NLP Infrastructure for the Lithuanian Language

Daiva Vitkutė-Adžgauskienė, Andrius Utka, Darius Amilevičius, Tomas Krilavičius
Vytautas Magnus University,
K. Donelaičio g. 58, 44248 Kaunas, Lithuania
E-mail: {d.vitkute-adzgauskiene, d.amilevicius, t.krilavicius}@if.vdu.lt, a.utka@hmf.vdu.lt

Abstract

The Information System for Syntactic and Semantic Analysis of the Lithuanian language (lith. Lietuvių kalbos sintaksinės ir semantinės analizės informacinė sistema, LKSSAIS) is the first infrastructure for the Lithuanian language combining Lithuanian language tools and resources for diverse linguistic research and applications tasks. It provides access to the basic as well as advanced natural language processing tools and resources, including tools for corpus creation and management, text preprocessing and annotation, ontology building, named entity recognition, morphosyntactic and semantic analysis, sentiment analysis, etc. It is an important platform for researchers and developers in the field of natural language technology.

Keywords: language technology infrastructure, Lithuanian language

1. Introduction

The paper presents LKSSAIS[1], the first infrastructure that provides online access to open-source tools for managing, processing, annotating, and analyzing Lithuanian language texts. The system was designed according to the best practices in the development of language technology infrastructures, keeping in mind the possibility of integration with the European CLARIN ERIC Infrastructure[2]. It attempts to address the needs of researchers working with the Lithuanian language, taking into account continuously growing amount of data and growing need for natural language processing applications in scientific, public, and business fields. It also provides online free access to public services for syntactic and semantic analysis of Lithuanian language texts.

Language tools and resources today are of paramount importance for research and teaching, as well as for different industrial applications. Two main factors leading to the development of Lithuanian language technology infrastructure were: 1) Lithuanian language has been considered to be extremely under-resourced on the NLP side and had a very weak support according to META-NET survey (Vaišnienė and Zabarskaitė 2012); 2) Lithuanian language is very rich in inflections and diacritics, therefore reuse or simple adaptation of widely-used English language technology tools and specifications was unreasonable.

Although prior to 2013, there had been a number of activities for developing language tools, resources and services for the Lithuanian language (Marcinkevičienė and Vitkutė-Adžgauskienė 2010), these activities were not well coordinated, separate tools and corpora were created using different standards, without well documented APIs and without open-access interfaces. Unsurprisingly the quality of most of NLP tools for Lithuanian was far away from the state of the art.

In order to improve such situation, a project for the development of the Information System for Syntactic and Semantic Analysis of the Lithuanian language was launched in the middle of 2012. The project was the collaboration of the Vytautas Magnus University (coordinator) and Kaunas Technology University, and its first results were presented in May, 2015. The main result of the project is an infrastructure combining Lithuanian language tools and resources for diverse language research and applications tasks, thus representing the first national NLP infrastructure for the Lithuanian language. It should be noted that most tools and services for the Lithuanian language were implemented for the first time. The fact that such infrastructure was built in just two years can come as a surprise. For this reason we will present some additional information about the context of the project. In the middle of 2012 Lithuania launched the program *Lithuanian language for information society*. The program consisted of 5 projects, which covered a vast range of language and speech technologies, and was funded by the European Regional Development Fund.

There were three main factors for the success of such a big project in such a short time: 1) sufficient amount of the funding, 2) state-of-the-art project management, and 3) effective collaboration between top-level academic researchers and top-level developers from the private sector.

The project implementation consisted of two stages: 1) development of the prototypes of different NLP components; 2) final implementation of the information system. The web information retrieval subsystem and the corpus repository were developed at the beginning of the project. It was important to do so, in order to ensure enough material for testing NLP components and pipelines of the system, while continuously crawling (scanning) the Lithuanian web. Whereas LKSSAIS is a modular system, during its implementation the leverage technology had to be used to ensure continuous development of sequential NLP components.

Due to the fact that in Lithuania the NLP expertize is concentrated in universities and the IT development in private companies, some top-level developers from

[1] http://semantika.lt
[2] http://clarin.eu

private sector enterprises had to be involved. The total number of staff employed by two universities for the project was around 75 persons (including researchers and developers).

