatic summarization, semantic technologies, advanced text analysis, compilation of language resources, and others, are defined as priorities. Two national infrastructures - *raštija.lt* (Integrated Information System of the Lithuanian language and language resources) and LKSSAIS (Information system for syntactical and semantical analysis of the Lithuanian language)[12] – support access to tools and resources created through national programs. Five new projects to support Lithuanian language in information society were launched in 2018. These projects aims at development of syntactic and semantic analyzers (SEMANTICS 2), Lithuanian language speech services (LIEPA 2) and machine translation systems and localization services, as well as, support creation

[8] https://www.sketchengine.eu/

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

[10] ELRC-SHARE repository: https://elrc-share.eu/info/

[11] https://keeleressursid.ee/en/

[12] http://semantika.lt/

of information systems for integrated Lithuanian language resources (RAŠTIJA 2) and Lithuanian Language Resources (E.kalba)[13] .

The importance of the language technologies for the long-term survival of the Latvian language has been recognized in the State Language Policy Guidelines for 2015-2020. Research in language technologies has been supported through the State Research Programmes, EU Structural Funds Programmes, grants from the Latvian Science Council, EU FP7 and Horizon 2020 Programmes. Although language technologies to some extent are presented in state research programmes for Latvian language support and ICT, Latvia lacks dedicated language technology program. As a result, research and development activities in human language technologies and creation of language resources are fragmented and in many cases insufficiently supported.

Currently two large-scale research and development projects support creation of missing language resources for Latvian and development of AI-based language technology solutions. The project “Full Stack of Language Resources for Natural Language Understanding and Generation in Latvian” aims to create a complex, multi-layered set of essential Latvian language resources (corpora, treebanks, lexicons, etc.) to demonstrate the potential of these advanced language resources though creation of an innovative NLU and NLG technology [8]. The project “Neural Network Modelling for Inflected Natural Languages” aims to research novel models for applying neural network technologies for core language technology tasks – written language processing, speech processing – and advanced applications – machine translation and human-computer interaction. Some results of these projects are presented in next chapters of this paper.

3. Language Resources

Usually life of the natural language in digital means starts with language resources – digitized books and newspapers, folklore materials, dictionaries, etc. These resources serve not only for a general public, but also for research and development of language technology solutions. In Language Whitepapers several categories of language resources were analyzed – corpora (text, speech, parallel), lexical resources and grammars. While monolingual written language corpora for languages of Baltic countries were rather well represented when compared to languages of similar size, availability of parallel corpora was weak. Moreover, speech corpora for Latvian and Lithuanian were not available. In this chapter some recent important achievements that advance language technologies in Baltic countries are presented.

3.1. Corpora

Text corpora have been developed for languages of Baltic countries already for several decades. Corpora as well as other language resources for Estonian are listed at the website of the Center of Estonian Language Resources. The repository lists 65 corpora, including general corpora and domain-specific corpora; monolingual corpora and

[13] https://www.cpva.lt/lt/veikla/paramos-administravimas/es-fondu-investicijos-q6t3/finansuojamos-sritys/informacine_visuomene/lietuviu-kalba-informacinese-technologijose.html

parallel corpora. The Web13 corpus (etTenTen)[14] is perhaps the largest Estonian language corpus. It is morphologically annotated and contains 270 million tokens (about 22 million sentences).

Different Latvian language corpora are listed at *korpuss.lv* website. The modern Latvian is presented through the Balanced Corpus of Modern Latvian [9], which is currently being extended to 10 million running words. The *korpuss.lv* website also lists corpus of historical texts, parliamentary debates corpus, speech corpus and some other corpora.

Corpus of Contemporary Lithuanian Language DLTK [10] is the largest corpus of the Lithuanian language which contains 102 million running words. The corpus represents modern Lithuanian language and includes different genres and domains.

First text corpora were mainly plain text documents. Today many corpora have a morphological annotation (mostly automatic, sometimes manual), some corpora are syntactically or even semantically annotated (mostly manual annotation). Such annotation is useful for linguistic studies as well as serves for the development of corpus-based language processing tools.

