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

doi:10.3233/978-1-61499-442-8-227

Language Resources and Technology in Latvia (2010-2014)

Inguna SKADIŅA[a,b,1], Ilze AUZIŅA[b], Guntis BĀRZDIŅŠ[b], Raivis SKADIŅŠ[a] and Andrejs VASIĻJEVS[a]
[a] *Tilde*
[b] *Institute of Mathematics and Computer Science, University of Latvia*

Abstract. Although human language technologies have a long history in Latvia, the Latvian language still belongs to under-resourced languages, as there are many gaps in basic language technologies and tools. However, despite difficulties, some of these gaps for both, resources and tools, have been filled in the last five years. The main goal of this paper is to report on recent achievements in language resources and technologies (LRT) for Latvian and to describe the current situation.

Keywords. Natural language processing, the Latvian language, language resources and tools, corpora, semantic analysis, speech technologies, machine translation

Introduction

Research and development of language resources and tools is a never-ending process driven by technological advancements, user needs, and research interests. Despite the long history of language technology research in Latvia, both CLARIN [[1]] and META-NET [[2], [3]] studies have identified serious gaps in basic language resources and technologies (LRT) for the Latvian language.

To advance Latvian language technologies, research and development in Latvia has significantly intensified during the last five years. Results of several recent projects and activities fill some major gaps in the availability of basic LRT for Latvian.

In this overview paper we present the main achievements in LRT field from 2010. By this we continue the series of reports on research and development activities on the Latvian language technologies ([[4]], [[5]] and [[6]]).

1. Language Policy and Major Activities

The importance of the language technologies for the long-term survival of Latvian has been recognized in the *State Language Policy Guidelines for 2005-2014.* Research in language technologies has been supported by the State Research Programmes, EU Structural Funds Programmes, EU FP7 and CIP ICT-PSP Programmes, was promoted

[1] Corresponding Author: Inguna Skadiņa: Vienības gatve 75^{a}, Riga, Latvia; e-mail: inguna.skadina@tilde.lv.

by the cooperative *Language Shore*[2] initiative, and also received a number of small grants from the Latvian Science Council. Major IT companies and research institutions have established the IT Competence Centre[3] co-funded by the Structural Funds Programme to create innovative language technology solutions.

Latvian researchers initiated and participated in several large scale international R&D projects that helped to advance Latvian language technologies and resources: ACCURAT[4], TTC[5] and TaaS[6] projects in the FP7 Programme, META-NORD[7], LetsMT![8] and EASTIN-CL[9] projects in the CIP ICT-PSP Programme.

Although Latvia is the only Baltic state which still has no dedicated Language technology programme, specific R&D actions are targeted in the *Information Society Development Guidelines 2014-2020* to advance support and usage of the Latvian language in the digital environment.

2. Infrastructural Developments

Latvia actively participated in the FP7 project CLARIN[10] to establish a common language resources and technology research infrastructure for the humanities. Unfortunately, political decisions have delayed taking the next step to establish a CLARIN ERIC centre in Latvia.

Latvia coordinates the Nordic and Baltic branch of the META-SHARE [[7]] infrastructure for language resources targeted at wider research and application development needs. Metadata for 100 Latvian language resources and tools have been catalogued in the META-SHARE, including 60 lexical resources of various types, 12 corpora and 3 language tools.

As part of the FP7 project MLi[11], Latvia is taking part in analysing the broader needs and requirements of an envisioned European public infrastructure of automated translation services, language resources and technologies.

3. Language Resources

3.1. Text Corpora

A great deal of work has been done in accumulating language resources. During the last five years special attention has been focused on creating corpora – both monolingual and multilingual, with and without annotation.

Work on the *Balanced Corpus of Modern Latvian* [[8]] has been continued and the size of the corpus has reached 4,5 million. Important step towards the preservation of the

[2] http://valodukrasts.lv
[3] http://www.itkc.lv/
[4] http://www.accurat-project.eu/
[5] http://www.ttc-project.eu/
[6] http://taas-project.eu/
[7] http://www.meta-nord.eu/
[8] http://project.letsmt.eu/
[9] http://www.eastin-cl.eu/project
[10] http://clarin.lv/
[11] http://mli-project.eu/

Latgalian language have been taken by creating the first balanced *Latgalian language corpus*[12] (MuLa). The corpus contains three types of texts: literal, technical and informative, selected by chronological principle for the time period 1988- 2012. The size of the corpus is 1 million running words.

