The META-NET Whitepaper Series on European Languages

Andrejs Vasiļjevs

Tilde, Latvia

andrejs@tilde.com

META-NET TEAM

META FORUM 2011 Solutions for Multilingual Europe
Budapest, Hungary, June 27/28

Co-funded by the 7th Framework Programme and the ICT Policy Support Programme of the European Commission through the contracts T4ME, CESAR, METANET4U, META-NORD (grant agreements no. 249119, 271022, 270893, 270899).

Outline

- “Yet another study”?
- General Approach
- Preliminary Results
- Conclusions

Why Yet Another Survey?

- META-NET aims at a concerted effort to improve monolingual and multilingual LT support for all European languages.
- The degree to which LT is used in Europe varies from language to language depending on the commercial relevance of the language, the problems the language poses for automatic processing, and the research already devoted to it.
- So far, no one has ever evaluated the state of European languages in regards to LT support or their state in the digital information age.

The Language White Papers

- Survey of the state of the respective language in the digital society.
- Provide expert estimations about the current status and availability of language resources and technologies.
- Are meant to inform politicians, journalists and the public at large about societal and technological problems, challenges, and economic opportunities.

29 Languages Covered so far

- Basque
- Bulgarian*
- Catalan
- Czech*
- Danish*
- Dutch*
- English*
- Estonian*
- Finnish*
- French*
- Galician
- German*
- Greek*
- Hungarian*
- Icelandic
- Irish*
- Italian*
- Latvian*
- Lithuanian*
- Maltese*
- Norwegian
- Polish*
- Portuguese*
- Romanian*
- Serbian
- Slovak*
- Slovene*
- Spanish*
- Swedish*

* = Official EU language

Structure of the White Papers

- Executive Summary
- Part 1: Introduction – A Risk for Our Languages and a Challenge for Language Technology
- Part 2: *Language* in the European Information Society
- Part 3: Language Technology Support for *Language*
- Part 4: About META-NET
- References

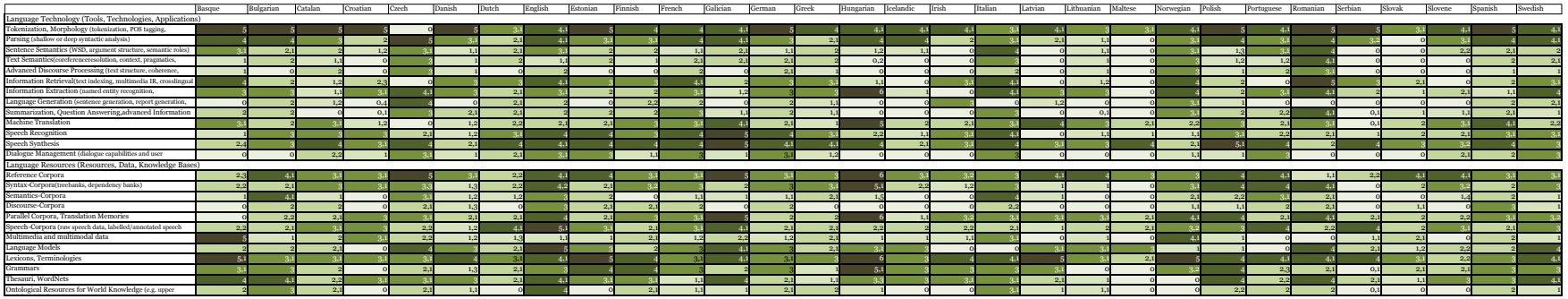
Assessing LT Support

- How to assess Language Technology support for a certain language?
- How to arrive at a result that can be communicated?
  - Count all existing tools and resources? => Does not result in a message.
  - Define quality criteria and perform a comparative evaluation? => Complicated, complex, time-consuming process, would take too long.
- For the White Papers, experts provided estimations condensed in a one table assessing core technology areas and resources such as:
  - Parsing, Information Retrieval, Machine Translation, Speech Recognition, Speech Synthesis, Reference Corpora, Language Models, Thesauri, etc.
- Assessment is done along criteria such as availability, quality, or coverage, maturity, sustainability and adaptability (see the example White Paper in your conference bag).

