VYTAUTO DIDŽIOJO
UNIVERSITETAS
Informatikos fakultetas

Centre of
Computational
Linguistics

Overview of the Development of Language Resources and Technologies in Lithuania (2012-2015)

ANDRIUS UTKA, DARIUS AMILEVIČIUS, TOMAS KRILAVIČIUS, DAIVA-VITKUTĖ ADŽGAUSKIENĖ

VYTAUTAS MAGNUS UNIVERSITY, KAUNAS, LITHUANIA

Plan of the presentation

1. Three stages of development of Language Technologies in Lithuania;

2. Policy of Language Technologies (projects and strategic documents);

3. Language technology research infrastructures;

4. Language resources and tools;

5. Future prospects.

Three Stages of Development

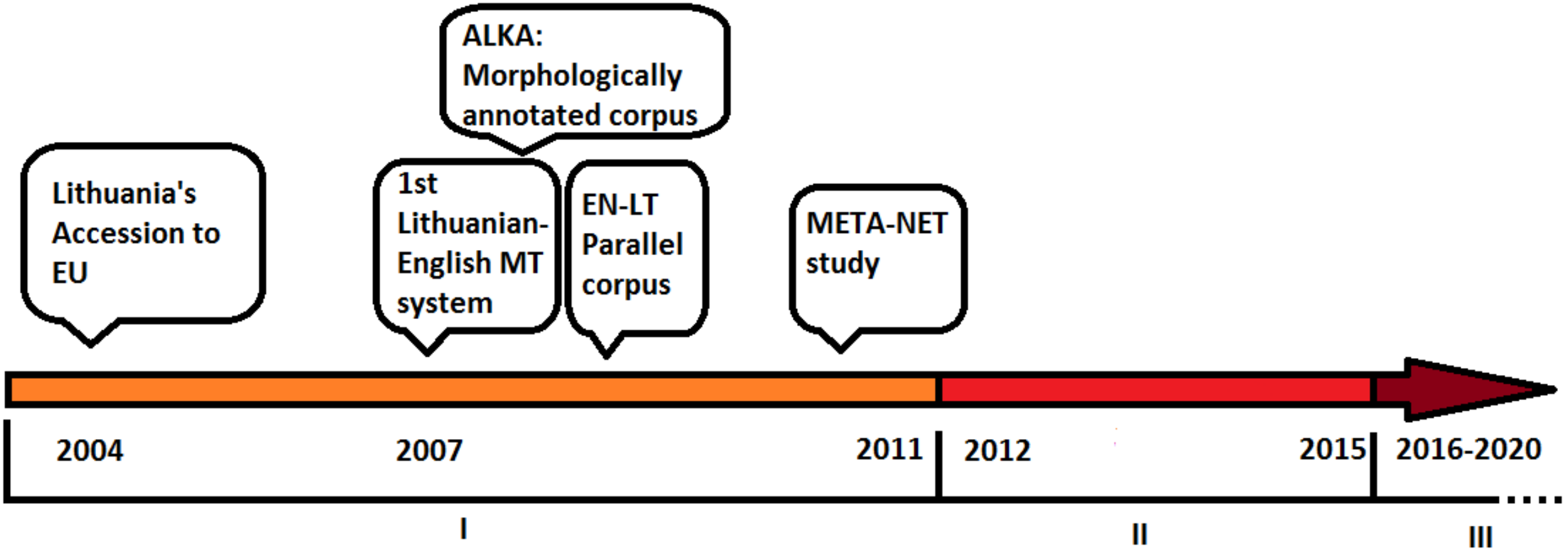
Judging from the current perspective and starting from Lithuania's joining the EU the development of language resources and technologies can be divided into 3 stages:

