Developing the Human Language Technology Infrastructure in Lithuania

Prof. Rūta Marcinkevičienė,
Dr. Daiva Vitkutė-Adžgauskienė

Vytautas Magnus University
Centre of Computational Linguistics
Kaunas, 2010

VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Presentation plan

- Introduction: EU policy and language technologies
- HLT in Lithuania: short overview
- HLT classification schemes – a tool for analysis
- Strategic issues in building the HLT infrastructure in Lithuania
- Conclusions

Introduction: EU policy and language technologies

- EU emphasizes the preservation of **linguistic diversity** and the promotion of **multilingualism**
  - FP7 Social Sciences and Humanities Research Roadmap
    - Sustaining cultural diversity 2011-2013
      - Translating Europe – a topic for 2011
      - Unity in linguistic diversity: a challenge for Europe – a challenge beyond 2011
  - HERA (ERANET for Humanities in European Research Area)
    - A report on existing infrastructural facilities and practices 2006
    - the strategy for the development of future infrastructure initiatives

VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Introduction: EU policy and language technologies

European Science Foundation

- ESF-EUROHORCs Roadmap
- ESF Member Organisation Forum on Medium-Sized Research Infrastructures

ESFRI (European Strategy Forum on Research Infrastructures) Roadmap 2008,

- CESSDA (Council of European Social Science Data Archives)
- DARIAH - Digital Research Infrastructure for the Arts and Humanities
- CLARIN (Common Language Resources and Technology Infrastructure)

Introduction: EU policy and language technologies

- SSH research infrastructures as language resources supporting cultural diversity

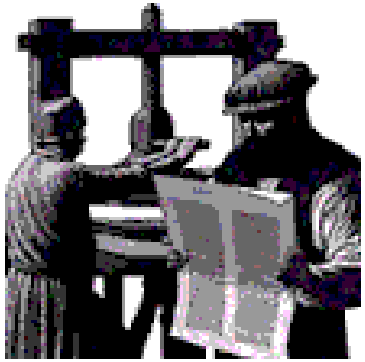
- Official **languages of EU differ in the level and availability of HLT resources**
  - Internationally used and thoroughly resourced languages are much more advanced in HLT tool availability
  - Scarce language resources for lesser used languages
  - However, latecomers benefit from the know-how and universally applicable technologies

Technologies and language survival (1)

- Wide-spread assumption - **barriers of survival for languages are related to technologies**:

  McLuhan (1994): "Physiologically, man in the normal use of technology is perpetually modified by it and in turn finds ever new ways of modifying his technology"

  - **Writing** - the first invention for knowledge dissemination (rather restricted, placed into the hands of church scholars)
  - **Printing** - Gutenberg's press put the power of knowledge dissemination into the hands of publishers (*first Gutenberg effect*)
  - **Computer mediated communication** - availability of self publishing put the power to produce and distribute knowledge and content in the hands of the author (*second Gutenberg effect*)



VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Technologies and language survival (2)

- Several oral languages died out unrecorded as a result of the first two inventions.

- The third one (computer mediated communication) might be crucial as it combines with socio-political and socio-economical factors:
  - domineering of global languages (96% of the world's languages being spoken by only 4% of the world's population)
  - high rate of migration
  - bilingualism
  - influence of mass media and mass culture (e.g. "netspeak" on Internet ☹)

David Crystal "The Language Revolution" (2004): "*Languages die when its speakers die due to natural disasters, genocide, and political persecution or* ***as they assimilate to the dominant culture***".

HLT in Lithuania: overview (1)

CORPORA OF WRITTEN LANGUAGE:

- **Corpus of Contemporary Lithuanian Language CCLL** (160 mln. running words) and its morphologically annotated version, as well as a set of parallel corpora (**bidirectional Czech-Lithuanian and Lithuanian-Czech corpus** of five millions words and **English-Lithuanian corpus** of 18 million words), compiled by the Centre of Computational Linguistics of Vytautas Magnus University (VMU);

- **Corpus Academicum Lithuanicum CorALit** (9 mln. running words), compiled by the Faculty of Philology of Vilnius University (VU);

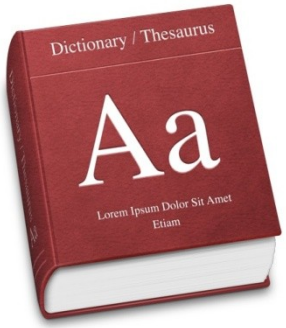
- **Corpus of Spoken Lithuanian language** (200 000 running words) and a universal annotated database of speech recordings, compiled by the Centre of Regional Studies of VMU;

- **Corpus of Lithuanian Dialects and Database of Old Lithuanian Writings**, compiled by the Lithuanian Language Institute (LLI).