Preliminary Results

Overall Evaluation

- The tools & resources tables in the current papers allow meaningful interpretation per language (Where are gaps? What is there?).
- Calibration between languages has happened in smaller groups, but still needs to be performed across all languages (after META-FORUM 2011).
- Resultants for each tool and resource have been derived from the two central features quality and coverage, resulting in a big table:



	Basque	Bulgarian	Catalan	Croatian	Czech	Danish	Dutch	English	Estonian	Finnish	French	Galician	German	Greek	Hungarian	Icelandic	Irish	Italian	Latvian	Lithuanian	Maltese	Norwegian	Polish	Portuguese	Romanian	Serbian	Slovak	Slovene	Spanish	Swedish
Language Technology (Tools, Technologies, Applications)																														
Tokenization, Morphology (tokenization, POS tagging,	5	5	5	5	5	5	5	5	4	5	4	4	4	5	4	4	4	4	4	3	3	4	5	4	5	5	5	4	5	4
Parsing (shallow or deep syntactic analysis)	4	4	3	2	5	3	2	4	3	3	4	4	5	2	4	4	2	3	2	1	0	3	4	3	4	3	0	3	4	4
Sentence Semantics (WSD, argument structure, semantic roles)	2	2	2	2	3	1	2	3	2	2	1	2	1	2	2	1	0	3	0	1	0	3	1	3	4	0	0	2	2	2
Text Semantics(coreferenceresolution, context, pragmatics,	1	2	1	0	3	1	2	1	2	1	2	2	2	2	0	0	0	3	0	1	0	3	1	2	4	0	0	0	2	2
Advanced Discourse Processing (text structure, coherence,	1	0	2	0	3	0	2	3	0	0	2	2	2	1	0	0	0	2	0	0	0	3	1	2	3	0	0	0	1	1
Information Retrieval(text indexing, multimedia IR, crosslingual	4	2	1,2	2,3	0	3	4	4	3	4	4,1	2	3	3,1	1,1	0	0,1	4,1	0	1,2	0	4	2	0	5	2,1	0	0	2	2,1
Information Extraction (named entity recognition,	3	3	1,1	3,1	4,1	3	2,1	3,1	2	2	3,1	1,2	3	3	6	1	0	4,1	3	3	0	4	2	3,1	4,1	3	1	2,1	1,1	4
Language Generation (sentence generation, report generation,	0	2	1,2	0,4	3	0	2,1	3	0	2,2	2	0	3	1,1	0	0	3	0	1,2	0	0	3	1	0	0	0	0	0	2	2,1
Summarization, Question Answering,advanced Information	2	2	0	0,1	3	2,1	2,1	2	2	2	3	1,1	2	1,1	0	0	0	3	0	0,1	0	3,1	2	2,2	4,1	0,1	1	1,1	2,1	1