The Universal Dependencies Framework[15] is widely used for syntactic annotation today. Estonian, Latvian and Lithuanian are among 60 languages represented in this format. The Estonian Treebank in a form of Universal Dependencies (UD)[16] currently contains 24,752 sentences (about 339,000 tokens). It is automatically created from the Estonian Dependency Treebank [11]. The Latvian UD treebank is created from the data in Latvian Treebank which is annotated according to a hybrid dependency-constituency grammar model [12]. The current treebank (v2.2) consists of 7,703 sentences (110,636 tokens). Lithuanian dependency treebank *Alksnis* was created in 2015-2016, it contains 2350 sentences annotated in Prague Markup Language format and PAULA XML format [13]. Recently these treebanks (together with treebanks of other EU official languages) were used in experiment to train parsers and automatically annotate JRC DGT parallel corpus of European law [14].

The most recent attempts in corpus annotation include multi-layered corpus. Such corpus contains several annotation layers, e.g., Universal Dependencies, FrameNet, PropBank and Abstract Meaning Representation [8]. Multi-layered corpus for Latvian will contain about 10-15 thousand sentences (annotated at all layers) by the end of 2019. Such corpus is useful for natural language understanding task, as well as, for cross-lingual and multilingual applications (see section 4.4).

Different parallel corpora are listed in CLARIN VLO and META-SHARE catalogues. Many of them include English or other widely spoken language as a source language. However, there are not so many parallel corpora between languages of Baltic countries. Moreover, parallel corpora that contain texts originally written in these languages, are rare. One such corpus is Latvian-Lithuanian parallel corpus LiLa, which contains about 8.7 million tokens, more than 56% of texts are originally written in one of Baltic languages [15].

[14] https://metashare.ut.ee/repository/browse/eesti-veeb-2013-ettenten-korpus-morfoloogiliselt-uhestatud/a59975e21a1011e7a6e4005056b400248c925e498de148909ee9b3e941de2aa0/

[15] http://universaldependencies.org/

[16] https://github.com/EstSyntax/EstUD

3.2. Lexical Resources

Digital lexical resources are among language resources that have been developed for the long time. Different lexical resources are available for each of languages. However, there are still many gaps in both – monolingual and multilingual lexicons (e.g., terminology databases or translation dictionaries between less used language pairs, etc.).

Today *tezaurs.lv* is the largest open lexical database for Latvian. It aims to be the central computational lexicon for Latvian, bringing together all Latvian words and frequently used multi-word units, allowing for the integration of other LT resources and tools. Today it contains 295,760 lexical entries that are compiled from more than 280 sources. *tezaurs.lv* is popular not only among researchers, but also widely used by general public – journalists, students and many others, receiving more than 2000 requests each day [16]. The dictionary is enriched with phonetic, morphological, semantic and other annotations and enhanced with language processing tools allowing generation of inflectional forms and selection of corpus examples on the fly. It is available also as an API for integration into third-party applications.

Estonian META-SHARE node lists 59 lexical conceptual resources – monolingual and bilingual dictionaries, wordlists, terminological databases, etc. The most downloaded resource is Estonian Frequency Dictionary (one of the first lexical resources for Estonian), while the most viewed is Estonian-Russian Dictionary.

Important lexical resource is WordNet – database of words grouped into sets of cognitive synonyms that are called synsets. These synsets are linked by conceptual-semantic and lexical relations. Development of the Estonian WordNet [17][17] has started already in 1996, currently it contains 115,318 keywords and 84,150 synsets. The WordNet is linked with other WordNets of Nordic countries [17]. There is no WordNet for Latvian, while Lithuanian WordNet [19] contains about 15 thousand synsets.

4. Technologies and Tools

Last few years have been challenging not only for language technology developers, but also for many other fields of computer science, since artificial intelligence, namely deep machine learning, has become popular. It has a great impact on almost every area of language technology, but especially on machine translation, human-computer interaction and automatic speech recognition.