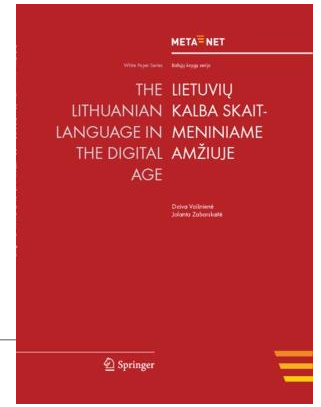
I. 2004-2011

II. 2012-2015

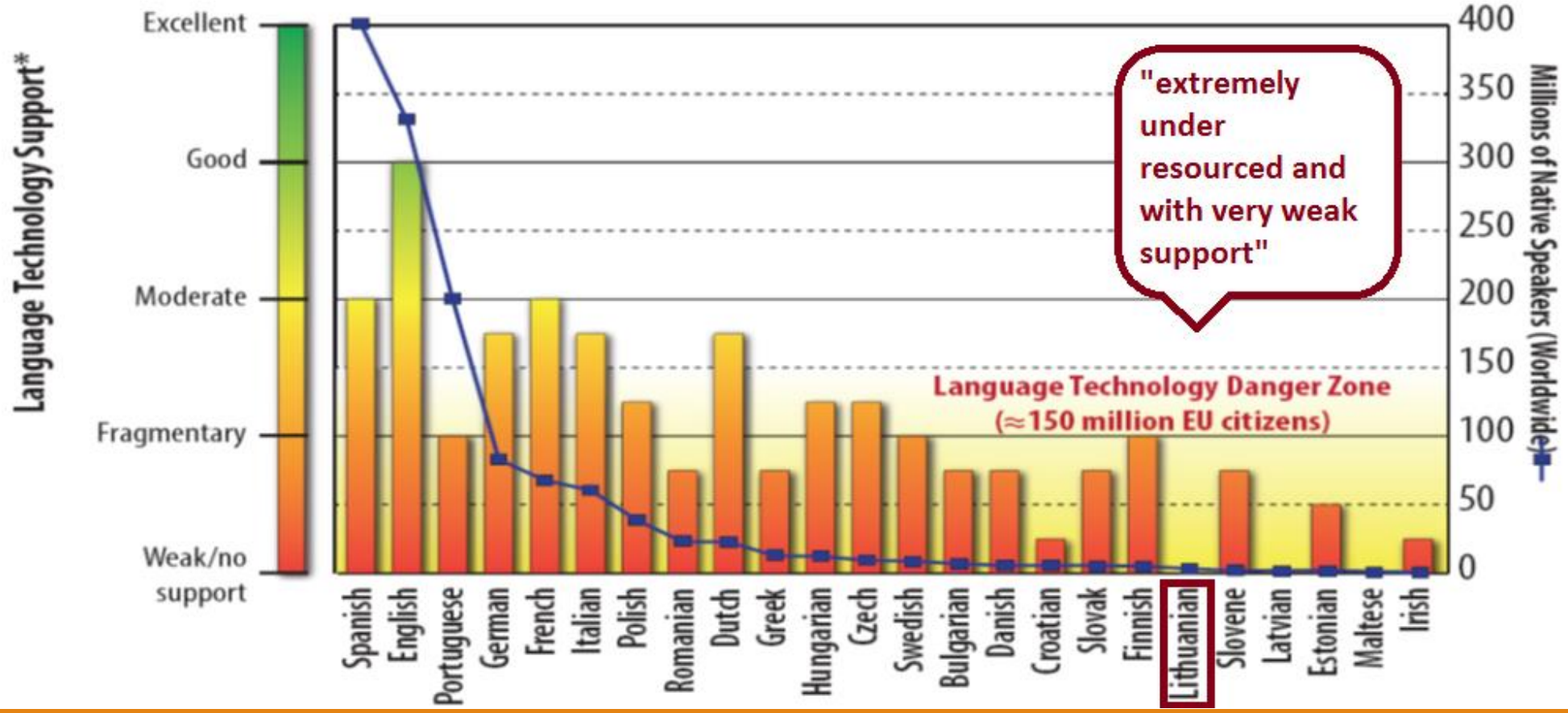
III. 2016-2020

Three Stages of Development (Timeline)