At the end of the program, in the middle of 2015, the main universities that were involved in implementation of the program (Vytautas Magnus University, Vilnius University and Kaunas University of Technology), formed the National consortium (CLARIN-LT), which became an official member of the European CLARIN ERIC infrastructure. It is hoped that the national infrastructure will greatly benefit from the membership to the international research infrastructure in terms of acquiring resources and competences from the competent CLARIN community.

2. Comparison with other Infrastructures

As previously mentioned, the LKSSAIS infrastructure for the Lithuanian language was designed according to the best international practices in the development of language technology infrastructures such as, CLARIN-DK[3], LINDAT[4], CLARIN-D[5] and META-SHARE[6].

While a detailed comparison of all the above mentioned infrastructures would be desirable, it would fall outside the scope of the paper. Thus we discuss a conceptually similar infrastructure, namely German CLARIN-D infrastructure.

When comparing both the feature list and the technical implementation model of the LKSSAIS infrastructure to the German CLARIN-D, it can be stated that there is a big correspondence in both cases. Regarding the feature list, both infrastructures allow free access for researchers to main analysis and annotation tools for national languages, covering sentence splitters, tokenizers, part-of-speech taggers, morphological analyzers and lemmatizers, syntactic analyzers, named entity recognizers, etc., as well as access to national linguistic resources, such as national language corpora or WordNet-type linguistic ontologies.

In both cases, service-oriented architecture of the system is used, implementing language technology tools such as HTTP/REST web services, and thus allowing the integration of corresponding tools into different applications. Besides, in both cases, a web interface is offered for easier access to corresponding tools, and, moreover, for using tool chains for solving specific text analysis tasks. However, it must be noted, that, while CLARIN-D already has a mature linguistic chaining tool WebLicht[7] (Hinrichs, Hinrichs and Zastrow 2010) with the web interface access, the web interface and tool chaining service for LKSSAIS is less flexible and still in its development phase, lacking convenient configuration management and visualization options.

3. Building Blocks of LKSSAIS

LKSSAIS is a modular, flexible, integrated, secure, open and interoperable information system. The backend of the system is the infrastructure of natural language processing and semantic tools, supported by necessary language resources, while the frontend is a set of online public services for information search and text analysis. The main focus of this paper is on the backend part of the system. It consists of the three main subsystems (see Figure 1):

1. **WEB Information Retrieval Subsystem** is used for building the web-based Lithuanian Corpus. It employs Apache Nutch[8] in line with Lithuanian language identification component for filtering out non-Lithuanian documents, supported by XPath (W3C 2015) for text and metadata extraction and boilerplate removal.
2. **Corpora Storage and Management Subsystem**. Corpora files and stand-off annotations are stored in the document-oriented database (MongoDB[9]). Presently there are two corpora in the system: 1) a dynamically growing web-based Lithuanian corpus (~1 billion tokens) and 2) the Corpus of Contemporary Lithuanian language (~200 million tokens). The database is paired with a repository management component, which ensures effective corpus data and metadata management. Texts are in a plain text format and remain intact after NLP processing. Each component produces its stand-off annotation in JSON (ECMA International 2013), which can be stored in the corpus storage and management subsystem or can be directly used by other LKSSAIS software components or external software. JSON was chosen over XML (W3C 2008) due to the fact that it ensures much better throughput and effective communication between different processing modules. If required, it is possible to export annotations and metadata into the TEI P5[10] standard file.

The *Repository Management Component* includes a classification module, which is used for populating domain-specific corpora. Presently two sub-corpora of politics and business are being compiled by the component. Indexing of corpus texts (full text indexing, inverted index, annotations indexing etc.) is performed by Lucene[11] based multiple-index engine.

2.1. **Basic NLP Pipeline** consists of text preprocessing (encoding converter (precision - 100% / recall - 100%), language filter 96/96), lexical processing (tokenizer 98/98, lemmatizer 95/95, spell checker 95/95, grammar checker 90/80), and morphosyntactic processing (POS tagger 95/95, morphological analyser 95/95, disambiguation tool, shallow syntactical parser (with precision 83%), deep syntactical parser 70/70 (Boizou and Zamblera 2014)) tools.

