Human Language Technologies – The Baltic Perspective
I. Skadiņa and R. Rozis (Eds.)

doi:10.3233/978-1-61499-701-6-3

Filling the Gaps in Latvian BLARK: Case of the Latvian IT Competence Centre

Inguna SKADIŅA[a], Ilze AUZIŅA[b], Daiga DEKSNE[a], Raivis SKADIŅŠ[a], Andrejs VASIĻJEVS[a], Madara GAIĻŪNA[c], Ieva PORTNAJA[c]
[a] *Tilde, Latvia*
[b] *Institute of Mathematics and Computer Science, University of Latvia*
[c] *Information Agency LETA*

Abstract. We present the results of the Latvian IT Competence Centre (IT CC) in developing several essential language technologies and applications. 11 language technology projects have been completed in the first phase of the IT CC work. We describe how IT CC has contributed to filling in the gaps and improving the quality of the basic language technologies for Latvian in speech processing, machine translation, parsing and grammar checking, intelligent media monitoring and multi-modal human-computer interaction.

Keywords: Latvian language resources and tools, speech technologies and corpora, machine translation, parsing, human-computer interaction

1. Introduction

In 2010, Latvian research institutions and major information technology (IT) companies founded the IT Competence Centre (IT CC) with the aim of supporting a long term cooperation between research organisations and industry in order to create innovative technologies and prototypes of internationally competitive IT products. The IT Competence Centre was one of the six Competence Centres that were co-funded by the European Structural Funds Programme from 2011 till 2015.

The IT Competence Centre has set two research directions as its main priorities: business process analysis and language technologies (LT). 11 projects have been completed so far in different disciplines related to language technologies and natural language processing: multilingual ontology-based e-learning, reading and comprehension analysis methods and tools, parsing and grammar checking, speech resources and technologies, intelligent media monitoring, machine translation and multi-modal human-computer interaction. For many of these topics, only some initial studies had been carried out before the CC programme started.

In this paper, we present the most important results achieved by the IT CC in the area of language resources and technologies (LRT). These results have enabled the IT CC to fill some key gaps in the Latvian Basic Language Resource Kit (BLARK) [1] – the minimal set of technologies, tools and resources that are needed for any language for research and application development.

We also show how the resources and technologies that have resulted from the research activities were used in developing novel IT applications by IT CC industry members – language technology company Tilde and news agency LETA.

2. Latvian BLARK

The progress in developing Latvian language resources and tools is rather well documented in the proceedings of previous Baltic HLT conferences ([2], [3], [4]) and the comparative study by META-NET [5]. A common conclusion of these studies has been that several basic language resources and tools are still missing for the Latvian language (see Table 1).

Table 1. Level of language technology support for Latvian in 2012 according to the META-NET study [5]: 0 – very low, 6 – very high

	Quantity	**Availability**	**Quality**	**Coverage**	**Maturity**	**Sustainability**	**Adaptability**
Language Technology: Tools, Technologies and Applications							
Speech Recognition	0	0	0	0	0	0	0
Speech Synthesis	2	3	4	3	4	3	4
Grammatical analysis	2.5	2	3	3.5	4	3	4
Semantic analysis	1	0	0	0	0	0	0
Text generation	1	2	1	2	2	1	2
Machine translation	3	4	3	3	4	3	4
Language Resources: Resources, Data and Knowledge Bases							
Text corpora	2	4	4	3	3	3	4.5
Speech corpora	1	0	1	1	1	1	3
Parallel corpora	1	3	2	2	3	4	4
Lexical resources	3	3.5	4	3	4.5	4.5	4.5
Grammars	2	1	3	2	3	4	3

In the comparative study of META-NET [5], Latvian was assessed as one of the three languages in the Baltic and Nordic region with weak or no support in all major LRT areas, as shown in Table 2.

Table 2. Availability of language resources and tools for languages of Baltic and Nordic countries [5].

