The META NORD Language Reports

Koenraad de Smedt
University of Bergen

Eiríkur Rögnvaldsson
University of Iceland

The META-NORD project

- The aim of the recently started META-NORD project is to make basic language resources for the Baltic and Nordic countries more accessible to developers, professionals and researchers to build language enabled applications
- One aspect of this work is to write language reports for the eight main languages in the area: Danish, Estonian, Finnish, Icelandic, Latvian, Lithuanian, Norwegian, and Swedish

Other languages in the area

- Minority languages in the geographic area are not explicitly addressed: Greenlandic, Faroese, Kven and Sami are mentioned only in passing
- Russian is not included
  - even if Northwestern Russia is a part of Northern Europe
  - and Slavic languages are important minority languages in the Baltic countries

The META-NET whitepaper series

- The META-NORD language reports are a part of the META-NET whitepaper series “Languages in the European Information Society”
- They survey the state of each European language with respect to Language Technology and explain the most urgent risks and chances
- They also intend to support the planning of international cooperation on LT
- The series will cover all official European languages

Why language reports?

- There have been a number of valuable and comprehensive scientific studies on certain aspects of languages and technology
  - In the Nordic countries *Vismansrapporten* (Lindén, Koskenniemi and Nordgård, 2006)
- However, there exists no generally understandable compendium that takes a stand by presenting the main findings and challenges for each language
- The META-NET whitepapers will fill this gap

Which information is collected?

- For each of the META-NORD languages, an analysis of the language community has been conducted and the role of the language in the respective language community is described
- The language technology research community and the language service and technology industry are identified
- The importance of language technology products and services in the language community is assessed
- Legal provisions related to language resources and tools are outlined

Form of the reports

- The information gathered in this survey will be presented in eight *parallel* reports
  - each 30-40 pages long
- The reports will be available in English and in the respective languages
- Final versions expected in June 2011

Function and target groups

- Raise awareness for language technology support and the benefits of sharing and exchanging resources by depicting the importance of language technology for every individual language as part of the European information society
- Target audiences are mainly nonexpert influential readers: politicians, national funding bodies and research councils, private companies in the technology sector, universities and research institutions, language councils, journalists

Introductory chapter

- Common report structure for all the META-NET languages, 3 chapters
- Identical first chapter written by experts from the DFKI
- Explain the opportunities and challenges for language technology in the modern information society

Language chapter

- The second chapter is different for each language and written by experts on the language
  - general facts on the language (number of speakers, official status, dialects, etc.)
  - particularities of the language
  - recent developments in the language
  - language cultivation
  - language in education,
  - international aspects
  - the role of the language on the Internet

Language technology chapter

- The third chapter contains subsections on the core application areas of language and speech technology, such as language checking, web search, speech interaction, machine translation, etc.
- The introductory parts are common to all the reports and written by experts from the DFKI
- The language particular parts of the subsections are written by local experts
- Furthermore, there are language particular subsections on language technology in education and language technology programs in each country

Table of tools and resources

- The third chapter ends with a detailed table of Language Technology tools and resources for each language
- Experts were asked to rate the existing tools and resources with respect to seven criteria:
- Quantity, availability, quality, coverage, maturity, sustainability, and adaptability
- The experts were asked to give 13 types of tools and 12 types of resources ratings ranging from 0-6 for each of these criteria