Another important resource is the *Lithuanian-Latvian-Lithuanian parallel corpus* (LiLa[13]) [[9]] that represents texts in both languages from 1990. The corpus is bidirectional – it contains Latvian texts and their translations into Lithuanian and vice versa. The corpus includes fiction, periodicals, documents, etc. The size of the corpus is 8 million running words.

Several corpora have been made available on the META-SHARE platform, e.g. ACCURAT multilingual corpora of comparable texts, corpora of comparable Wikipedia sentences, Latvian-English Ngram corpus of legislative texts (1,3 million Latvian tokens). Processing of the EU Bookshop data resulted in parallel Latvian and other EU language corpora with 14,9 million Latvian tokens [[10]].

3.2. Treebanks

Latvian Treebank [[11]] is under development since 2010 and currently it contains ~3,700 sentences. The treebank is annotated according to the SemTi-Kamols dependency-based grammar model [[12]]. In essence, each tree is a dependency structure where some nodes are phrases instead of single words. This Treebank has recently been used to create a state-of-art syntactic dependency parser for Latvian [[13]] and various Treebank transformations have been explored.

3.3. Speech Corpus

An important achievement is the first Latvian speech corpus (*The Latvian Speech Recognition Corpus*) [[14]]. The corpus has been designed specifically for speech recognition purposes. It consists of two parts: an orthographically annotated corpus containing 100 hours of orthographically transcribed audio data and a phonetically annotated corpus containing 4 hours of phonetically transcribed audio data.

The corpus is designed to satisfy a set of criteria (audio signal quality, distribution of noise, speech styles, physical and linguistic characteristics of speakers, phonetic coverage etc.) which specify the required quality of speech data and the proportional distribution of data with different speaker characteristics[14].

3.4. Other Resources

The first resources for semantic analysis of the Latvian language have been developed – a Latvian FrameNet of 26 of the most popular frames has been created and a corpus of 5000 sentences has been fully annotated. Based on these resources a state-of-the-art Frame-semantic parser [[15]] for Latvian has been implemented.

For development and assessment of the grammar checker the *Error annotated corpus of Latvian* has been created [[16]]. It consists of two parts: the corpus of student papers and the balanced text corpus. The size of the corpus is 20,877 sentences.

[12] http://hipilatlit.ru.lv/bonito2/

[13] http://www.korpuss.lv/lila/

[14] http://runa.korpuss.lv

The development of the database of the valence of Latvian verbs has been started [[17]]. The database currently contains 300 verbs. The set of the semantic roles has been developed taking into account several approaches to case grammar and verb valency. Semantic roles are chosen and classified according to the language corpora used in annotation. The corpus contains 47,000 manually annotated sentences. The description of each semantic role includes the information about its grammatical form [[18]].

The work on the *Explanatory Dictionary*[15] is being continued. Currently the dictionary contains more than 232,000 entries from about 225 Latvian dictionaries of different times and domains. A new *Dictionary of the Modern Latvian* (44,760 entries)[16], the drafting of which has been completed this year, is integrated into *the Explanatory Dictionary* as well.

4. Tools and Technologies

Significant progress has also been made in language technologies. Tools for all language processing levels (morphology, syntax and semantics) have been developed in this period. Several important tools that have been developed for Latvian for the first time are named entity recognizers [[19], [20]], anaphora resolution [[21]], and a transliteration tool.

4.1. Morphological Analysers and Taggers

Morphological analysis as a basic technology for the Latvian language has been developed a long time ago. Some recent elaborations include the transformation of the previously developed morphological analyser into a finite state transducer making morphological analysis much faster [[22]].

One of the gaps in basic technologies for Latvian was the lack of a morphological tagger, which was not available till 2010. However now there are three different Latvian taggers available based on perceptron[17] and other machine learning algorithms [[23], [24]].

4.2. Syntactic Analysis

For syntactic analysis dependency parsers based on the *MaltParser* toolset and re-implementation of the Collins parser have been implemented resulting in several Latvian language parsers. To further improve parsing accuracy, research in dependency parsing has shifted to finding the best formalization [[13]] for complex language constructs (permissible word omissions, complex coordination, idiomatic constructs).

Another parser, based on context-free grammar has been extended for application in a grammar checking task [[25]]. The proposed grammar formalism is rather general and easily adaptable to similar languages, e.g. Lithuanian.

[15] http://www.tezaurs.lv/sv

[16] http://www.tezaurs.lv/mlvv/

[17] https://github.com/pdonald/latvian#part-of-speech-tagging

4.3. Semantic Analysis

A frame-semantic parser [[15]] along with Named Entity Linking and coreference resolution [[21]] has been used to create the first large-scale information extraction system for Latvian. This system was commissioned by the national news agency LETA to develop innovative tools to better utilize the Latvian newswire article archive.