	Excellent	**Good**	**Moderate**	**Fragmentary**	**Weak/None**
Speech Processing			Finnish	Danish, Estonian, Norwegian, Swedish	Icelandic, **Latvian,** Lithuanian
Machine Translation					Danish, Estonian, Finnish, Icelandic, **Latvian,** Lithuanian, Norwegian, Swedish
Text Analysis				Danish, Finnish, Norwegian, Swedish	Estonian, Icelandic, **Latvian,** Lithuanian
Resources			Swedish	Danish, Estonian, Finnish, Norwegian	Icelandic, **Latvian,** Lithuanian

3. Basic Language Resources and Tools Developed in the IT Competence Centre Programme

To narrow the gaps and develop missing technologies, IT CC focused on all key LRT areas – speech processing, machine translation, text analysis and language resources.

3.1. Speech Resources and Technologies

Until now, the major gap in BLARK for Latvian was the missing Automatic Speech Recognition (ASR) technology. This important gap is now filled through the results of four IT CC projects, which are presented in the following sections.

3.1.1. Speech Corpus

The main reason why there was no speech recognition system for Latvian was the lack of a sufficiently big annotated speech corpus, which is a prerequisite for ASR development. Therefore, creation of such a corpus was initiated by IT CC industry members Tilde and LETA with the involvement of the Institute of Mathematics and Computer Science of the University of Latvia (IMCS) as a cooperation partner. It was decided to build a 100-hour orthographically annotated and 4-hour phonetically annotated Latvian Speech Recognition Corpus [6]. Specification and creation of the corpus took about a year and was finished by the end of 2013. The statistics of the orthographically annotated data in the Latvian Speech Recognition Corpus are given in Table 3; the proportional distribution of speech data with respect to the gender and age of speakers and the style of speech is given in Figure 1.

Table 3. The statistics of the Latvian Speech Recognition Corpus [6]

Number of unique words	~72.5 k
Number of running words	~837 k
Total number of speakers	1,851
Men	1,016
Women	835

Table 4. Data distribution with respect to different speech segment types [6]

Type	**Total length**
Inhalation, exhalation	3 h 45 min (13,538 s)
Pauses	1 h 55 min (6,911 s)
Non-verbal segments	19 min (1,137 s)
Verbal segments	94 h 1 min (338,500 s)
The whole corpus	100 h 1 min (360,086 s)

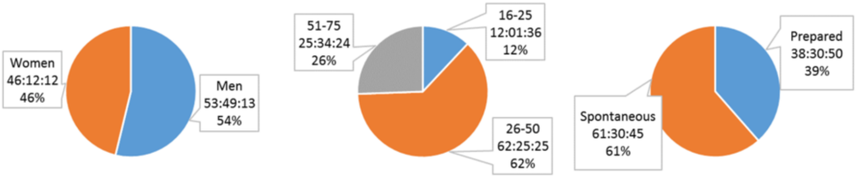

Figure 1. Data distribution with respect to the 1) gender and 2) age of speakers and the 3) style of speech [6].

The corpus includes both verbal and non-verbal segments (see Table 4). Most audio data included in the corpus have a frequency of 44.1 kHz with 16 bits allocated per sample. The data included in the corpus have been selected to contain different noise types: (1) audio data without background noise (studio recordings / outside a studio without background noise), (2) data recorded in a studio with background noise (e.g. physiological noise), (3) data recorded outside a studio with background noise (e.g.

office noise), (4) street noise (i.e. noise caused by vehicles, pedestrians, etc.), (5) noise inside a car and (6) loud music as background noise.

3.1.2. Automatic Speech Recognition

The Latvian Speech Recognition Corpus was used in two separate IT CC projects for building Latvian ASR systems. Language technology company Tilde created Latvian ASR systems for audio transcription [7], [8], [9], [10] and for text dictation [11]. The news agency LETA, together with IMCS, created an ASR system for keyword recognition in media monitoring [12], [13] (see Section 4.1).

Tilde's Latvian ASR for audio transcription was evaluated on two test sets: (1) about 1 hour of lecture recordings collected from the Web and (2) a very small (23 min.) corpus of Latvian speech that was obtained by recording various people reading articles from Web news portals. The results of the evaluation are summarised in Table 5. Word error rates (WER) of 19-21% can be considered to be rather high. However, our analysis of misrecognised words showed that only 47% of them make utterances difficult or impossible to understand. This is because 42% of errors are in the word endings, which in most cases are easy to identify and correct.