[3] https://clarin.dk
[4] https://lindat.mff.cuni.cz
[5] http://www.clarin-d.de/
[6] http://www.meta-share.eu/
[7] http://weblicht.sfs.uni-tuebingen.de

[8] http://nutch.apache.org/
[9] https://www.mongodb.org/
[10] http://www.tei-c.org/
[11] https://lucene.apache.org

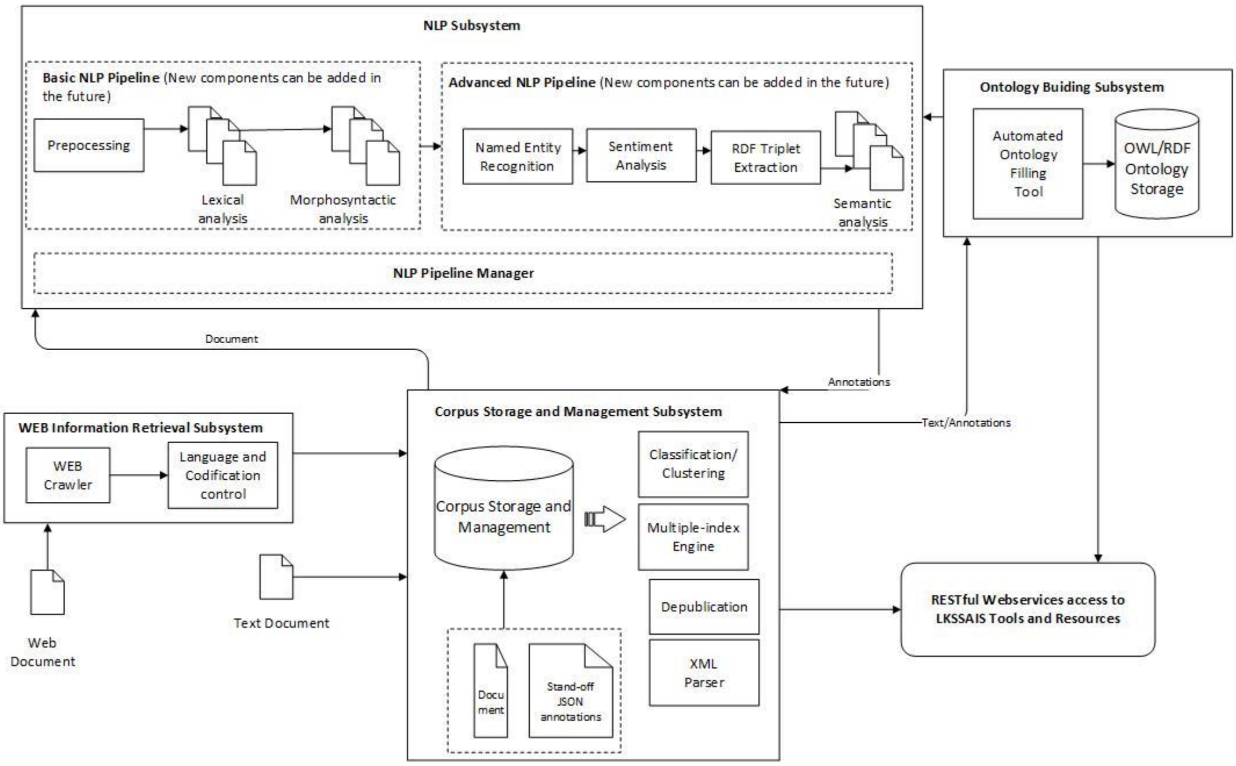
NLP Subsystem
Basic NLP Pipeline (New components can be added in the future)
Prepocessing
Lexical analysis
Morphosyntactic analysis
Advanced NLP Pipeline (New components can be added in the future)
Named Entity Recognition
Sentiment Analysis
RDF Triplet Extraction
Semantic analysis
NLP Pipeline Manager
Ontology Buiding Subsystem
Automated Ontology Filling Tool
OWL/RDF Ontology Storage
Document
Annotations
Text/Annotations
WEB Information Retrieval Subsystem
WEB Crawler
Language and Codification control
Web Document
Text Document
Corpus Storage and Management Subsystem
Corpus Storage and Management
Classification/ Clustering
Multiple-index Engine
Depublication
XML Parser
Docu ment
Stand-off JSON annotations
RESTful Webservices access to LKSSAIS Tools and Resources