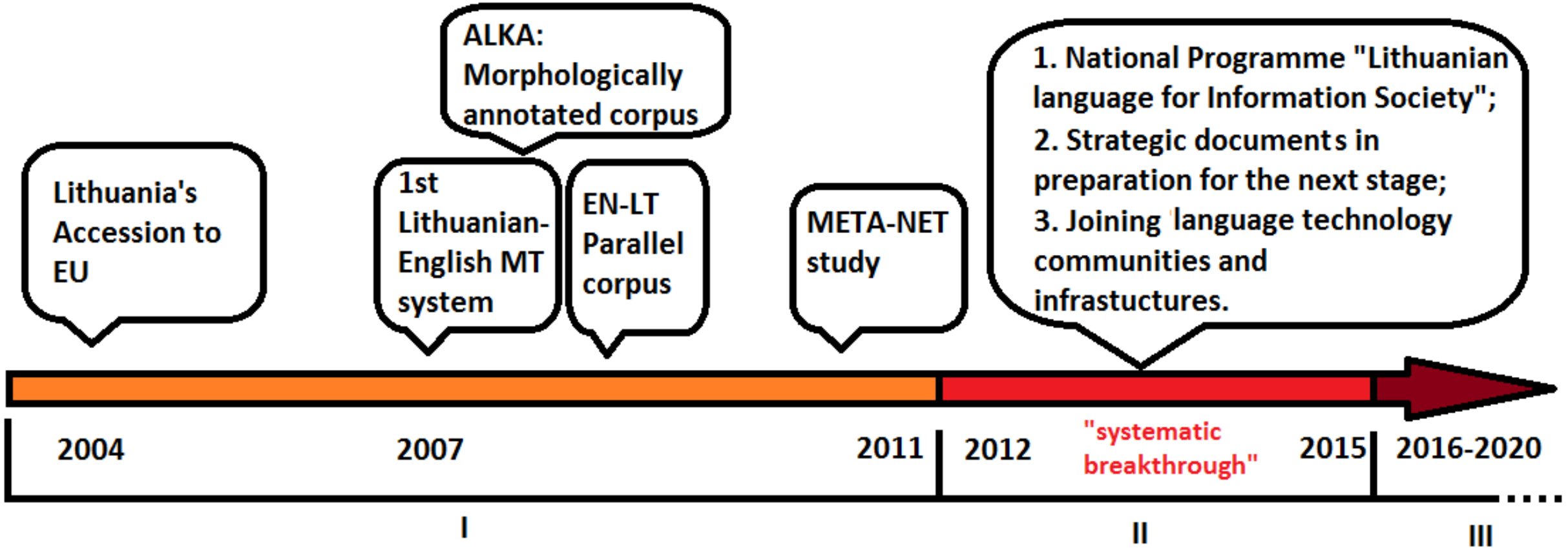




1 stage: META-NET Study (2010/2011)



Three Stages of Development (Timeline)



Policy of Language Technologies (projects)

- The national programme "The Lithuanian Language for Information Society (2012-2015)" funded 6 large language technology projects:

  1) Semantika LT (Language technology services for institutions and private users) (Vytautas Magnus University and Kaunas Technology University),

  2) LIEPA (Speech recognition and synthesis) (Vilnius University, LEU, Šiauliai University, Institute of Lithuanian language),

  3) Software localization and development of localization tools (Vilnius University),

  4) English-Lithuanian-English and French-Lithuanian-French statistical MT system (Vilnius University),

  5) Solutions and resources for preservation of Lithuanian language in public sphere (Institute of Lithuanian Language),

  6) the platform *www.raštija.lt* (Vilnius university).