Table 5. Evaluation of Word error rate in Tilde's Latvian ASR for audio transcription

Test Set	**OOV (out of vocabulary)**	**WER**
Lectures	3%	20.71%
News	6%	19.63%

The evaluation results encouraged to integrate Latvian ASR in real-world applications. Tilde created a public online service for audio transcription[1] and also integrated Latvian speech recognition in the software product *Tildes Birojs*[2]. In addition, an annotated audio corpus of dictated text [14] was created and used to create a dictation system [11].

3.2. Machine Translation

IT CC work in machine translation (MT) focused on researching novel multilingual methods for Latvian and other under-resourced languages. Several methods were investigated how to improve statistical MT by complementing statistical models with various linguistic knowledge. The research focused on integration of knowledge about word morphology, translation dictionaries and term dictionaries [15], treatment of special tokens (e.g. numbers, measurement units, file names, URLs etc.). A special research activity was devoted to automation of corpora collection and its processing to address the problem of data sparseness [16], [17].

The research results were validated in practical application for the localization industry by integrating MT in computer aided translation tools. Effect of MT on productivity of translators was evaluated by comparing traditional translation process using only translation memories with enhanced model where translation memories are supplemented with machine translation results. Significant productivity improvement (15-32%) was achieved for all investigated language pairs [18], [19]. The created

[1] https://www.tilde.lv/runas-atpazinejs

[2] http://www.tilde.lv/tildes-birojs-2016

general purpose MT systems outperform other known MT systems (including *Google Translate*) for all languages of Baltic states [20]. The Table 6 and the Figure 2 give more details on training data and automatic quality evaluation.

Table 6. Amount of training data and results of the automatic evaluation

MT systems	**Corpora size (sentences)**		**BLEU**
	Parallel	**Monolingual**	
English – Latvian	8.9 M	60.9 M	37.38
Latvian – English	12.7 M	66.6 M	44.15
English – Lithuanian	5.3 M	24.1 M	28.80
Lithuanian – English	5.3 M	81.0 M	38.42
English – Estonian	12.5 M	33.1 M	24.22
Estonian – English	11.5 M	107.9 M	37.97

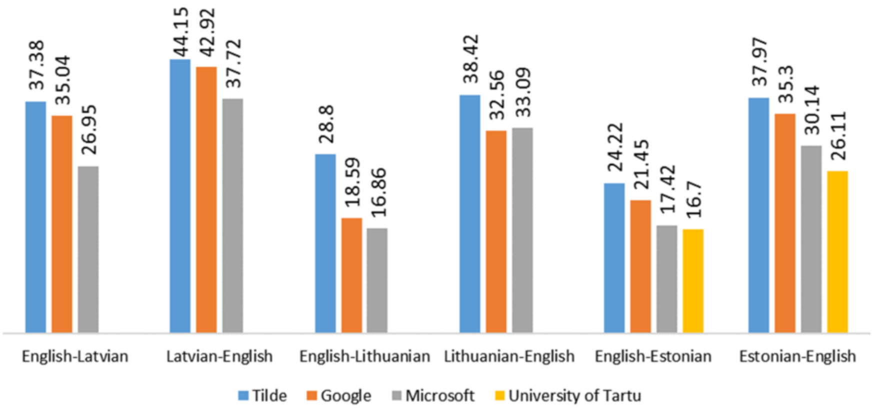

Figure 2. MT systems compared to MT systems of Google, Microsoft and University of Tartu [20].

3.3. Parsing and Grammar Checking

In scope of the IT CC project a prototype of syntactic analysis system and a prototype of grammar error correction system were created. After analysis of different formal grammars used for various languages, it was decided to base the rule format on context free grammar formalism. As Latvian has a rich morphology, which is hard to describe with non-terminal symbol names, the classic context-free grammar formalism was extended by adding syntactic roles to constituents, enabling to specify lexical constraints and constraints on morpho-syntactic feature values. 580 handwritten rules describing the correct syntactic constructions of the Latvian language were created. They cover all main constituents of a sentence, starting from simple nominal, prepositional, adverbial and verbal phrases and ending with description of complex and compound sentences. More than 1,000 textual examples and corresponding syntactic rules were collected in a database of syntactic construction examples.