Figure 1: Structure of the LKSSAIS infrastructure

Different preprocessing and processing tasks are supported by diverse linguistic resources, such as ontologies, grammars, lexicons, and sets of linguistic rules. The selection of tools in the basic pipeline can be changed according to analysis or research needs. The modular architecture permits adding new components to the pipeline, e.g. preprocessing and processing of social media texts may require different tools than those needed for the standard language.

2.2. **Advanced NLP Pipeline** is used for semantic analysis. It can perform the following text analysis tasks: 1) named entity recognition (main component NER recognizer, F-80%), identifying such entities as personal names, company names, geographical names, expressions of time, quantities, monetary values, percentages, etc.; 2) sentiment analysis (main component sentiment analyzer), identifying and extracting subjective information that describes attitudes of an author including his/her judgment (precision 80 % - positive / negative / neutral text) and emotional state in the text (precision 60 % - anger, fear, sadness, surprise, satisfaction, joy); 3) RDF triplet extraction (main component RDF triplet extractor) is through subject-predicate-object triplet identification in Lithuanian sentences. Each of these components can be used to extend any basic NLP pipeline, provided that prerequisite requirements for component application are met with respect to annotation levels.

2.3. **NLP Pipeline Manager** is a Java framework that manages the pipeline of NLP annotations and serves the requests from *Corpora Storage Manager* to annotate new texts. Its GUI allows an administrator to monitor the annotation process, to check the status of pipeline components, to register new components, to remove the old ones, to build new pipelines, and to make changes to existing ones. It also logs errors of annotation process with their descriptions. In case of multiple requests the message bus of the Pipeline Manager handles their logistics. Presently, the system does not support parallel annotation.

3. **Ontology Building Subsystem** is dedicated to ontology building from corpus texts, as well as other external sources (ontologies, dictionaries, linguistic databases, etc.). *Automated Ontology Filling Tool* allows populating the ontology, using named entity and RDF triplet information extracted and annotated in the Advanced NLP pipeline as well as using information extracted directly from corpus texts, using different statistical analysis, classification and rule-based methods. Corpus texts are employed for extracting and comparing contexts of candidate words for the ontology in the terms of their neighboring word collections, as well as for extracting hierarchical dependency information for ontology concepts. Ontology Building Subsystem also accommodates several special-purpose ontologies: *Common Ontology of Lithuanian Language* (Lithuanian WordNet (Vitkutė-Adžgauskienė, et al. 2016)), *Ontology of Geo-names* (10,000 instances), *Ontology of Personal Names* (40,000 instances), and *Ontology of Abbreviations* (600 instances).

4. Users

The users of LKSSAIS system can be divided into the 3 major categories: non-experts, experts, and IT systems. While online services of the infrastructure is of interest to the wide non-expert group for casual multi-level text analysis, the back end of the system is very important for expert users, namely developers, researchers, and IT engineers. A multi-component architecture of LKSSAIS that communicates via a simple REST web interface meets the growing interest in cloud-based NLP services for enterprise and public IT systems.

5. Problems

The design and exploitation of LKSSAIS has revealed two major problems:

1. Scalability and pipeline management issues. The initial challenge is to support a theoretically unlimited number of tokens with a theoretically unlimited number of annotation layers built upon those tokens. The similar challenge has already been tackled for German corpora in IDS in Manheim (Kupietz, et al. 2014). However, since there was no preceding experience in building such an NLP framework for Lithuanian, the exploitation of non-distributed LKSSAIS in real-life conditions has revealed that quality of service can be provided only on up to 3 billion word corpora. Due to the rapidly growing amount of data the capacity limit of the system will be reached by the end of 2018. There are two possible solutions: 1) migrating the system into the servers with more resources (processing power, memory, storage), or 2) modifying the system into to a distributed version, using frameworks such as Hadoop and Spark with distributed storage solutions.