4.1. Toolkits for Natural Language Processing

The text analysis category in the Whitepapers was assessed by quality and coverage of (1) existing text analysis technologies (morphology, syntax, semantics) and (2) text corpora, lexical resources and grammars. For languages of Baltic countries different basic language processing tools, such as tokenizers, morphological analyzers, taggers and spelling checkers, etc., are created. From the user perspective (both computer scientists and digital humanities) it is useful to compile tools for common language processing tasks in a single toolkit. The well know toolkit is Natural Language Toolkit

[17] http://www.cl.ut.ee/ressursid/teksaurus/

(NLTK) – a platform for building Python programs to work with a natural language [20]. The similar Python library – EstNLTK toolkit – has been developed now for Estonian [21]. The EstNLTK includes tools for sentence and word tokenization, morphological analysis and synthesis, correction of spelling errors and named entity recognition. Different Latvian processing tools are included in *nlp-pipe*[18] – a modular pipeline for text tokenization, tagging, parsing and named entity recognition [22].

4.2. Best Machine Translation for Complex Less Resourced Languages

Machine translation (MT) was among the areas that was mentioned in Whitepapers as insufficiently supported for all three languages of Baltic countries. However, this situation has changed recently – machine translation solutions that translate between English and languages of the Baltic states not only provide better translation as MT engines by global companies, but also have been recognized among the best systems in international news translation shared tasks (Figure 1).

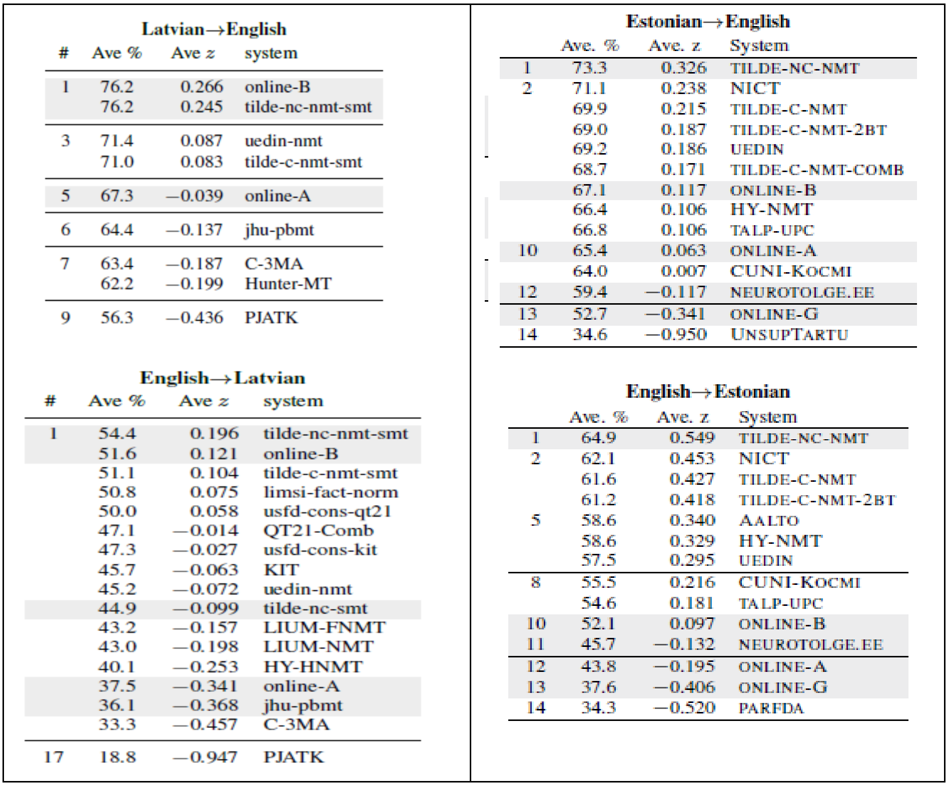
Translation into under-resourced morphologically rich languages, always has been recognized as a problem, often Baltic and Finno-Ugric languages are mentioned as most complicated cases. First achievements using statistical machine translation (SMT) were reported in 2014, when SMT systems created in Latvia were applied for translation between English and languages of Baltic states and demonstrated better results as *Google* and *Microsoft* in translation of general domain texts [23]. Obtained results demonstrated usefulness of machine translation and allowed to create public MT platforms. In Latvia *hugo.lv* is a free public sector's machine translation service to translate texts, documents and websites from Latvian into English and vice versa, as well as from Latvian into Russian. Moreover, English-Latvian-English MT system was specially designed for the 2015 Presidency of the Council of the European Union. The tool assisted staff members, translators, EU delegates, journalists, and other visitors at EU Council Presidency events.