HLT in Lithuania: overview (2)

DICTIONARIES AND ON-LINE DATABASES:

- Bilingual English-Lithuanian, French-Lithuanian and international-word dictionaries Alkonas, Anglonas 2, Frankonas and Interleksis, maintained by Fotonija;
- English-Lithuanian, Lithuanian-English, German-Lithuanian and Lithuanian-German dictionaries LED and WinLED, maintained by the TEV Publishing House;
- Dictionary of Contemporary Lithuanian Language and the Dictionary of Lithuanian Language, maintained by the Lithuanian Language Institute;
- „Tildės biuras" software package, incuding dictionaries for English, German, Latvian, Russian and Lithuanian languagies;
- Cobuild English-Lithuanian-Czech Dictionary;
- Multilingual dictionaries Stella and Etoile by Akelote;
- Database of Old Lithuanian Writings and the Dictionary of Toponyms maintained by the Lithuanian Language Institute;
- Open terminological database monitored by the State Commission of the Lithuanian Language;
- Database combining digitalized term dictionaries from 27 different branches, monitored by the Institute of Mathematics and Informatics;
- Database of Lithuanian nominal collocations built using the CCLL corpus
- etc.



HLT in Lithuania: overview (3)

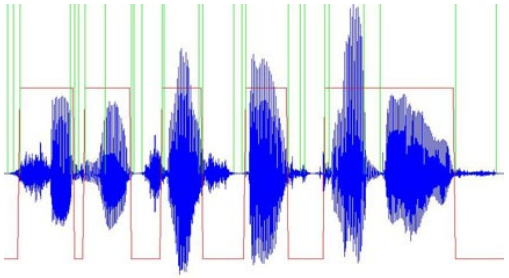
Tools:

- Corpus query and concordance extraction tools;
- Collocation extraction tool using Gravity Counts method;
- Spellcheckers and grammar checkers „Juodos avys“ and “Tildės biuras”;
- Lemmatizer;
- Aligner for compiling parallel corpuses;
- Automatic morphological annotator/tagger;
- Automatic accentuation tool;
- Autonomous translation system „Vertimo vedlys“ (part of „Tildės biuras“);
- On-line English-Lithuanian machine translation tool;
- Semantic ontological annotation tool (prototype).

HLT in Lithuania: overview (4)

Speech processing and synthesis:

- ***Technologies:***
  - Speaker identification and verification technology - Institute of Mathematics and Informatics;
  - Speech recognition (voice commands) and spoken dialogue systems – Kaunas University of Technology;
  - Text-to-speech synthesizer "Aistis" (Vilnius University and Kaunas University of Technology)
  - Large vocabulary continuous speech recognition (Vytautas Magnus University)
- ***Speech resources:***
  - LTDIGITS corpus of digit sequences and voice commands (6h, 350 speakers)
  - LRN corpus of radio news (10 h, 23 speakers);
  - LAB50 corpus of read speech (50h, 50 speakers);
  - VMU corpus of spontaneous speech (10h, 18 speakers);
  - Archive of Lithuanian Dialects;
  - Corpus of diphones/



VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

HLT classification: the Sarasola scheme

- Sarasola's typology of language technology resources (2000):

Foundations	Raw or untagged corpus, lexicons, machine-readable dictionaries, speech databases
Basic tools	Statistical tools for corpus treatment, morphological analyzers/generators and lemmatizers, POS taggers and POS-tagged corpora, speech recognition systems dealing with isolated words
Medium-complexity tools	Spell checkers, structured lexical databases including multiword lexical units, surface syntax analyzers, Web crawler managing languages, environment for tool integration
Advanced tools	Syntactically annotated corpora (treebanks), grammar and style checkers, lexical-semantic knowledge bases or concept taxonomies such as WordNet, word sense disambiguators, speech processing tools functioning at sentence level
Multilinguality and general applications	Semantically annotated corpora, information retrieval and extraction, dialogue systems, language learning systems, machine translation systems, multilingual lexical-semantic databases

VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Lithuanian HLT resources by the Sarasola scheme

Foundations

Raw or untagged corpus:
CCLL (160 mln words); CorALit (9 mln words); Parallel bidirectional Lithuanian-Czech corpus (5 mln words); Parallel English-Lithuanian corpus (18 mln words); Corpus of Spoken Lithuanian Language (200000 words); Corpus of Lithuanian Dialects; DB of Old Lithuanian Writings

Lexicons, machine-readable dictionaries:
•*Explanatory*: Dict. of Lithuanian Language, Dict. of Contemporary Lithuanian ;
•*Translation*: Alkonas, Anglonas, Frankonas, LED, WinLED, "Tildės Biuras", Cobuild, Etoile, etc.;
•*Special*: Lithuanian Term Bank, DB of digitalised term dictionaries from 27 branches, Dict. of international words *Interleksis,* Dict. of Toponyms, DB of nominal collocations, etc.