4.4. Natural Language Generation

The first Controlled Natural Language resource for Latvian has been created. Latvian resource grammar now is a part of the *Grammatical Framework* distribution[18], allowing easy multilingual text generation in over 20 different languages, including Latvian [[26]].

4.5. Terminology Processing

In recent years the work in terminology has expanded from online databases such as EuroTermBank to cloud-based terminology services [[27]]. The Latvian language is among the best supported languages in the TaaS project where tools and workflows for terminology work have been developed. The TaaS platform[19] provides tools and services for monolingual terminology extraction, terminology mark-up using the new ITS 2.0 standard by W3C, bilingual term mapping [[28]], generation of canonical term form, terminology lookup and other term processing tasks. In particular, these methods are able to deal with multiword and compound terms taking into account their inflectional variations.

4.6. Machine Translation

During the last five years intensive work and significant progress has been achieved in machine translation (MT) for Latvian and other under-resourced languages. Active research was carried out in the collection and application of comparable data for statistical machine translation. Methods and tools created in the ACCURAT and TTC projects have helped to overcome the lack of parallel data for Latvian and other under-resourced languages [[29]].

Analysis of English-Latvian MT output reveals that for current state of art techniques the main difficulties are inflectional forms and word order [[30]]. Taking into account the morphological richness of Latvian and its complex syntax, special attention is paid to the methods for improving statistical MT by adding linguistic knowledge [[31]].

Besides work on general purpose MT solutions, MT adaptation for particular domains has been investigated [[32]]. Finally, the possible application of MT in the professional localization workflow has been evaluated showing productivity increase with no substantial degradation of quality [[33]]. As a result of this work machine translation systems for English-Latvian-English have been created showing higher quality than *Google Translate* in both automatic (BLEU score) and human evaluation [[34]].

[18] http://www.grammaticalframework.org/
[19] https://demo.taas-project.eu/

4.7. Speech Technologies

One of the research areas which has received less attention in previous years is speech technologies. Due to the lack of a speech corpus, only text-to-speech systems were developed in the first decade. In 2013 the newly created speech corpus opened opportunities to start new research activities in speech recognition.

The possibilities of both limited vocabulary and free speech recognition are currently under research [[35]]. Research on limited vocabulary ASR has led to the first experimental mobile apps prototyping speech based interface with Latvian text-to-speech and speech recognition[20]. Another active research direction is keyword recognition [[36]].

4.8. Multimedia

Novel research has been started to apply language technologies to dialog systems for smartphones. Two virtual agents are currently under development in Latvia. An intelligent virtual agent *Laura* [[37]] can hold conversations in English and provide information about several topics. She already can translate words, phrases and sentences into several languages and development of support for Latvian is in progress. The first Latvian speaking agent *Ēriks* [[38]] was designed for a currency converting task which was very actual in Latvia when the local currency was replaced by the euro. Limiting the lexicon of this virtual agent enabled the achievement of an accuracy of automatic speech recognition that is sufficient for practical application.

5. Conclusion

During the last five years several important basic language resources and tools have been developed narrowing the so called technological gap. In several areas novel methods have been researched to efficiently produce language resources and deal with the language complexities that can also be used for other under-resourced languages. Latvian language technologies are successfully applied to create innovative applications and solutions. However, language technology research in Latvia is still fragmented and there is an urgent need for a dedicated language technology programme to fill the remaining gap.

6. Acknowledgments

The research leading to these results have received funding from the research project “Information and Communication Technology Competence Center” of EU Structural funds, contract nr. L-KC-11-0003 signed between ICT Competence Centre (www.itkc.lv) and Investment and Development Agency of Latvia.

We would also like thank our colleagues, especially Andrejs Spektors, Indra Sāmīte, Baiba Saulīte and Roberts Rozis, for their support and contributions.

[20] http://www.tilde.com/chatterbots

References

[1] Krauwer, S. (2008). CLARIN: Common Language Resources and Technology Infrastructure, In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA).

[2] Skadiņa, I., Veisbergs, A., Vasiļjevs, A., Gornostaja, T., Keiša, I., Rudzīte, A. (2012). Language Technology Support for Latvian. In *The Latvian Language in the Digital Age* (pp. 60-79). Springer Berlin Heidelberg.