Machine Translation	2,1	2	3,1	1,2	0	1,2	2,2	2,1	2,1	3	3,1	4,1	2,1	1	0	2	2,1	3,1	4	3	2,1	2,2	3	2,1	3,1	0,1	2	3,1	4,1	2,2
Speech Recognition	1	3	3	3	2,1	1,2	3,1	4	4	4	4	5	4	2,1	2,2	1,1	3,1	4,1	0	1,1	0	1,1	3,1	2,2	2,1	2	2	2,1	3,1	3
Speech Synthesis	2,1	3	4	3,1	2,1	4	4,1	4	4	4	4	5	4,1	4	4	2,1	3,1	4	3,1	3	2	2,1	3,1	4	2	4	3	3,1	4	3
Dialogue Management (dialogue capabilities and user	0	0	2,2	1	3,1	1	2,1	3,1	3	1,1	3	1	3,1	1,2	0	0	0	3	0	0	0	0	1,1	1	3	0	0	0	2,1	2
Language Resources (Resources, Data, Knowledge Bases)																														
Reference Corpora	2,3	4,1	3,1	3,1	5	3,1	2,2	4,1	4	3,1	3,1	5	3,1	3	6	3,1	3,2	4	4,1	4	3	3	4	4,1	1,1	2,2	4,1	4,1	3,1	3,1
Syntax-Corpora(treebanks, dependency banks)	2,2	2,1	1	3,1	3,1	1,3	2,2	4,2	2,1	3,2	3	2	3	3,1	5,1	2,2	1,2	3	1	1	0	3,1	4	4	4,1	0	2	3,2	2	3
Semantics-Corpora	1	3,1	1	0	3,1	1,2	1,2	3	2	2	3	1,1	3	2,1	1,2	0	0	3	1	0	0	2,1	2,2	3,1	3,1	0	0	1,1	2	1
Discourse-Corpora	0	2	2	0	2,1	1,3	0	3	2,1	2,1	2	2	2	0	0	0	0	2,2	0	0	0	1,1	1,1	2	3,1	0	1,1	0	3	1
Parallel Corpora, Translation Memories	0	2,2	2,1	3	3,1	2,1	2,1	4	2,1	3	3,1	6	2	2	6	1,1	3,2	3,1	3,1	3,1	2,1	4,1	4	2,1	4,1	2,1	2	2,2	3,1	3,2
Speech-Corpora (raw speech data, labelled/annotated speech	2,2	2,1	3,1	3	2,2	1,2	4,1	5,1	3,1	2,1	3,1	4,1	2,1	2,1	2,2	3	3,2	2,1	1	2	2,1	3,2	3	4	2,2	2	2	3,1	2,1	3
Multimedia and multimodal data	5	1	2	3,1	2,2	1,2	1,3	1,1	1	2,1	1,2	2,2	1,2	2,1	1	1	1,1	3,1	0	1	0	4,1	1	0	0	1,1	2,1	0	2	1
Language Models	2	2	2,1	0	4	3	2,1	5	3	2	3	4,1	3	2,1	3,1	3	0	0	3,1	2,1	1	1	1	0	4	2,1	1,2	2,2	2	4
Lexicons, Terminologies	3,1	3,1	3,1	3,1	3,1	3	3,1	4,1	5	4	3,1	4,1	3,1	3	6	3	4	4,1	5	3,1	2,1	5	4	4,1	4,1	4	3,1	2,2	3	4,1
Grammars	3,1	3	3	0	2,1	1,3	2,1	3	4	4	3	2	3	1	5,1	3	3	3	3,1	0	0	3,2	4	2,3	2,1	0,1	2,1	2,1	3	3
Thesauri, WordNets	4	4,1	2,2	3,1	3	3	2,1	4,1	3,1	3,1	1,1	4	2,1	1,1	3,2	3	3,1	3,1	2,1	1	0	0	4	2,2	4	2,1	1,1	3	3	4,1
Ontological Resources for World Knowledge (e.g. upper	2	3	2,1	0	2,1	1,1	0	4	0	2,1	1,1	1	2,1	2	1	0	0	3,1	1	1,1	0	0	2,2	2	2	0,1	0	0	2	1