For grammar checking, 263 error rules that describe incorrect syntax and depend on phrases described by correct syntax rules and 239 error rules that contain only terminal symbols were created. The four main areas (orthography, punctuation, grammar and lexicon) where errors occur and where grammar checking would be helpful were identified, and errors were clustered in 22 error types. For parsing Latvian texts, the CYK (Cocke-Younger-Kasami) algorithm was chosen as it allows partial parsing, which is useful for grammar checking [21].

The parser was evaluated using the PARSEVAL measures of precision, recall and crossing brackets [22]. The gold standard for the parser evaluation was created semi-automatically. At first, 484 sentences were parsed by a parser, and then they were corrected by a human editor. The parser achieved 89.1% recall and 89.13% precision, and brackets do not cross in 87.6% of the cases. For the grammar checker, we collected and manually annotated a corpus containing sentences with different types of errors [23]. The predefined error type and the suggestion for correction were assigned to every erroneous sentence in the corpus. The grammar checker achieved 18.7% recall and 65% precision on the student paper test corpus. We also performed a human evaluation of the prototype for the grammar error correction system.

4. From BLARK to Innovative Applications and Products

4.1. System for Monitoring of Latvian Radio and Television Broadcasts

The experimental developments carried out during the IT CC project enabled the development of an automatic system for monitoring of Latvian radio and television broadcasts [12].

The system for automatic monitoring of Latvian radio and television broadcasts (more than 300 TV and radio programs) uses an automatic speech recognition (ASR) module to convert audio and video files to text and to extract more than 5000 keywords of interest. The system performs several tasks: it automatically downloads audio and video files from servers that record radio and television broadcasts; then the pre-processing and splitting of these files for ASR on multiple server instances are performed; recognised audio fragments are merged from divided fragments into a single resulting audio. In the end, the system exports processed files to a separate media monitoring system that allows to monitor media based on the preferences of the client's keyword lists, date and time [13].

The automatic broadcast monitoring system also provides a user interface for editing transcribed text – inaccurately recognised keywords and metadata (e.g. title, date, time, source and broadcast name). Although the system works only for Latvian, it could be adapted to be used for other languages as well.

4.2. System of Text Summarisation and Information Extraction

In the scope of the IT CC project, the experimental system of text summarisation and information extraction for obtaining CV-style structured information about publicly mentioned persons, organisations and their relations by analysing newswire archives in the Latvian language was developed by LETA and IMCS [24], [25]. The text analysis and CV-style text summarisation system consists of morpho-syntactic analysis, named entity recognition, coreference resolution and a semantic role labelling system based on FrameNet principles [25].

The main goal for creation of this system was to automatically process, analyse and extract information about persons and organisations from news articles. Before this system was implemented, there were approx. 25,000 unstructured person and organisation profiles in the LETA archive. As a result of the implementation of this system, previous profiles are structured, new facts are automatically added, and the number of profiles has grown three times.

The developed system allowed LETA and IMCS to successfully complete the EU ERAF funded project “Identification of relations in newswire texts and graph visualisation of the extracted relation database”[3]. Integration and development of the system continues under the Horizon 2020 Research and Innovation Action "Scalable Understanding of Multilingual Media (SUMMA)"[4].

4.3. Multimodal-interaction

As dialog-based applications like *Apple Siri* and *Microsoft Cortana* have attracted many users, human-computer interaction using a conversational interface has become a hot research topic. In Latvia, dialog-based interaction is a relatively new field of research. Thus, an IT CC project was initiated to research novel interfaces for human-computer interaction on mobile devices.

The goal of this project was to create animated virtual agents and research their usability for different applications. In the first experiments, several English-speaking, humanlike 3D agents were created to evaluate the state of the art of the existing freely available dialogue management platforms. These humanlike characters are able to hold simple natural language based conversations on predefined topics. Several virtual assistants that can be used for simple conversation or as a virtual guide were created for Android, iOS and Windows mobile platforms [26], [27].

Inspired by the first positive results and user interest (conversational agent Laura has more than 120 thousand downloads from *Google Play* store), task specific, Latvian-speaking virtual assistants were designed. Two use scenarios were investigated: (1) teaching multiplication to kids and (2) asking/telling facts about Latvia. Our main concern was the usability of the Latvian speech recogniser [7] for mobile devices, particular domains and different speakers. Another challenge was dialogue management since Latvian is an inflected free word order language, which significantly complicates the processing of user input and generation of a response.