For Lithuanian MT system *versti.eu* translates general, IT and legal domain texts between Lithuanian and English, and general and legal texts between Lithuanian and French.

After the paradigm shift from SMT to neural machine translation (NMT), EU presidency translators were developed for Estonian, Bulgarian and Austrian Presidency of the Council of the EU [24]. The systems combine the European Commission's eTranslation service with a set of customized, domain adapted NMT systems.

Complexity and small amount of training data has attracted organizers of WMT (workshop/conference on MT) shared task to include Latvian (in 2017) and Estonian (in 2018) in news translation shared task. The machine translation solutions between English and Latvian developed in Latvia for WMT 2017 were among the best, wining not only systems developed by well-known research teams, but also by global industry players [25]. Systems developed by researchers of Baltic countries for WMT 2018 were again among the best for English-Estonian and Estonian-English translation pairs [26].

[18] http://nlp.ailab.lv/



Latvian→English

#	Ave %	Ave *z*	system
1	76.2	0.266	online-B
	76.2	0.245	tilde-nc-nmt-smt
3	71.4	0.087	uedin-nmt
	71.0	0.083	tilde-c-nmt-smt
5	67.3	−0.039	online-A
6	64.4	−0.137	jhu-pbmt
7	63.4	−0.187	C-3MA
	62.2	−0.199	Hunter-MT
9	56.3	−0.436	PJATK

English→Latvian

#	Ave %	Ave *z*	system
1	54.4	0.196	tilde-nc-nmt-smt
	51.6	0.121	online-B
	51.1	0.104	tilde-c-nmt-smt
	50.8	0.075	limsi-fact-norm
	50.0	0.058	usfd-cons-qt21
	47.1	−0.014	QT21-Comb
	47.3	−0.027	usfd-cons-kit
	45.7	−0.063	KIT
	45.2	−0.072	uedin-nmt
	44.9	−0.099	tilde-nc-smt
	43.2	−0.157	LIUM-FNMT
	43.0	−0.198	LIUM-NMT
	40.1	−0.253	HY-HNMT
	37.5	−0.341	online-A
	36.1	−0.368	jhu-pbmt
	33.3	−0.457	C-3MA
17	18.8	−0.947	PJATK

Estonian→English

	Ave. %	Ave. z	System
1	73.3	0.326	TILDE-NC-NMT
2	71.1	0.238	NICT
	69.9	0.215	TILDE-C-NMT
	69.0	0.187	TILDE-C-NMT-2BT
	69.2	0.186	UEDIN
	68.7	0.171	TILDE-C-NMT-COMB
	67.1	0.117	ONLINE-B
	66.4	0.106	HY-NMT
	66.8	0.106	TALP-UPC
10	65.4	0.063	ONLINE-A
	64.0	0.007	CUNI-KOCMI
12	59.4	−0.117	NEUROTOLGE.EE
13	52.7	−0.341	ONLINE-G
14	34.6	−0.950	UNSUPTARTU

English→Estonian

	Ave. %	Ave. z	System
1	64.9	0.549	TILDE-NC-NMT
2	62.1	0.453	NICT
	61.6	0.427	TILDE-C-NMT
	61.2	0.418	TILDE-C-NMT-2BT
5	58.6	0.340	AALTO
	58.6	0.329	HY-NMT
	57.5	0.295	UEDIN
8	55.5	0.216	CUNI-KOCMI
	54.6	0.181	TALP-UPC
10	52.1	0.097	ONLINE-B
11	45.7	−0.132	NEUROTOLGE.EE
12	43.8	−0.195	ONLINE-A
13	37.6	−0.406	ONLINE-G
14	34.3	−0.520	PARFDA

Figure 1. Results of WMT17 [26] and WMT18 [26] News Translation Task between English and Estonian/Latvian.

