META NORD

Review Meeting
Luxembourg
April 11, 2013

Key Findings of the Language White Papers

Eiríkur Rögnvaldsson
HI

META-NET

The META-NET Language White Papers

- 31 Language White Papers
  - published by Springer in September 2012
- 9 White Papers for the 8 META-NORD languages
  - two for Norwegian (Bokmål and Nynorsk)
- The White Papers contain information regarding
  - general facts about each language and its particularities
  - recent developments in the language
  - Language Technology support for the language
  - core application areas of language and speech technology

State of Language Technology Support

- The language white papers present a cross-language comparison ranking the respective language within four key areas:
  - machine translation, speech processing, text analysis, and language resources
- Experts were asked to rate the existing tools and resources with respect to seven criteria:
  - quantity, availability, quality, coverage, maturity, sustainability, and adaptability
  - on a scale of 0 (no tools/resources) to 6 (well presented)

METANORD

Danish

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technology: Tools, Technologies and Applications							
Speech Recognition	4	2	–*	4	4	3	3
Speech Synthesis	5	2	–*	3	3	2	3
Grammatical analysis	3	2	4	4	3	2	3
Semantic analysis	1	0	1	1	1	1	1
Text generation	0	0	0	0	0	0	0
Machine translation	3	2	2	3	3	1	2
Language Resources: Resources, Data and Knowledge Bases							
Text corpora	4	3	4	3	4	4	3
Speech corpora	3	3	3	1	2	2	2
Parallel corpora	3	1	3	2	2	2	2
Lexical resources	3	3	4	4	3	3	3
Grammars	2	1	4	1	2	2	2

METANET

META NORD

Estonian

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technology: Tools, Technologies and Applications							
Speech Recognition	2	5	2.8	2.8	3	3	3
Speech Synthesis	2	5	2.8	2.8	3	2	3
Grammatical analysis	2.5	3.5	3.2	2.8	4	2.5	3.5
Semantic analysis	1	1.3	0.9	0.9	1.3	1.3	1.7
Text generation	0	0	0	0	0	0	0
Machine translation	3	3	1.4	2.1	3	4	2
Language Resources: Resources, Data and Knowledge Bases							
Text corpora	3	5	2.5	2.1	3	2.5	
Speech corpora	2	5	2.1	2.8	4	4	4
Parallel corpora	2	2	2.1	1.4	3	3	2
Lexical resources	3.5	4	3.2	2.8	3.5	3.5	3.5
Grammars	2	5	2.8	2.8	3	3	3

META NET

META NORD

Finnish

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technology: Tools, Technologies and Applications							
Speech Recognition	3	2	4	3	3	3	4
Speech Synthesis	3	3	5	4	4	4	4
Grammatical analysis	3,5	3,5	3,5	4	4	3,5	3,5
Semantic analysis	0,4	0,4	1	1	1	1,4	0,7
Text generation	3	3	4	2	3	3	4
Machine translation	3	1	4	2	3	1	2
Language Resources: Resources, Data and Knowledge Bases							
Text corpora	3	4	4	3,5	3,5	3,5	4
Speech corpora	2	3	3	2	2	2	2
Parallel corpora	1	2	3	2	2	3	3
Lexical resources	3	4	3,5	4	3,5	3,5	3,5
Grammars	2	5	4	4	4	3	3

META NET

META NORD

Icelandic

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technology: Tools, Technologies and Applications							
Speech Recognition	1	1	1	1.5	1	0	1
Speech Synthesis	1	1	2.5	2.5	2	1	1
Grammatical analysis	2	5.5	4	3	3.5	3.5	3
Semantic analysis	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Text generation	0	0	0	0	0	0	0
Machine translation	1	4	1	1.5	1.5	1.5	2
Language Resources: Resources, Data and Knowledge Bases							
Text corpora	1.5	4	3	2.5	2.5	4.5	3
Speech corpora	1	2	1.5	1.5	1	1.5	1.5
Parallel corpora	1	1	1	0.5	1	1	1
Lexical resources	1	2	2.5	2.5	2	2	2
Grammars	1	4	2.5	2	2.5	2.5	2