Speech databases:
DB of Speech Recordings for the Common Lithuanian language; Archive of Lithuanian dialects

Basic Tools

On-line corpus query tools (concordances, etc.); Morphological analyzer and tagger; Statistical tools (frequency lists, etc.); Collocation extraction tools; Automatic accentuation tool; Automatic identification of text functions; Morphologically annotated corpus of Lithuanian; Universal annotated DB of speech recordings, etc.

Medium -compl. Tools

Spellcheckers; Morpho-syntactic analyzer; DB of Lithuanian nominal collocations

Adv. Tools

Grammar checkers; Semantic-ontological annotation tool; Experimental concept ontology for semantic-ontological mark-up

Gen. appl.

Rule-based autonomous and on-line English-Lithuanian translation systems;
Learning tools (word analyser and synthesizer, etc.)

VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Analysis of Lithuanian HLT resources using the Sarasola Scheme

- Foundations – a collection of text and speech corpus exists, though:
  - Their expansion is still a task
  - Corpora and other resources are designed by different institutions and available only via specialized, individual access tools, or even unaccessible
  - In many cases, standards, such as TEI P5, are underused
- The collection of resources and tools is rather fragmented – lack of coordination in development
- Obvious lack of tools, especially those belonging to the "advanced" and "multilinguality and general applications" categories
- Only first attempts to create tools for semantic annotation and analysis as well as to compile semantically annotated resources.
  - This bravch should be more developed taking into account the needs of emerging Semantic Web.
- Tools mainly oriented towards researchers, very few on-line tools for users

VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

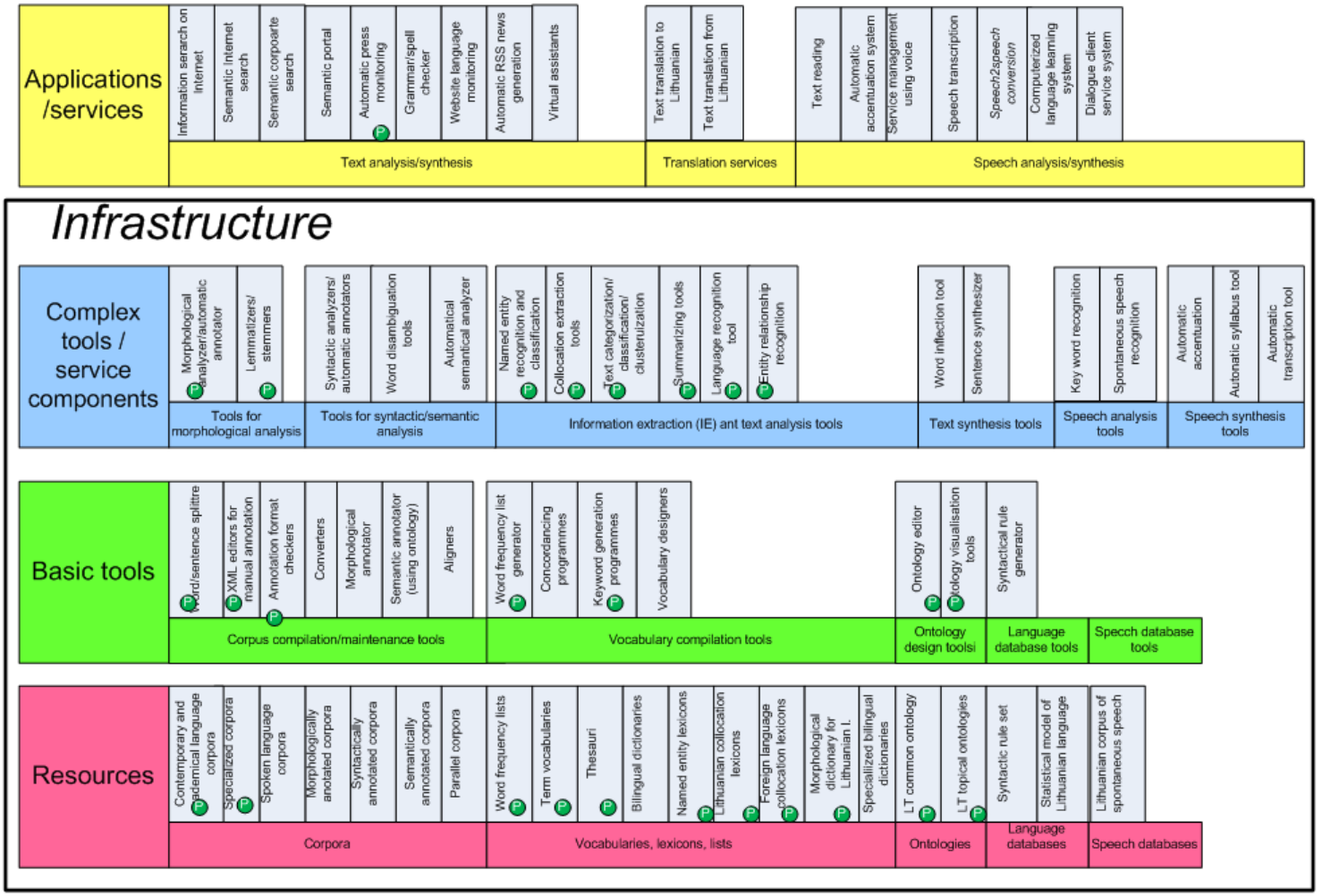
HLT classification: "Resource-Tool-Application" scheme