Preliminary Results

- Overall ranking of the quality and coverage of tools and resources is plausible (see the top and bottom tools and technologies on the right).
- Scale:
  0 = non-existent
  6 = perfect

3.8	Tokenization, Morphology
3.3	Speech Synthesis
2.7	Parsing
2.4	Machine Translation
2.4	Speech Recognition
	…
1.2	Text Semantics
0.9	Language Generation
0.8	Advanced Discourse Processing

Preliminary Results

- If one takes a value of at least 4 as threshold for practical usability of a technology/resource, then, e.g.,:
  - 13 of the 30 languages lack sufficient support in speech synthesis
  - 18 of the 30 languages lack sufficient support in parsing
  - 26 of the 30 languages lack sufficient support in machine translation

Example: MT Results

- For an example ranking and comparison,
  - a few values were manually adjusted and
  - normalization was carried out for each language, all technologies: $(x\text{-}m)/d$ with m mean and d standard deviation.

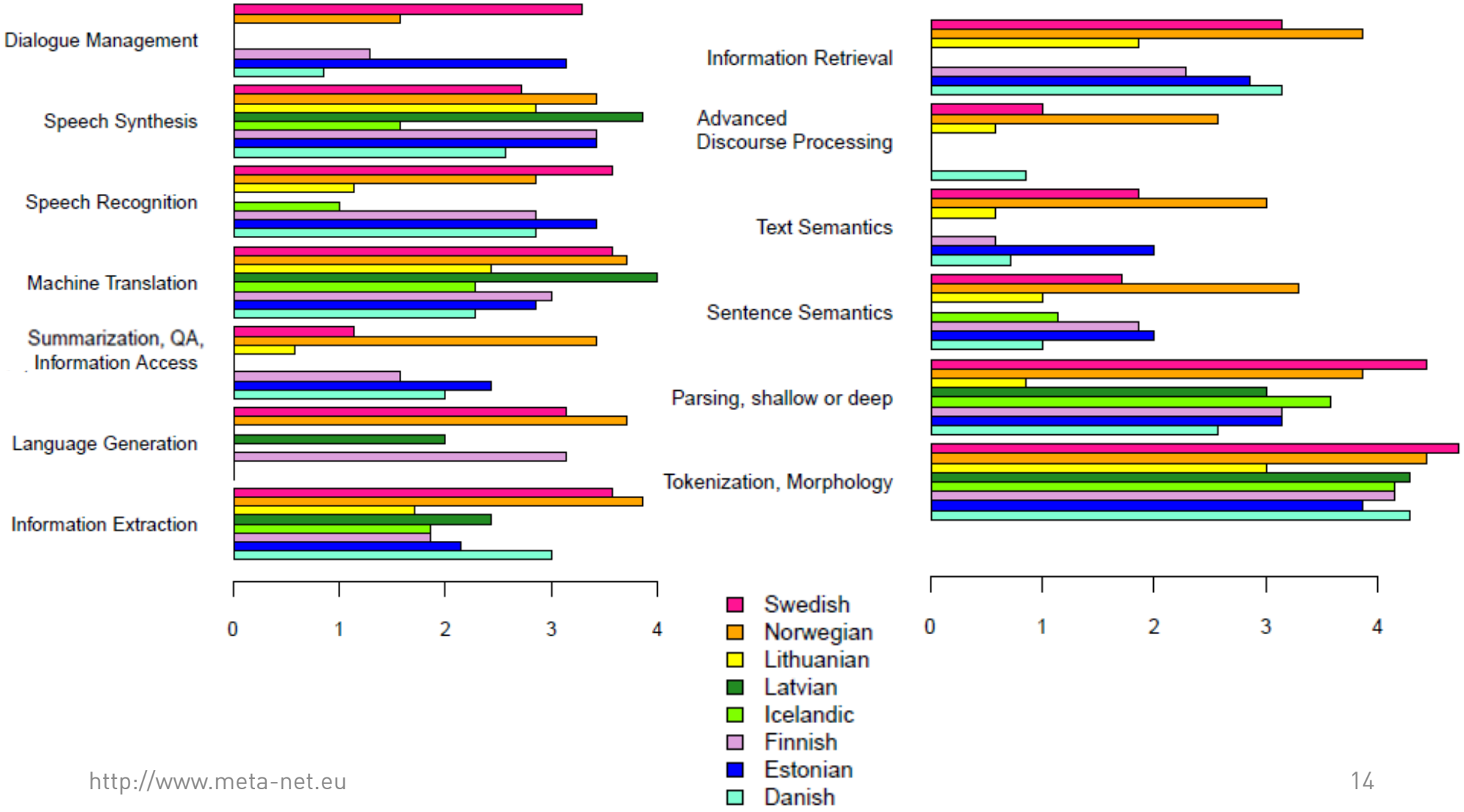
Cluster1 Higher ranks	Cluster2 Medium ranks	Cluster 3 Lower ranks
Spanish, English, Latvian, Lithuanian, Maltese, Galician, Slovene, Hungarian, Catalan	Polish, Icelandic, French, Irish, Basque, Italian, Romanian, Bulgarian, Dutch, Portuguese, Estonian, Finnish	Croatian, German, Swedish, Czech, Norwegian, Danish, Serbian, Greek
Includes romanian languages with active research.		*Difficult languages and languages with little research.*