[3] Vasiljevs, A., & Skadina, I. (2012). Latvian Language Resources and Tools: Assessment, Description and Sharing. In *Proceedings of the 5th Conference on Human Language Technologies – the Baltic Perspective (BalticHLT)*. IOS Press, Frontiers in Artificial Intelligence and Applications 247, 265-272.

[4] Language & Technology in Europe 2000, Reports of Seminar, Riga, November 10-11, 1994.

[5] Milčonoka, E.,Grūzītis, N., Spektors, A. (2004). Natural Language Processing at the Institute of Mathematics and Computer Science: 10 Years Later, In *Proceedings of the first Baltic conference Human Language Technologies – the Baltic Perspective*, 6–12.

[6] Skadiņa, I., Auziņa I., Grūzītis N., Levāne-Petrova K., Nešpore G., Skadiņš R., Vasiļjevs A. (2010). Language Resources and Technology for the Humanities in Latvia (2004–2010). In *Frontiers in Artificial Intelligence and Applications*, 219, 15-22, IOS Press,

[7] Skadiņa I., Vasiļjevs A., Borin L., Lindén K., Losnegaard G., Olsen S., Pedersen B., Rozis R., de Smedt K. (2013). Baltic and Nordic Parts of the European Linguistic Infrastructure, In *Proceedings of Nodalida 2013*, 195-211.

[8] Levāne-Petrova K. (2012). Līdzsvarots mūsdienu latviešu valodas tekstu korpuss un tā tekstu atlases kritēriji. In *Baltistica VIII priedas*, Vilnius, 89-98.

[9] Utka, A., Levane-Petrova, K., Bielinskiene, A., Kovalevskaite,J., Rimkute, E. and Vevere, D. (2012). Lithuanian-Latvian-Lithuanian Parallel Corpus, In *Frontiers in Artificial Intelligence and Applications* 247, 260-264, IOS Press.

[10] Skadiņš, R., Tiedemann, J., Rozis, R., Deksne, D. (2014) Billions of Parallel Words for Free: Building and Using the EU Bookshop Corpus, In *Proceedings of LREC 2014*, Reykjavik, Iceland.

[11] Pretkalnina, L., & Rituma, L. (2012). Syntactic Issues Identified Developing the Latvian Treebank. *Proceedings of the 5th Conference on Human Language Technologies – the Baltic Perspective* . IOS Press, Frontiers in Artificial Intelligence and Applications 247, 185-192.

[12] Bārzdiņš, G., Grūzītis, N., Nešpore, G., Saulīte, B. (2007). Dependency-Based Hybrid Model of Syntactic Analysis for the Languages with a Rather Free Word Order. In *Proceedings of the 16th Nordic Conference of Computational Linguistics* 2007, 13-20.

[13] Pretkalnina, L., Znotins, A., Rituma, L., Gosko, D. (2014). Dependency parsing representation effects on the accuracy of semantic applications — an ex-ample of an inflective language. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC'14)*.

[14] Pinnis, M., Auziņa, I., & Goba, K. (2014). Designing the Latvian Speech Recognition Corpus. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC'14)*.

[15] Barzdins, G., Gosko, D., Rituma, L., Paikens, P. (2014). Using C5.0 and Exhaustive Search for Boosting Frame-Semantic Parsing Accuracy. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC'14)*.

[16] Deksne, D., Skadiņa I. (2014) Error-Annotated Corpus of Latvian. In *Human Language Technologies – The Baltic Perspective - Proceedings of the Sixth International Conference Baltic HLT 2014*. Kaunas, Lithuania: IOS Press.

[17] Nešpore, G., Saulīte, B., Grūzītis, N., Garkāje, G. (2012) Towards a Latvian Valency Lexicon. In *Proceedings of the 5th Conference on Human Language Technologies – the Baltic Perspective (BalticHLT)*. IOS Press, Frontiers in Artificial Intelligence and Applications 247, 154-161.

[18] Saulīte, B., Nešpore, G., Auziņa, I. (2014) Latviešu valodas verbu valences marķēšanās izmantojamās semantiskās lomas. In *Valoda: nozīme un forma 2. Kategoriju robežas gramatikā: LU Humanitāro zinātņu fakultātes Latviešu un vispārīgās valodniecības rakstu krājums*; red. A. Kalnača un I. Lokmane. Rīga : LU Akadēmiskais apgāds.

[19] Paikens, P., Auziņa, I., Garkāje, G., Paegle, M. (2012). Towards named entity annotation of Latvian National Library corpus. In *Human Language Technologies - The Baltic Perspective: Proceedings of the Fifth International Conference Baltic HLT 2012 (2012)*, 169-175.