Table 7. Evaluation results for multiplication and Latvian facts dialogues

Dialogue system	**Speech recognition**			**Correctness of dialogue**	
	Incorrect	**Partly correct**	**Correct**	**Correct**	**Incorrect**
Multiplication	9%	22%	69%	87%	13%
Latvian facts	14%	26%	14%	74%	26%

The dialogue system for the multiplication scenario was evaluated by 21 evaluators – 10 children and 11 adults. The dialogue system that tells and asks facts about Latvia was evaluated by 20 adults - 10 women and 10 men. In both scenarios, correctness of ASR output and correctness of the answer generated by the dialogue system were evaluated (see Table 7). Although the multiplication scenario requires

[3] Nr.2DP/2.1.1.1.0/13/APIA/VIAA/014

[4] http://www.summa-project.eu/

understanding of children's language, it showed better results. This could be explained by less variability in the user's answers.

5. Conclusions

In this paper, we presented an overview of Latvian language resources and tools developed in the IT Competence Centre programme in 2011-2015. We described how this programme helped to advance Latvian language technologies and fill major gaps in the Latvian BLARK. A particularly important achievement is the creation of a large annotated Latvian speech corpora and the first speech recognition systems for Latvian – the components that were completely missing in the Latvian BLARK.

Among the major success factors are focus on practical results that are applicable in real-world applications, close cooperation between IT companies and research institutions, efficient research with rapid prototyping, continuous evaluation of the applicability of proposed methods and validation of research results in applications and use scenarios.

IT CC wants to continue to be a major driver of the technological development for the Latvian language and has applied for continuous co-funding in the next Competence Centre programme 2016-2021.

Acknowledgements

The research leading to these results has received funding from the research project "Information and Communication Technology Competence Centre" of EU Structural funds, contract nr. L-KC-11-0003 signed between ICT Competence Centre and the Investment and Development Agency of Latvia.

References

[1] S. Krauwer, The Basic Language Resource Kit as the First Milestone for the Language Resources Roadmap. http://www.elsnet.org/dox/krauwer-specom2003.pdf , 2003.

[2] I. Skadiņa, I.Auziņa, N. Grūzītis, K. Levane-Petrova, G. Nešpore, R. Skadiņš, A. Vasiļjevs, *Language Resources and Technology for the Humanities in Latvia (2004–2010).* Proceedings of the Fourth International Conference Baltic HLT 2010, Frontiers in Artificial Intelligence and Applications 219 (2010), 15-22.

[3] A. Vasiļjevs, I. Skadiņa*, Latvian Language Resources and Tools: Assessment, Description and Sharing.* Human Language Technologies – The Baltic Perspective. Proceedings of the Fifth International Conference Baltic HLT 2012, Frontiers in Artificial Intelligence and Applications 247 (2012), 265-272.

[4] I. Skadiņa, I. Auziņa, G. Bārzdiņš, R. Skadiņš, A. Vasiļjevs, *Language Resources and Technology in Latvia (2010-2014).* Human Language Technologies – The Baltic Perspective. Proceedings of the Sixth International Conference Baltic HLT 2014, Frontiers in Artificial Intelligence and Applications 268 (2014), 227-235.

[5] I. Skadiņa, A. Veisbergs, A. Vasiļjevs, T. Gornostaja, I. Keiša, A. Rudzīte, *Latvian Language in the Digital Age*, Springer, 2012.

[6] M. Pinnis, I. Auziņa, K. Goba, *Designing the Latvian Speech Recognition Corpus*. Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14) (2014), 1547–1553

[7] A. Salimbajevs, J. Strigins, *Latvian Speech-To-Text Transcription Service*. Proceedings of Interspeech (2015), 722–723.

[8] A. Salimbajevs, M. Pinnis, *Towards Large Vocabulary Automatic Speech Recognition for Latvian*. Human Language Technologies – The Baltic Perspective. Proceedings of the Sixth International Conference Baltic HLT 2014 (2014), 236–243.

[9] A. Salimbajevs, J. Strigins, *Error Analysis and Improving Speech Recognition for Latvian language*. Proceedings of 10th International Conference on Recent Advances in Natural Language Processing (RANLP 2015) (2015), 563–569.