Example: Comparing the Situation for Parsing

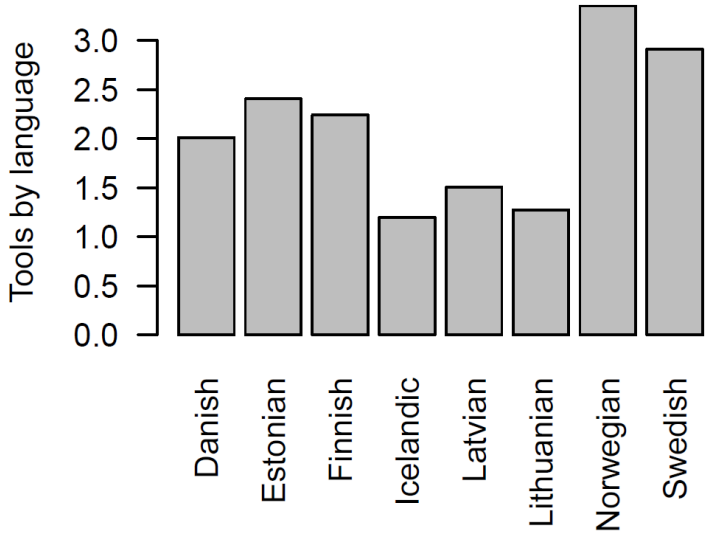
Cluster1	Cluster2	Cluster 3	Cluster 4
Czech, Dutch, English, German	Bulgarian, French, Norwegian, Polish, Portuguese, Spanish, Swedish	Basque, Catalan, Danish, Finnish, Galician, Hungarian, Irish, Italian, Romanian	Croatian, Estonian, Greek, Icelandic, Latvian, Lithuanian, Maltese, Serbian, Slovak, Slovene

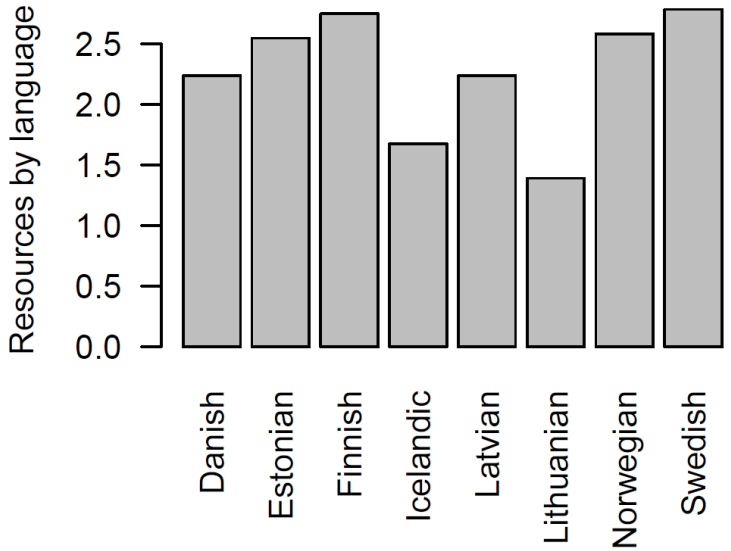
For this example ranking, a few values were manually adjusted.

Example: Findings from the Nordic and Baltic White Papers



Example: Findings from the Nordic and Baltic White Papers





General Conclusions

- Speech processing and synthesis appears to be more mature than processing of written text. Advanced information access technology is in its infancy.
- Research was successful in designing particular high quality results in some areas, but many of the resources lack standardization, i.e., even if they exist, sustainability is not given; concerted programmes and initiatives are needed to standardize data and interchange formats.
- Most (very) large companies have stopped working in the area, leaving the field to SMEs, which can hardly attack an international market.

Q/A

Thank you.

office@meta-net.eu

http://www.meta-net.eu
http://www.facebook.com/META.Alliance