- Sarasola scheme does not draw a clear separating line between resources and tools
- Such a scheme is not convenient enough for analysing the structure and resources needed for composite services
- A "Resource-Tool-Application" classification scheme is proposed (here different complexity levels for tools and resources can be separated)

Applications/ Services	**1) Institutional (for researchers) 2) On-line (all users) E.g. – search, machine translation, dialogue systems, etc.**
Complex tools	**Diverse analysis tools (information extraction, syntactical, semantical); synthesis tools (text and speech); service components**
Basic tools	**Tools for resource compilation; annotation tools (formating, morphosyntactical, semantical annotation)**
Resources	**Corpora speech databases, vocabukaries and ontologies, lexical databases, etc. 1) Institutional 2) national**

VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

"Resource-Tool-Application" scheme – tool/resource need evaluation for a service



VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Analysis using the "Resource-Tool-Application" scheme

- Allows to evaluate the needs for resources and tools of different complexity levels for designing new applications and services
- Systematic approach towards tools and resourses gives a possibility to track different sets or combinations that enable their users to reach their goals in different ways
- Such classification scheme can help to compare several implementation alternatives for the same application or service, based on their complexity, depending on the availability and complexity of their constituent parts.
- Also, such a scheme could be useful for prioritizing the compilation of new HLT resources or tools by visualizing their need in application and service plans.

VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Strategic issues in building the Lithuanian HLT infrastructure

- Standardization of HLT resources and tools
- Design of a federated system for joint HLT resource access and reuse
- Rational planning of the most needed HLT resources and tools as well as their implementation strategies
- Alignment with European initiatives

VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Standardization of HLT resources and tools

- ***Driving needs:***
  - Exploiting possibilities to design services and applications, combining various resources and tools
  - Participation in large-scale international projects
  - Use of open-source and other available tools
- ***Main activities:***
  - Adaptation of the largest corpora (CCLL and CorALit) to the TEI (Text Encoding Initiative) P5 encoding standard
  - On-line access to the functionality of the main HLT tools via SOAP web service interface

VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Design of a federated system for joint HLT resource access and reuse

- Standardization – first step in designing an open national infrastructure of Lithuanian HLT resources
- National eLingua project addresses the design issues of a common and highly interoparable virtual system of resources
  - Federation-oriented architecture planned for resource storage
  - Both centralized and institution-specific repositories
  - Interoperable machine-to-machine interaction over the Internet
  - Open access to resources and tools prevails
  - Flexible intellectual property protection mechanisms upon need
- Design of common national infrastructure also opens possibilities for planning

VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Alignment with European initiatives

- Development and support of resources in the framework of National Research Infrastructure compatible with ESFRI requirements for national states
- The strategy of NRI includes documentation and unification of existing national resources as well as support for trans-national initiatives such as CLARIN, CESSDA and other similar joint infrastructures
- Processes of joining the CLARIN at the national level are rather slow due to decision and financing problems
- First steps are easier to implement at the institutional level by accomplishing the following actions:
  - Institutional CLARIN membership (LLI and VMU agreements);
  - Establishment of CLARIN C and B type centres by preparing and opening access to the CLARIN-compatible metadata system for HLT resources and corresponding services (corresponding project started by VMU)

VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Participation in CLARIN network – open questions

- Detailed guidelines (requirements) for CLARIN centre establishment
- Recommendations regarding standards, data formats, etc.
- Unified legal bases for copyright issues

VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Conlusions

- Lithuanian HLT community that started from scratch two decades ago has advanced in creating tools and resources

- In that way it contributed to the preservation of the Lithuanian language in its digital format

- However, there is much still to be done:
  - The existing tools and resources have to be compatible and accessible as one national HLT infrastructure
  - New advanced tools and resources have to be created to fill in the gaps
  - The national infrastructure has to be integrated into EU and transnational networks in order to enable multilingual HLT applications

- The major strategic steps towards the ultimate goal are:
  - **become digital,**
  - **become standard and integrated,**
  - **apply tools and communicate.**

VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Thank you!

Contacts:

r.marcinkeviciene@hmf.vdu.lt

d.vitkute@if.vdu.lt