4.3. Speech Technologies

In Whitepapers speech processing was evaluated by quality and existence of speech recognition and synthesis technologies, as well existence and quality of speech corpora. Research and development on speech technologies have been known as a success of Estonia for a long time. Work on speech synthesis started already in 1980-ies and has received national scientific prize. Today speech synthesis solutions are available for all three languages and are integrated in different use cases. For instance, Estonian speech synthesis is used at Estonian National library to produce an audio version of electronic text [34], while Lithuanian synthesizer is integrated in a website of newspaper *Lietuvos Žinos*[19].

Where it concerns Estonian speech recognition, some most recent achievements include real time speech recognition, content search in audio and speech transcription system for Estonian (e.g. [28], [29], [30]).

Latvian and Lithuanian for many years were not so well represented in speech technologies. Both Baltic languages were assessed as “weak or no support” in speech processing mainly because of lack/weak speech recognition support.

The lack of speech corpus was among the reasons, why speech recognition solutions in these countries were not available for a long time. The situation changed

[19] https://www.lzinios.lt/lzinios/index.php

when transcribed corpus of spoken Latvian was created [31]. Although the corpus is rather small – it contains only 100 hours of transcribed speech, it was good starting point for development of several speech recognition systems for Latvian [32], [33]. The output of these systems are comparable with the state of the art. Moreover, the systems demonstrate significantly better results as speech recognition solution developed by *Google* [32].

Table 2. Evaluation of Latvian and Lithuanian speech recognition solutions on general domain test sets [32]

Language	**WER[20]: Google Cloud Speech**	**WER: Tilde**
Latvian	33-44	16.9
Lithuanian	27-40	23.3

In Lithuania work on speech technologies has started already in between 70ies-80ies of 20-century. However, the Lithuanian speech recognition research and development activities were significantly advanced after creation of the speech corpus Liepa[21] allowing to create speech recognition solutions for Lithuanian in better quality as systems developed by global companies [32] (Table 2).

4.4. Natural Language Understanding and Generation

Semantic analysis and representation of meaning are indispensable constituents for natural language understanding task (NLU). Abstract Meaning Representation, AMR [35], is a semantic representation language used for logical representation of sentence meaning. Researchers from Latvia have successfully participated in international competitions related to natural language understanding and generation (NLG) tasks. At SemEval-2016 the top result was achieved in the task on Meaning Representation Parsing [36], while at SemEval-2017 - the top result was in the subtask on AMR-to-English Generation [37].

Initially AMR Bank was created manually for English. Later this representation was adapted and validated for other languages, e.g. French, Spanish, Czech, and others. Currently AMR is being tested also for Latvian. Results in SemEval competitions give confidence that it is worth to develop further the combined machine learning and grammar-based approach for NLU and NLG. Moreover, they demonstrate that AMR, complemented by FrameNet, Universal Dependencies, Grammatical Framework and other state-of-the-art syntactic and semantic representations, is emerging as a powerful interlingua for cross-lingual applications.

4.5. Human-Computer Interaction

With the renaissance of artificial intelligence and availability of computational resources that have made deep learning techniques applicable to natural language processing tasks, the human-computer interaction, and in particular virtual employees communicating in natural language have become actual topics again. The global success stories, such as IBM Watson, Apple Siri, Microsoft Cortana, Amazon Alexa and IPsoft Amelia, have raised a global interest in this field, including Baltic states.