META NET

META NORD

Latvian

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technology: Tools, Technologies and Applications							
Speech recognition	0	0	0	0	0	0	0
Speech synthesis	2	3	4	3	4	3	4
Grammatical analysis	2,5	2	3	3,5	4	3	4
Semantic analysis	1	0	0	0	0	0	0
Text generation	1	2	1	1	2	1	1
Machine translation	3	4	3	2,5	4	3	4
Language Resources: Resources, Data and Knowledge Bases							
Text corpora	2	4	4	3	3	3	4,5
Speech corpora	1	0	1	1	1	1	3
Parallel corpora	1	3	2	2	3	4	4
Lexical resources	3	3,5	4	3	4,5	4,5	4,5
Grammars	2	1	2,5	2	3	4	3

META NET

META NORD

Lithuanian

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technology: Tools, Technologies and Applications							
Speech recognition	2	0	2	1	1	0	2
Speech synthesis	3	2	2,5	2,5	1,5	1	2
Grammatical analysis	2	1,5	2,5	2	1,5	1	2
Semantic analysis	1,3	1	1,3	1	0	0	0,3
Text generation	0	0	0	0	0	0	0
Machine translation	2	3	2,5	2,5	2	2	2
Language Resources (Resources, Data and Knowledge Bases)							
Text corpora	1,5	1,5	2,5	2,5	2	2,5	2,5
Speech corpora	2	1	2	2	1	1	2
Parallel corpora	2	2	1,5	1,5	2	2	4
Lexical resources	2,5	2	2,5	2	2	0,5	2,5
Grammars	0	0	0	0	0	0	0

META NET

META NORD

Norwegian

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technology: Tools, Technologies and Applications							
Speech Recognition	4	2	2	1	2	3	3
Speech Synthesis	3	2	3	2	3	3	3
Grammatical analysis	4	4,5	4	4	4,5	4,5	5
Semantic analysis	2	2	3,3	3	3,7	3,3	3,7
Text generation	1	4	4	3	5	4	5
Machine translation	4	4	2	2	3	5	3
Language Resources: Resources, Data and Knowledge Bases							
Text corpora	4,5	3,5	3,5	3	4	4,5	4
Speech corpora	5	4	3	5	4	5	5
Parallel corpora	5	3	2	2	4	3	3
Lexical resources	2,5	2	2	2	2	2	2,5
Grammars	2	4	5	3	4	5	3

META NET

META NORD

Swedish

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technology: Tools, Technologies and Applications							
Speech Recognition	2	1	3	4	5	5	5
Speech Synthesis	3	1	3	3	3	3	3
Grammatical analysis	4.5	3.5	5	4	5	5	5
Semantic analysis	1.5	1	2	1.5	1.5	1	1.5
Text generation	3	3	3	2	4	3	4
Machine translation	3	1	3	1	4	3	3
Language Resources: Resources, Data and Knowledge Bases							
Text corpora	2	2.5	3.5	3	5	5	5
Speech corpora	4	3	3	3	5	4	4
Parallel corpora	3	1	5	3	5	5	5
Lexical resources	4	2	5	4	3.5	4	4
Grammars	3	2	3	3	3	4	5

META NET

Results - Good Coverage

- Only for the most basic tools and resources, such as tokenisers, PoS taggers, morphological analysers / generators, syntactic parsers, reference corpora, lexical resources and termbanks, the status is reasonably positive for all of the META-NORD languages

Results - Limited Coverage

- All META-NORD languages have some tools for information extraction, machine translation, and speech synthesis
- There are parallel corpora, speech corpora and computational grammars for some of the META-NORD languages
  - **though these are limited in coverage and functionality and are not always sufficiently tested and documented**

Results - Little or no Coverage

- When it comes to the more advanced areas (e.g., sentence and text semantics, information retrieval, language generation, and multimodal data) it appears that one or more of the languages lack tools and resources for these areas

Cross-language Comparison

- An initial comparison across all 30 META-NET languages places three small languages of the Nordic and Baltic region – Icelandic, Latvian, and Lithuanian – in the bottom cluster, defined as having major gaps in all of the four key areas
- The relative ranking of the remaining five META-NORD languages is slightly higher, although none of them come close to the so-called “big” languages (English, French, Spanish, German)

Machine Translation

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	▪ English	▪ French ▪ Spanish	▪ Catalan ▪ Dutch ▪ German ▪ Hungarian ▪ Italian ▪ Polish ▪ Romanian	▪ Basque ▪ Bulgarian ▪ Croatian ▪ Czech ▪ Danish ▪ Estonian ▪ Finnish ▪ Galician ▪ Greek ▪ Icelandic ▪ Irish ▪ Latvian ▪ Lithuanian ▪ Maltese ▪ Norwegian (Bokmål, Nynorsk) ▪ Portuguese ▪ Serbian ▪ Slovak ▪ Slovene ▪ Swedish