[10] A. Salimbajevs, J. Strigins, *Using sub-word n-gram models for dealing with OOV in large vocabulary speech recognition for Latvian*. Proceedings of the 20th Nordic Conference of Computational Linguistics NODALIDA 2015 (2015), 281–286.

[11] A. Salimbajevs, *Towards First Dictation System for Latvian Language*. Proceedings of the Seventh International Conference Baltic HLT 2016 (2016).

[12] R. Darģis, A. Znotiņš, *Baseline for Keyword Spotting in Latvian Broadcast Speech*, Human Language Technologies-The Baltic Perspective: Proceedings of the Sixth International Conference Baltic HLT 2014 (Vol. 268) (2014), 75-82.

[13] A. Znotiņš, K. Polis, R. Darģis, *Media Monitoring System for Latvian Radio and TV Broadcasts*. Sixteenth Annual Conference of the International Speech Communication Association (2015).

[14] M. Pinnis, A. Salimbajevs, I. Auziņa, *Designing a Speech Corpus for the Development and Evaluation of Dictation Systems in Latvian*. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016) (2016), 775–780.

[15] M. Pinnis, R. Skadiņš, *MT Adaptation for Under-Resourced Domains – What Works and What Not*. Frontiers in Artificial Intelligence and Applications, Volume 247: Human Language Technologies – The Baltic Perspective (2012), 176–184

[16] R. Skadiņš, J. Tiedemann, R. Rozis, D. Deksne, *Billions of Parallel Words for Free: Building and Using the EU Bookshop Corpus*. Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14) (2014), 1850–1855.

[17] R. Rozis, A. Vasiļjevs, R. Skadiņš, *Collecting Language Resources for the Latvian e-Government Machine Translation Platform*. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016) (2016), 1270–1276

[18] R. Skadiņš, M. Pinnis, A. Vasiļjevs, I. Skadiņa, T. Hudik, *Application of Machine Translation in Localization into Low-Resourced Languages*. Proceedings of the 17th Annual Conference of the European Association for Machine Translation EAMT 2014 (2014), 209–216

[19] M. Pinnis, R. Skadiņš, A. Vasiļjevs, *Real-world challenges in application of MT for localization: The Baltic case*. Proceedings of AMTA 2014, vol. 2: MT Users (2014), 66–79

[20] R. Skadiņš, V. Šics, R. Rozis, *Building the World's Best General Domain MT for Baltic Languages*. Frontiers in Artificial Intelligence and Applications: Volume 286. Human Language Technologies – The Baltic Perspective - Proceedings of the Sixth International Conference Baltic HLT 2014 (2014), 141–148.

[21] D. Deksne, I. Skadiņa, R. Skadiņš, *Extended CFG formalism for grammar checker and parser development*. Computational Linguistics and Intelligent Text Processing. Springer, Berlin, Heidelberg, (2014), 237–249.

[22] E. Black, S. Abney, D. Flickinger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini, and T. Strzalkowski, A procedure for quantitatively comparing the syntactic coverage of english grammars. *Proceedings of the DARPA Speech and Natural Language Workshop*, (1991), 306–311.

[23] D. Deksne, I. Skadiņa, *Error-Annotated Corpus for Latvian*. Human Language Technologies – The Baltic Perspective. Proceedings of the Sixth International Conference Baltic HLT 2014, Vol.268 (2014), IOS Press, 163-166.

[24] A. Znotiņš, P. Paikens, *Coreference resolutian for Latvian*. Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC'14) (2014), 3209–3213.

[25] P. Paikens, *Latvian Newswire Information Extraction System and Entity Knowledge Base, Speech*, Human Language Technologies-The Baltic Perspective: Proceedings of the Sixth International Conference Baltic HLT 2014, Vol.268 (2014), IOS Press, 119–125.

[26] A. Vasiļjevs, I. Vīra, *The Development of Conversational Agent Based Interface*. Human Language Technologies – The Baltic Perspective. Proceedings of the Sixth International Conference Baltic HLT 2014, Frontiers in Artificial Intelligence and Applications 268 (2014), 46–53.

[27] I. Vīra, J. Teseļskis, I. Skadiņa, *Towards the development of the multilingual multimodal virtual agent.* 9th International Conference on NLP PolTAL 2014. Springer, (2014), 470-477.