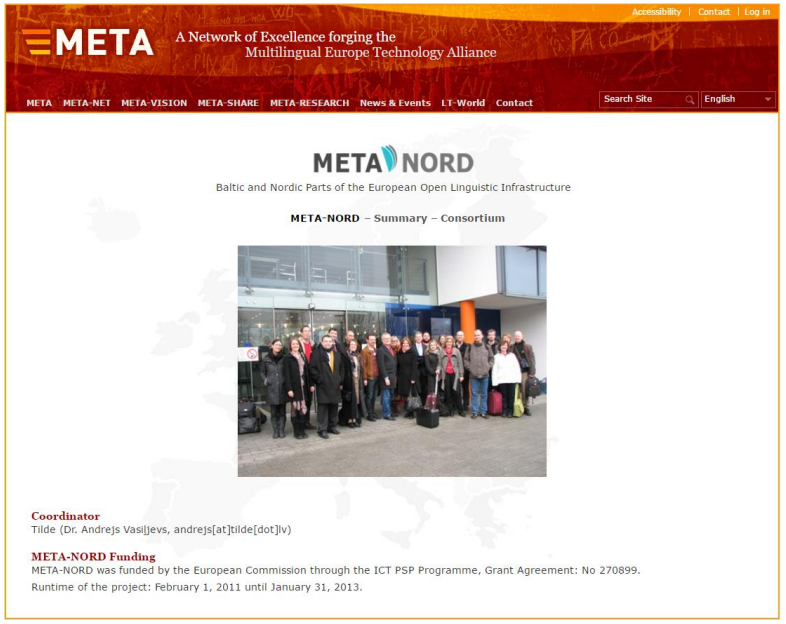
- A few smaller important research projects were funded by the Research Council of Lithuania, EU Cross-Border programme (e.g. ŠIMTAI – terminology extraction, ASTRA –authorship attribution, etc.)

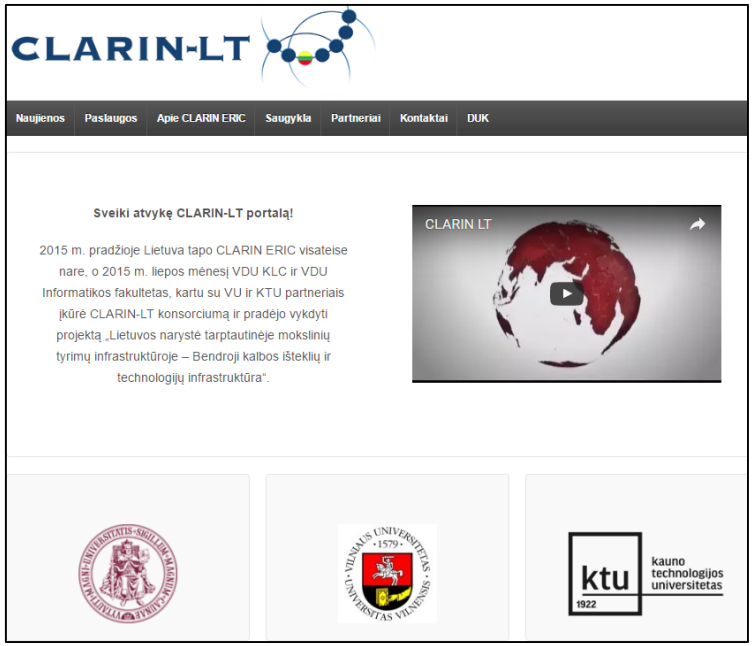
Policy of Language Technologies (strategic documents)

- 2013 State Commission of the Lithuanian language issued "Guidelines for Lithuanian Language Technologies development 2014-2020". Prioritised fields are: machine translation, speech analysis, dialogue systems, automatic summarization, semantic technologies, advanced text analysis, compilation of language resources, and others.
- 2014 “The Lithuanian Information Society Development Programme 2014–2020” (“Digital Agenda of Lithuania”) was approved by the Government of Lithuania, which is an important document for the next funding period.
- 2015 the Government renewed “Lithuanian Roadmap for Research Infrastructures", which introduced 22 initiatives of national significance. In the SSH field, the Roadmap lists 3 infrastructures: E-Lingua (CLARIN-LT), ESS LT, and LIDA.