Speech Processing

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	▪ English	▪ Czech ▪ Dutch ▪ Finnish ▪ French ▪ German ▪ Italian ▪ Portuguese ▪ Spanish	▪ Basque ▪ Bulgarian ▪ Catalan ▪ Danish ▪ Estonian ▪ Galician ▪ Greek ▪ Hungarian ▪ Irish ▪ Norwegian (Bokmål, Nynorsk) ▪ Polish ▪ Serbian ▪ Slovak ▪ Slovene ▪ Swedish	▪ Croatian ▪ Icelandic ▪ Latvian ▪ Lithuanian ▪ Maltese ▪ Romanian

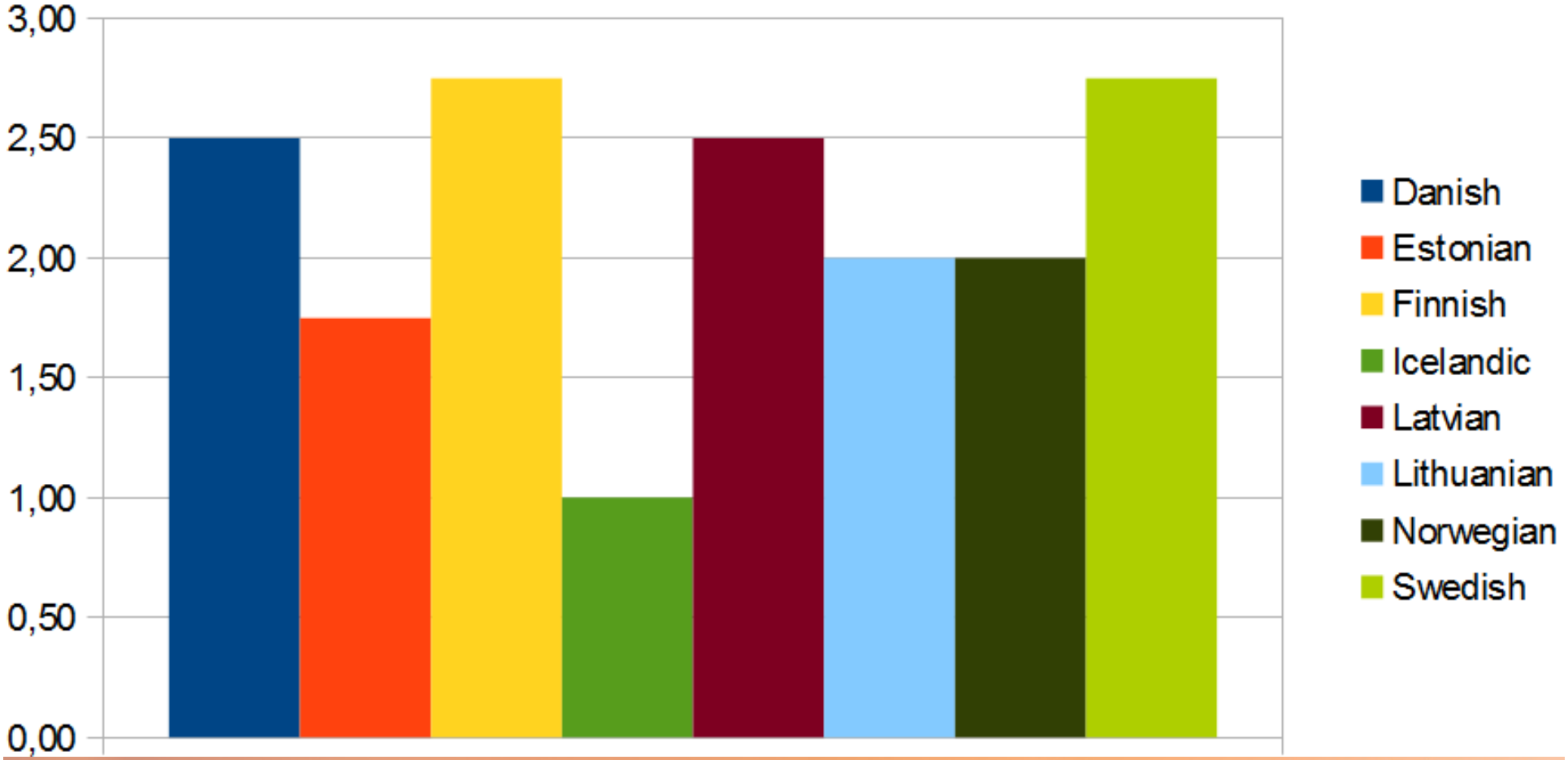
Text Analysis

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	▪ English	▪ Dutch ▪ French ▪ German ▪ Italian ▪ Spanish	▪ Basque ▪ Bulgarian ▪ Catalan ▪ Czech ▪ Danish ▪ Finnish ▪ Galician ▪ Greek ▪ Hungarian ▪ Norwegian (Bokmål, Nynorsk) ▪ Polish ▪ Portuguese ▪ Romanian ▪ Slovak ▪ Slovene ▪ Swedish	▪ Croatian ▪ Estonian ▪ Icelandic ▪ Irish ▪ Latvian ▪ Lithuanian ▪ Maltese ▪ Serbian