[20] WER - word error rate

[21] https://www.raštija.lt/liepa

Communication between human and computer in Baltic countries have been studied for many years (e.g. [38], [39]). Today several task oriented virtual assistants can communicate in Latvian [40], Lithuanian or Estonian[22] helping users to find answers for particular problem. There are virtual assistants that teach multiplication to children, helps in library or helps to learn foreign language. It becomes popular to use virtual assistants in customer care, especially in telecommunications, insurance, education and travel domains.

Several virtual assistants are developed for public sector. Bilingual (Lithuanian and English) virtual assistant serves at the Migration Department at the Ministry of the Interior in Lithuania[23]. Recently Latvian virtual assistant Una[24] started work at the Register of Enterprises of the Republic of Latvia helping users to find answers to the questions related to registration or closing enterprise.

However, natural language understanding is still not solved problem and thus a lot of work needs to be done to create technologies for deeper language understanding and human-computer interaction.

5. Conclusion

Although languages of Baltic countries – Estonia, Latvia and Lithuania – are represented by a rather small number of speakers and often are called under-resourced, all three languages are represented in digital world not only by digital libraries of texts and language resources (corpora, lexicons, etc.), but also by fundamental language technologies, such as spelling checkers, morphological analyzers, taggers and parsers.

However, the situation with more advanced technologies is different for each language. Automated translation support between English and languages of Baltic states has reached rather good quality for domains with sufficient data, while MT for domains and especially language pairs with limited language resources still needs support.

During last five years significant advancements are made in speech technologies, especially for Lithuanian and Latvian. These technologies have reached state of art quality in some use cases, but need further research and adaptation for specific use cases, domains and environments.

The natural language understanding, which is a key for successful life of languages in a digital world and currently is a hot research topics in a world, has also reached some initial achievements in Baltic states, but needs much more attention and deeper research activities in a nearest future. It is also important to fill the gaps in basic language resources (e.g. WordNet for Latvian) and technologies (e.g. tools for deep language analysis) that needs to be developed for support of NLU task.

There are significant achievements in all three Baltic countries during last five years. However, language technologies still needs to be priority with strong national support for research and development activities in all three Baltic states to facilitate life of languages in digital world and help to narrow technological gap with widely spoken languages.

[22] https://alphablues.com/

[23] http://www.migracija.lt/index.php?2044534709

[24] https://www.ur.gov.lv/lv/

Acknowledgements

The research was developed under the University of Latvia contract no. AAP2016/B032 "Innovative information technologies" and the European Regional Development Fund under the grant agreement No.1.1.1.5/18/I/016. I am very grateful to my colleagues in three Baltic countries - Mark Fishel, Normunds Grūzītis, Tomas Krilavičius, Raivis Skadiņš, Andrius Utka, Jurgita Vaičenonienė and Kadri Vider – who helped me to identify and highlight most important achievements in Baltic states.

References

[1] I. Skadiņa. Languages of Baltic Countries in Digital Age. Proceedings of 13th International Baltic Conference, DB&IS 2018, Communications in Computer and Information Science **838** (2018), 32-40.

[2] K. Liin, K. Muischnek, K. Müürisep, K. Vider. *Eesti keel digiajastul -- The Estonian Language in the Digital Age*. Springer, 2012.

[3] I. Skadiņa, A. Veisbergs, A. Vasiļjevs, T. Gornostaja, I. Keiša, A. Rudzīte. *Latviešu valoda digitālajā laikmetā -- The Latvian Language in the Digital Age*. Springer, 2012.

[4] D. Vaišnienė, J. Zabarskaitė. *Lietuvių kalba skaitmeniniame amžiuje -- The Lithuanian Language in the Digital Age*. Springer, 2012.

[5] F. de Jong, B. Maegaard, K. de Smedt, D. Fišer and D. van Uytvanck. CLARIN: Towards FAIR and Responsible Data Science Using Language Resources. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (2018), 3259-3264.

[6] S. Krek, I.Kosem, J. P. McCrae, R. Navigli, B.S. Pedersen, C.Tiberius, T. Wissik. European Lexicographic Infrastructure (ELEXIS). *Proceedings of the 18th EURALEX International Congress* (2018), 881-891.