2. Licensing of open source tools, resources and corpora. All NLP tools and language resources in LKSSAIS are open source, but their free distribution is hindered by the Lithuanian legal framework. The Lithuanian legal framework does not accept widely recognized open source licensing practices (Creative Commons, GPL, LGPL etc.), and thus the distribution of tools and resources can only be provided by complicated procedures of legacy licensing. For this reason presently LKSSAIS provides only GUI and API based services for developers and CLARIN ERIC community.

However, it is expected that the Lithuanian legal framework will change in the near future. The prospective legislation will require that almost all tools and resources in state information systems will need to be freely distributed and open access. In case of LKSSAIS, exceptions Except the WEB corpus and Corpus of Contemporary Lithuanian language, because both contain texts covered by copyright limitations. It must be noted that these limitations do not prevent making open access and free of charge research of their content and produce derived resources (e.g. dictionaries etc.).

6. Discussion and Next Step

Large scale infrastructures such as LKSSAIS for Lithuanian are complex systems that in some cases reach limits of software and hardware capabilities and thus often require innovative and unique engineering solutions. While building LKSSAIS, two of the most important lessons learned are: (1) realizing the importance of a compact and efficient format (JSON) that is essential for the effectiveness of the system; (2) an importance of distributed technical solutions when processing Big Data. Of course, well-known in NLP/CL communities licensing issues should not be underestimated, as their proper usage ensures necessary flow of ideas and competence, while at the same time protects the rights of authors.

In spite of difficulties such national infrastructures and their integration into the European context are very important, as the accumulation of basic and complex language tools provides a backbone for NLP and the environment for further improvement.

As far as future steps for developing LKSSAIS are concerned, they will take three main directions: (1) technical solutions for a distributed system and their implementation into LKSSAIS; (2) building the annotated Lithuanian web-social-media and medical corpora and (3) Developing the pipeline for analysis of informal and medical texts.

7. Acknowledgements

LKSSAIS was funded by the European Regional Development Fund.

8. Bibliographical References

Boizou, L., and Francesco Z. (2014). Syntactic Engine for the Lithuanian Language. *Proceedings of the 6th Baltic HLT conference*, Kaunas: IOS Press, pp. 69-74.

ECMA International. 2013. Standard ECMA-404: The JSON Data Interchange Format.

Henriksen, L., Dorte H.H., Maegaard B., Pedersen B.S., and Povlsen C. (2014). Encopassing a spectrum of LT users in the CLARIN-DK Infrastructure. LREC 2014, pp. 2175-2181.

Hinrichs E. , Hinrichs M., and Zastrow T. (2010). WebLicht: web-based LRT services for German. *Proc. of the ACL 2010*, pp. 25-29.

Kupietz, M., Lüngen H., Bański P., and Belica C. (2014). Maximizing the Potential of Very Large Corpora: 50 Years of Big Language Data at IDS Mannheim. *LREC. Challenges in the Management of Very Large Corpora (CMLC-2)*, Reykjavik.

Marcinkevičienė, R., and Vitkutė-Adžgauskienė. D. (2010). Developing the Human Language Technology Infrastructure in Lithuania. *Proc. of the 4th Int. Conf. Baltic HLT*, Riga. Amsterdam: IOS Press, pp. 3-10.

Vaišnienė, D., and Zabarskaitė J. (2012). The Lithuanian Language in the Digital Age / Lietuvių kalba skaitmeniniame amžiuje. *META-NET White Paper Series, Europe's Languages in the Digital Age*, ed. G. Rehm and H. Uszkoreit. Heidelberg: Springer.

Vitkutė-Adžgauskienė D., Dainauskas J, Amilevičius D, and Utka A. (2016). Lietuvių kalbos žodžių tinklas (Lithuanian Word Net). *Darbai ir dienos* 64, pp. 101-114.

W3C. 2008. "Extensible Markup Language (XML) 1.0." W3C. Accessed March 1, 2013. http://www.w3.org/TR/2008/REC-xml-20081126/.

—. 2015. XML Path Language (XPath) 2.0. *W3C.* September 7. Accessed October 22, 2015. http://www.w3.org/TR/xpath20/.