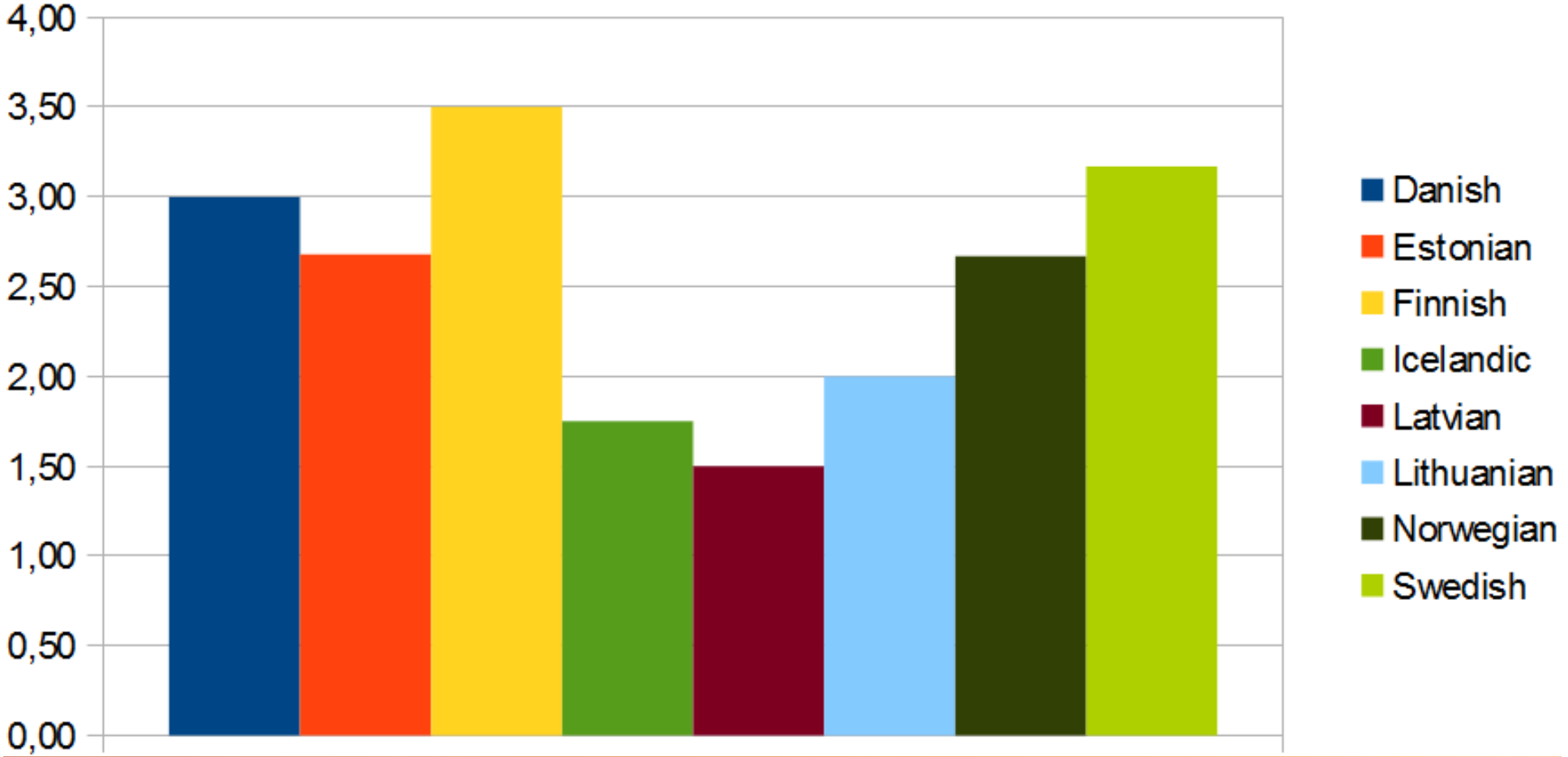
Language Resources

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	▪ English	▪ Czech ▪ Dutch ▪ French ▪ German ▪ Hungarian ▪ Italian ▪ Polish ▪ Spanish ▪ Swedish	▪ Basque ▪ Bulgarian ▪ Catalan ▪ Croatian ▪ Danish ▪ Estonian ▪ Finnish ▪ Galician ▪ Greek ▪ Norwegian (Bokmål, Nynorsk) ▪ Portuguese ▪ Romanian ▪ Serbian ▪ Slovak ▪ Slovene	▪ Icelandic ▪ Irish ▪ Latvian ▪ Lithuanian ▪ Maltese