[7] L. Andrea, V. Mapelli, S. Piperidis, A. Vasiļjevs, L. Smal, T. Declerck, E. Schnur, K. Choukri and J. Van Genabith. European Language Resource Coordination: Collecting Language Resources for Public Sector Multilingual Information Management. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (2018), 1339-1343.

[8] N. Gruzitis, L. Pretkalnina, B. Saulite, L. Rituma, G. Nespore-Berzkalne, A. Znotins. and P. Paikens. Creation of a Balanced State-of-the-Art Multilayer Corpus for NLU. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation - LREC 2018* (2018), 4506-3264.

[9] K. Levane-Petrova, K. Līdzsvarots mūsdienu latviešu valodas tekstu korpuss un tā tekstu atlases kritēriji (The balanced corpus of modern Latvian and the text selection criteria). *Baltistica,* **8** (2012), 89-98.

[10] E. Rimkutė, J. Kovalevskaitė, V. Melninkaitė, A. Utka, D. Vitkutė-Adžgauskienė. Corpus of Contemporary Lithuanian Language – the Standardised Way. *Human Language Technologies – The Baltic Perspective: Proceedings of the Fourth International Conference Baltic HLT* (2010), 154-160.

[11] K. Muischnek, K. Müürisep, T. Puolakainen, Dependency Parsing of Estonian: Statistical and Rule-based Approaches. *Human Language Technologies -- The Baltic Perspective* (2014), 111-118.

[12] L. Pretkalnina, L. Rituma, B. Saulite Deriving Enhanced Universal Dependencies from a Hybrid Dependency-Constituency Treebank. *Text, Speech, and Dialogue* (Springer).

[13] A. Bielinskienė, L. Boizou, J. Kovalevskaitė, E. Rimkutė. Lithuanian Dependency Treebank ALKSNIS. *Human Language Technologies -- The Baltic Perspective* (2016), 107-114.

[14] N. Ljubešić and T. DGT-UD: a Parallel 23-language Parsebank. *CLARIN Annual Conference 2018 Proceedings* (2018), 147-150.

[15] A. Utka, K. Levane-Petrova, A. Bielinskiene, J. Kovalevskaite, E. Rimkute, D. Vevere. Lithuanian-Latvian-Lithuanian parallel corpus. *Human Language Technologies -- The Baltic Perspective* (2012), 260 – 264.

[16] A. Spektors, I. Auzina, R. Dargis, N. Gruzitis, P. Paikens, L. Pretkalnina, L. Rituma and B. Saulite. Tezaurs.lv: the largest open lexical database for Latvian. *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC) (2016)*, 2568-2571.

[17] N. Kahusk and K. Vider. The Revision History of Estonian Wordnet. *LDK Workshop Challenges of Wordnets* (2017), 164-173.

[18] B.S. Pedersen, L. Borin, M. Forsberg, K. Linden, H. Orav, E. Rögnvaldsson. Linking and Validating Nordic and Baltic Wordnets - A Multilingual Action in META-NORD. *Proceedings of 6th International Global Wordnet Conference: 6th International Global Wordnet Conference* (2012), 254-259.

[19] R. Garabík and I. Pileckytė, Indrė. From Multilingual Dictionary to LithuanianWordNet. *Natural Language Processing, CorpusLinguistics, E-Learning* (2013), 74-80.

[20] S. Bird, E. Klein, and E. Loper. *Natural Language Processing with Python– Analyzing Text with the Natural Language Toolkit.* O'Reilly Media, 2009

[21] S. Orasmaa, T. Petmanson, A. Tkachenko, S. Laur and H.J. Kaalep. *EstNLTK - NLP Toolkit for Estonian. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (2016).

[22] A. Znotins, E. Cirule. NLP-PIPE: Latvian NLP Tool Pipeline. *Human Language Technologies - The Baltic Perspective. Frontiers in Artificial Intelligence and Applications* **307** (2018).