[20] Pinnis, M. Latvian and Lithuanian Named Entity Recognition with TildeNER. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, 2012.

[21] Znotins, A., Paikens, P. Coreference Resolution for Latvian. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC'14)* (2014).

[22] Deksne D. (2013). Finite State Morphology Tool for Latvian. In *Proceedings of the 11th International Conference on Finite State Methods and Natural Language Processing*, 49-53*,* Association for Computational Linguistics.

[23] Pinnis, M., Goba, K. (2011). Maximum Entropy Model for Disambiguation of Rich Morphological Tags. In *Systems and Frameworks for Computational Morphology - Second International Workshop, SFCM 2011*, 14-22.

[24] Paikens, P., Rituma, L., Pretkalniņa, L. (2013). Morphological analysis with limited resources: Latvian example. In *Proceedings of NODALIDA 2013*. 267-278.

[25] Deksne, D., Skadiņa, I., & Skadiņš, R. (2014) Extended CFG Formalism for Grammar Checker and Parser Development. *Computational Linguistics and Intelligent Text Processing.* Springer, 237-249.

[26] Grūzītis N., Barzdins, G. (2010). Polysemy in Controlled Natural Language Texts. *Lecture Notes in Computer Science 5972*, 102-120.

[27] Vasiļjevs, A., Pinnis, M., Gornostay, T. (2014). Service model for semi-automatic generation of multilingual terminology resources. In *Proceedings of the 11th International Conference on Terminology and Knowledge Engineering 2014*, 67-76.

[28] Pinnis, M. (2013). Context Independent Term Mapper for European Languages. In *Recent Advances in Natural Language Processing (2013),* Hissar, Bulgaria.

[29] Skadiņa, I., (2012). Analysis and Evaluation of Comparable Corpora for Under-Resourced Areas of Machine Translation. In *Proceedings of the 5th Workshop on Building and Using Comparable Corpora (BUCC 2012)*, Istanbul, Turkey, 26 May 2012, (pp. 17-19).

[30] Skadiņa I., K. Levāne-Petrova, G.Rābante. (2012). Linguistically Motivated Evaluation of English-Latvian Statistical Machine Translation. In *Human Language Technologies – The Baltic Perspective - Proceedings of the Fifth International Conference Baltic HLT 2012*, IOS Press, Frontiers in Artificial Intelligence and Applications, Vol. 247, pp. 221-229.

[31] Skadiņš, R., Goba, K., Šics, V. (2011). Improving SMT with morphology knowledge for Baltic languages. In *Research Workshop of the Israel Science Foundation*, University of Haifa, Israel.

[32] Pinnis, M., Skadiņš, R. (2012). MT Adaptation for Under-Resourced Domains – What Works and What Not. In *Human Language Technologies – The Baltic Perspective - Proceedings of the Fifth International Conference Baltic HLT 2012*, 176-184, Tartu, Estonia.

[33] Skadiņš, R., Skadiņa, I., Sics, V., Pinnis, M., Vasiljevs, A., Hudik, T. (2014). Application of Machine Translation in Localization into low-resourced language. In *Proceedings of EAMT 2014 Conference*, Dubrovnik, Croatia

[34] Skadiņš, R., Šics, V., Rozis, R. (2014). Building the world's best general-domain MT for Baltic languages. In *Human Language Technologies – The Baltic Perspective - Proceedings of the Sixth International Conference Baltic HLT 2014*. Kaunas, Lithuania: IOS Press.

[35] Salimbajevs, A., Pinnis, M. (2014). Towards Large Vocabulary Automatic Speech Recognition for Latvian. In *Human Language Technologies – The Baltic Perspective - Proceedings of the Sixth International Conference Baltic HLT 2014*. Kaunas, Lithuania: IOS Press.

[36] Znotins, A., Dargis, R. (2014) Baseline for keyword spotting in Latvian broadcast speech. In *Human Language Technologies – The Baltic Perspective - Proceedings of the Sixth International Conference Baltic HLT 2014*. Kaunas, Lithuania: IOS Press.

[37] Vīra, I., Teseļskis, J., Skadiņa I. (2014). Towards the development of the multilingual multimodal virtual agent. In *9th International Conference on Natural Language Processing PolTal 2014*, 470-477, LNAI 8686, Springer.

[38] Vīra, I., Vasiļjevs A. (2014). The Development of Conversational Agent Based Interface. In *Human Language Technologies – The Baltic Perspective - Proceedings of the Sixth International Conference Baltic HLT 2014*. Kaunas, Lithuania: IOS Press.