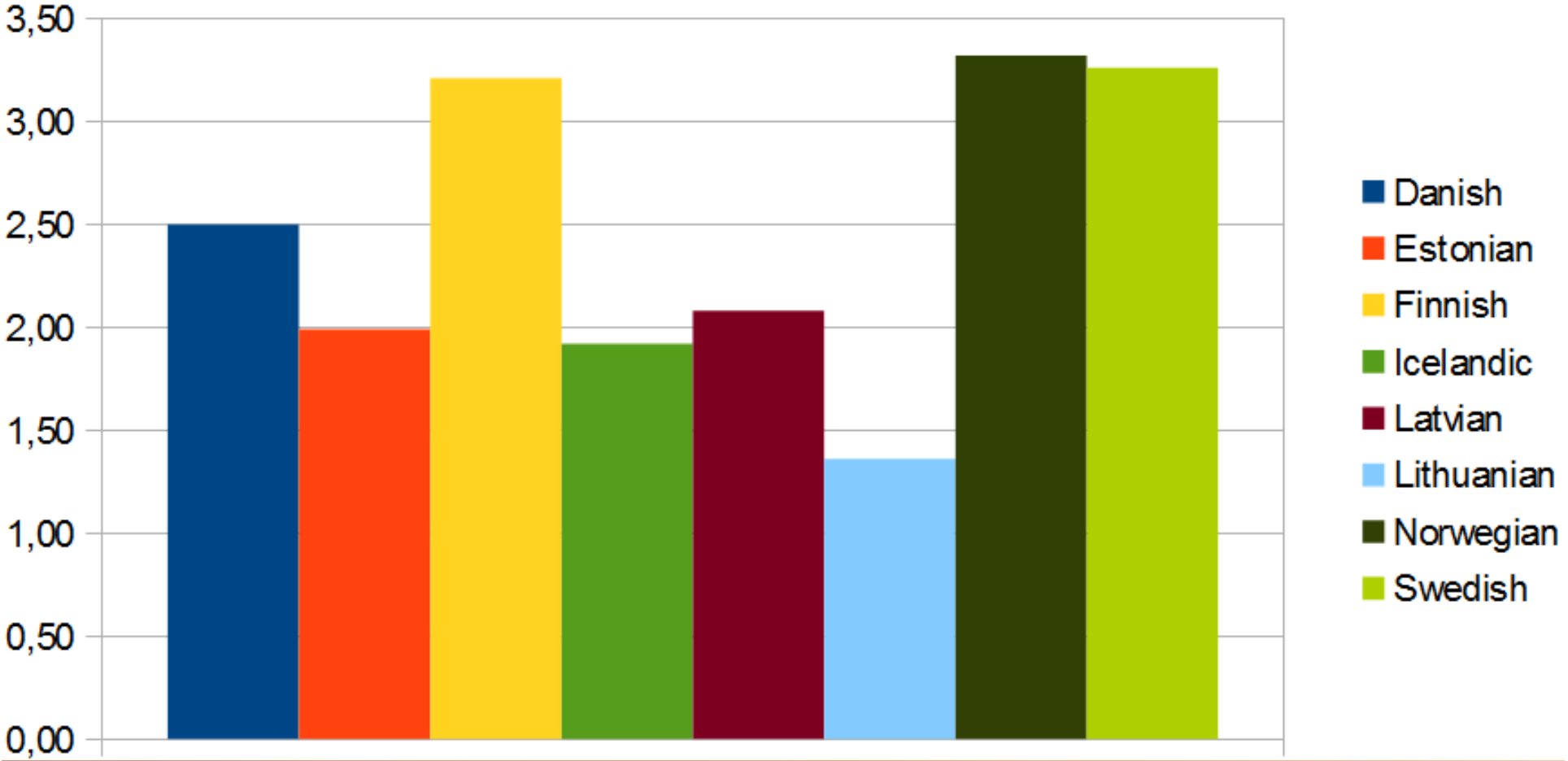
META-NORD - Machine Translation



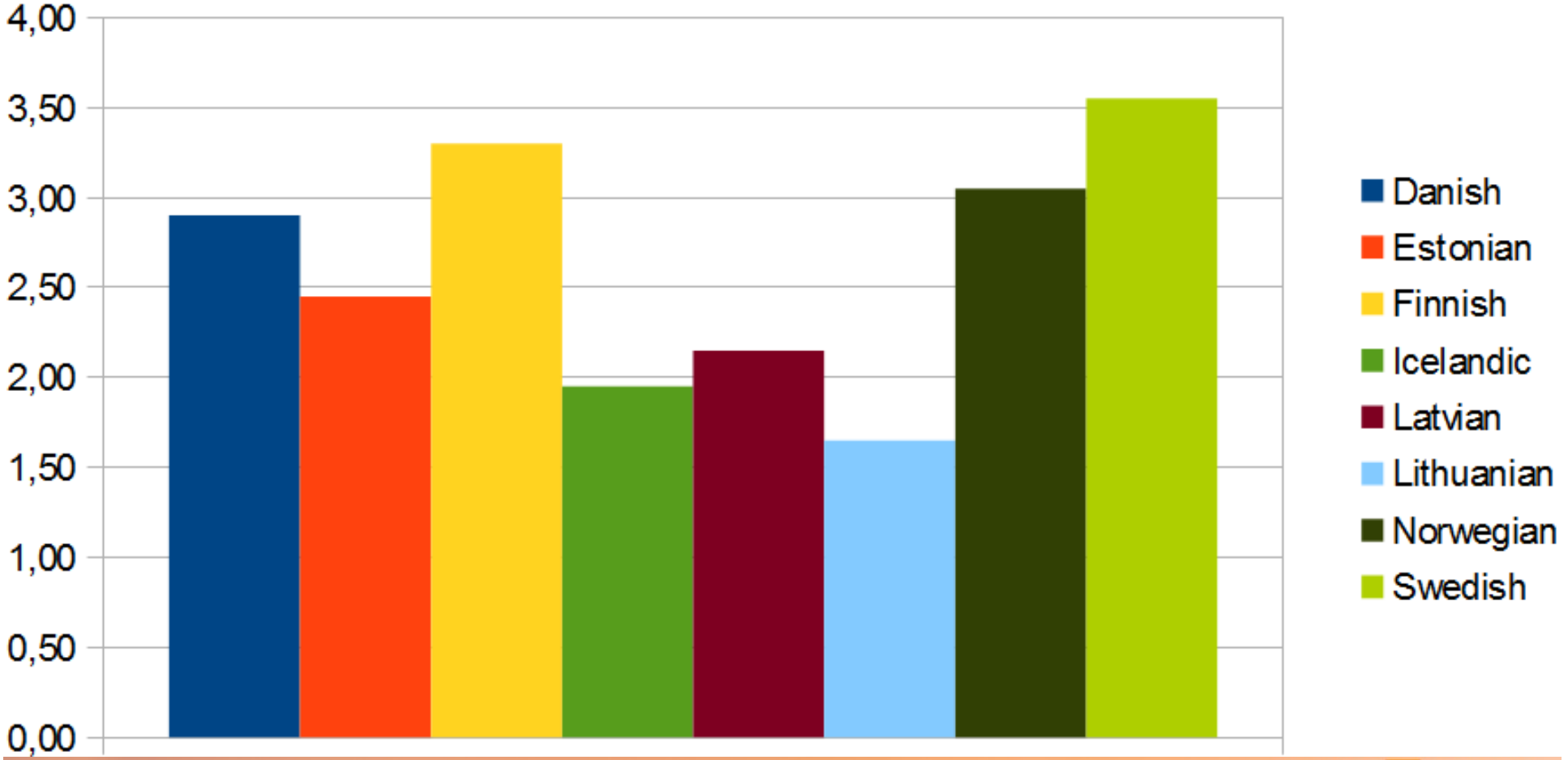
META-NORD - Speech Processing



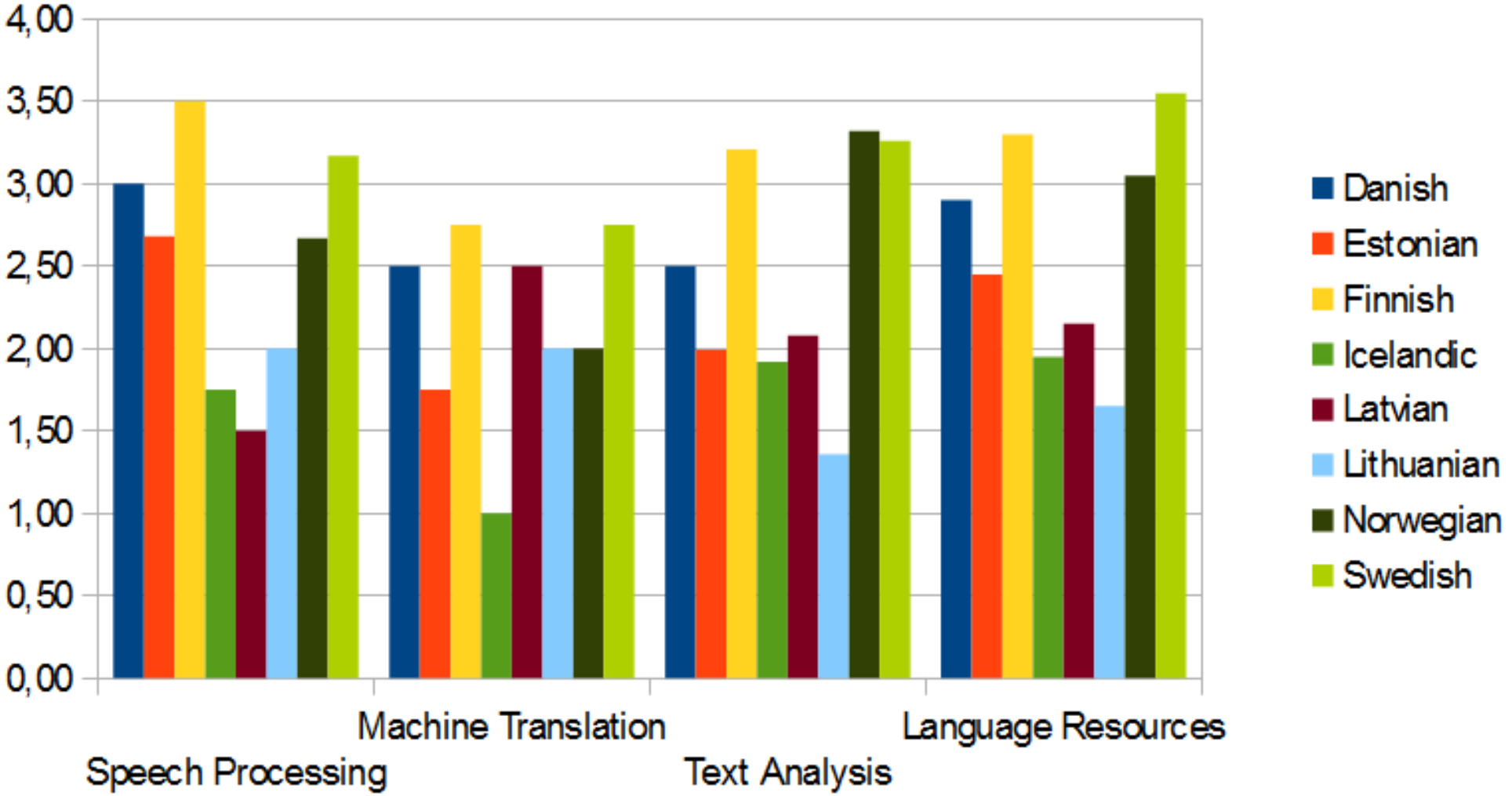
META-NORD - Text Analysis



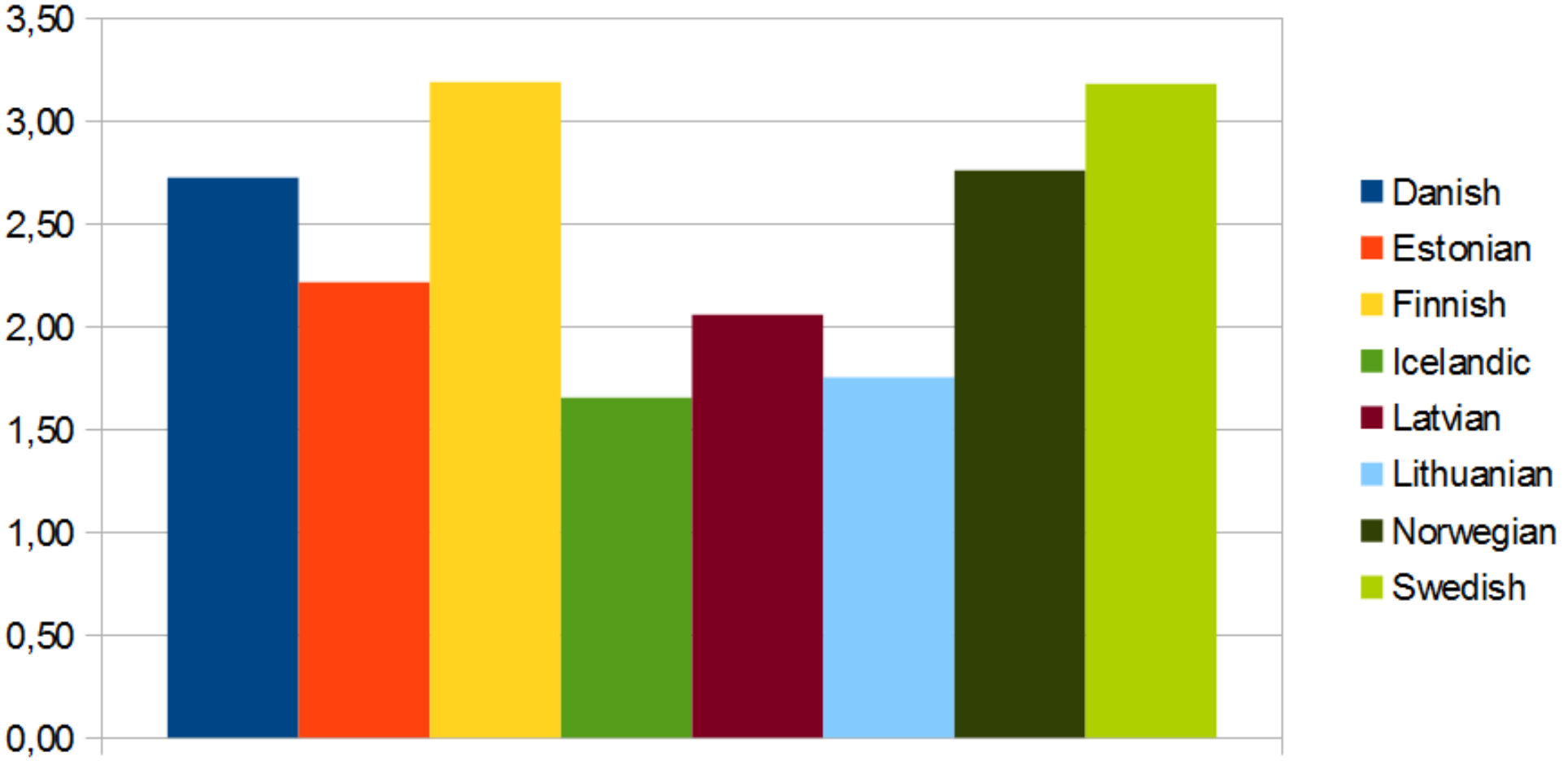
META-NORD - Language Resources



META-NORD - All Categories



META-NORD - Mean Ratings



LWP Conclusions

- “The results of this white paper series show that there is a dramatic difference in language technology support between the various European languages. While there are good quality software and resources available for some languages and application areas, others, usually smaller languages, have substantial gaps. Many languages lack basic technologies for text analysis and the essential resources. Others have basic tools and resources but the implementation of for example semantic methods is still far away. Therefore a large-scale effort is needed to attain the ambitious goal of providing high-quality language technology support for all European languages, for example through high quality machine translation”