[23] R. Skadiņš, V. Šics and R. Rozis. Building the World's Best General Domain MT for Baltic Languages. *Human Language Technologies – The Baltic Perspective.Proceedings of the Sixth International Conference Baltic HLT 2014* (2014), 141-148.

[24] M. Pinnis and R. Kalniņš. Developing a Neural Machine Translation Service for the 2017-2018 European Union Presidency. *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (AMTA 2018),* **vol. 2: MT Users** (2018), 72-83.

[25] M. Pinnis, R. Krišlauks, T. Miks, D. Deksne and V. Šics. Tilde's Machine Translation Systems for WMT 2017. *Proceedings of the Second Conference on Machine Translation* **2**, Volume 2: Shared Task Papers (2017), 374-381.

[26] O. Bojar, R. Chatterjee, Ch. Federmann, Y. Graham, B. Haddow, S. Huang, M. Huck, P. Koehn, Q. Liu, V. Logacheva, Ch. Monz, M. Negri, M. Post, R. Rubino, L. Specia and M. Turchi. Findings of the 2017 Conference on Machine Translation (WMT17). *Proceedings of the Second Conference on Machine Translation* (2017), 169-214.

[27] O. Bojar, Ch. Federmann, M. Mark, Y. Graham, B. Haddow, M. Huck, P. Koehn, Ch. Monz. Findings of the 2018 Conference on Machine Translation (WMT18). *Proceedings of the Third Conference on Machine Translation* (2018), 272-307.

[28] K. Kurimo, S. Enarvi, O. Tilk, M. Varjokallio, A. Mansikkaniemi, T. Alumäe. Modeling under-resourced languages for speech recognition. *Language Resources and Evaluation* (2017), 961-987.

[29] A. Paats, T. Alumäe, E. Meister, I. Fridolin. Evaluation of automatic speech recognition prototype for Estonian language in radiology domain: a pilot study. *16th Nordic-Baltic Conference on Biomedical Engineering* (2015), 96-99.

[30] T. Alumäe, O. A. Tilk, Advanced Rich Transcription System for Estonian Speech. *Frontiers in Artificial Intelligence and Applications* (307: Human Language Technologies – The Baltic Perspective) (2018), 1-8.

[31] M. Pinnis, I. Auziņa, K. Goba. Designing the Latvian Speech Recognition Corpus. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (2014)*,* 1547-1553.

[32] A. Salimbajevs. Creating Lithuanian and Latvian Speech Corpora from Inaccurately Annotated Web Data. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (2018), 2871-2875.

[33] A. Znotins, K. Polis, R. Dargis. Media monitoring system for Latvian radio and TV broadcasts. *Proceedings of the 16th Annual Conference of the International Speech Communication Association (INTERSPEECH)* (2015)*,* 732-733.

[34] M. Mihkla, I. Hein, I. Kiissel. Self-Reading Texts and Books. *Frontiers in Artificial Intelligence and Applications* (307: Human Language Technologies – The Baltic Perspective) (2018), 79-87.

[35] L. Banarescu, C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and N. Schneider. Abstract Meaning Representation for Sembanking. *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse* (2013).

[36] G. Barzdins and D. Gosko. RIGA at SemEval-2016 Task 8: Impact of Smatch extensions and character-level neural translation on AMR parsing accuracy. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval)* (2016).

[37] N. Gruzitis, D. Gosko, G. Barzdins. RIGOTRIO at SemEval-2017 Task 9: Combining machine learning and grammar engineering for AMR parsing and generation. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval)* (2017)*,* 924-928.

[38] M. Koit. Modelling Human-Computer Interaction. *SCAI* (1997), 275-276.

[39] M. Koit, T. Roosmaa and H. Õim. Knowledge representation for human-machine interaction. *Proceedings of the international conference on knowledge engineering and ontology development: International conference on knowledge engineering and ontology development* (2009), 396-399.

[40] A.Vasiljevs, I. Skadina, D. Deksne, M. Kalis and I. Vira, Application of Virtual Agents for Delivery of Information Services. *New Challenges of Economic and Business Development – 2017* (2017), 667-678.