Language Technology Research Infrastructures

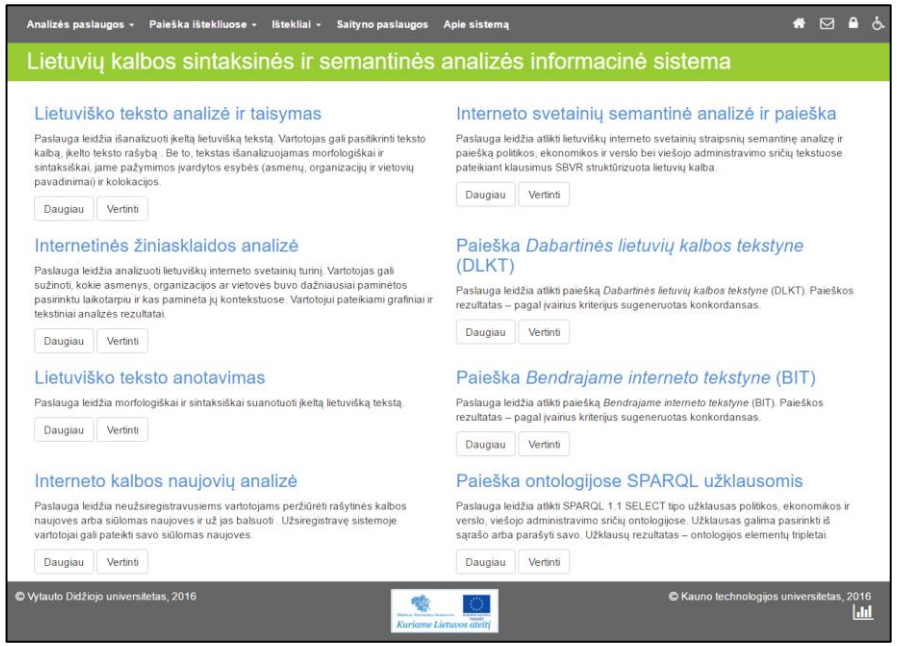
- **International Research Infrastructures**
  - CLARIN - Common Language Resources and Technology Infrastructure,
  - META-NET - Multilingual Europe Technology Alliance.
- **National Research Infrastructures**
  - LKSSAIS (semantika.lt) – Lithuanian language services for syntactic and semantic language analysis,
  - Raštija.lt - the access point for all projects of the National programme.





International Infrastructures

National Infrastructures



Main Language Resources (corpora)

Corpora

- **DLKT (Corpus of Contemporary Lithuanian Language)** (200 m words),
- **BIT (Lithuanian Internet News Texts corpus)** (~1 b words),
- **LILA (Lithuanian-Latvian-Lithuanian Corpus**) 10 m words (*tekstynas.vdu.lt*),
- **ASTRA corpora** designed for authorship attribution and author profiling experiments,
- **Multilingual corpora** for NER, classification, clustering and language identification tasks (LT, RU, AZ),
- Etc.

Main Language resources (treebank and speech corpora)

Treebank

ALKSNIS (a treebank) - 2,300 sentences encoded in the PML (Prague Mark-up Language).

Speech corpora

113 h of recorded speech (Vilnius University);

15 h of recorded speech (Vytautas Magnus university, Kaunas).

Ontology

LitWordNet 44,640 words and 54,446 word senses, organized into 43,903 synsets, and is compliant with main structural principles of Princeton WordNet.

Main Language Analysis Tools

A number of language tools have been created during the implementation of the National programme:

1. new morphological analyser based on Hunspell engine,

2. syntactic parser,

3. EN-LT-EN, FR-LT-FR – machine translation system

4. semantic tools (NER recogniser, sentiment analyzer, RDF triplet recognizer, SPARQL question answering tool);

5. new engine for speech command and speech synthesis;

6. the pipeline for Lithuanian text annotation;

Etc.

Future Prospects

1. The 2nd National Programme "The Lithuanian Language for Information Society (2014-2020)", which will be funded by EU Structural funds and the State budget, is about to start in 2017. It is planned that 5 large language technology projects will be funded. The projects will aim to implement new technologies in machine translation, speech recognition, automatic transcription, semantic technologies, automatic summarisation, social media texts analysis and others.

2. Although national CLARIN-LT project will finish in 2016, we hope that the funding will be prolonged until 2020, as CLARIN-LT is included in the New Roadmap of Lithuanian research infrastructures of national importance. However, the extent of the funding remains vague;

3. There are signs that businesses become increasingly interested in automatizing their processes with the help of language technologies. So we expect that there will be more work for implementing new practical applications.

Ačiū!

Paldies!

Tänan!