Types of tools

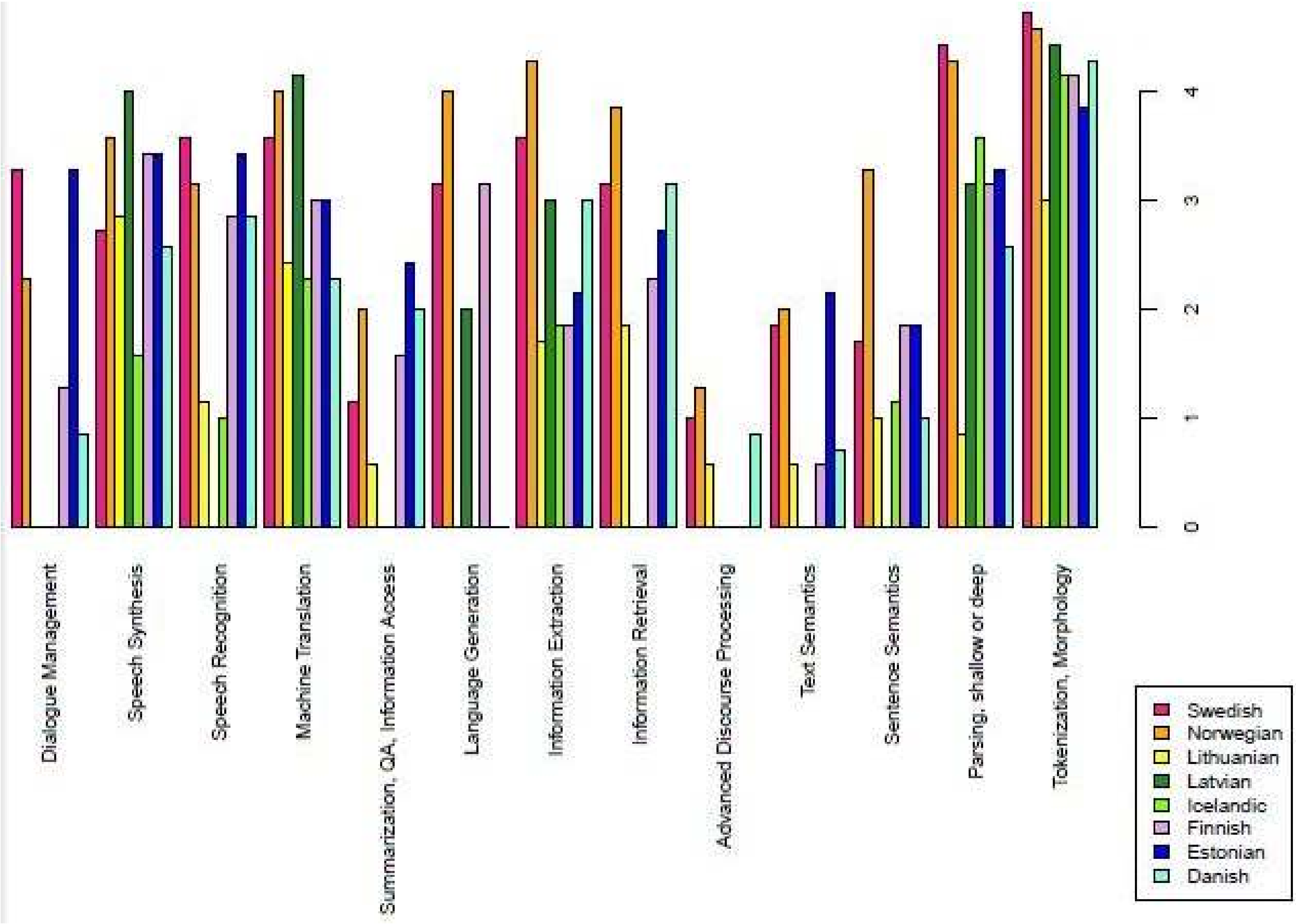
1. Tokenization, Morphology (tokenization, PoS tagging, morphological analysis/ generation)
2. Parsing (shallow or deep syntactic analysis)
3. Sentence Semantics (WSD, argument structure, semantic roles)
4. Text Semantics (coreference resolution, context, pragmatics, inference)
5. Advanced Discourse Processing (rhetorical structure, coherence, argumentative zoning, argumentation, text patterns)
6. Information Retrieval (text indexing, multimedia IR, crosslingual IR)
7. Information Extraction (NER, event/relation extraction, opinion/sentiment recognition)
8. Language Generation (sentence generation, report generation, text generation)
9. Summarization, Question Answering, Advanced Information Access Technologies
10. Machine Translation
11. Speech Recognition
12. Speech Synthesis
13. Dialogue Management (dialogue capabilities and user modelling)

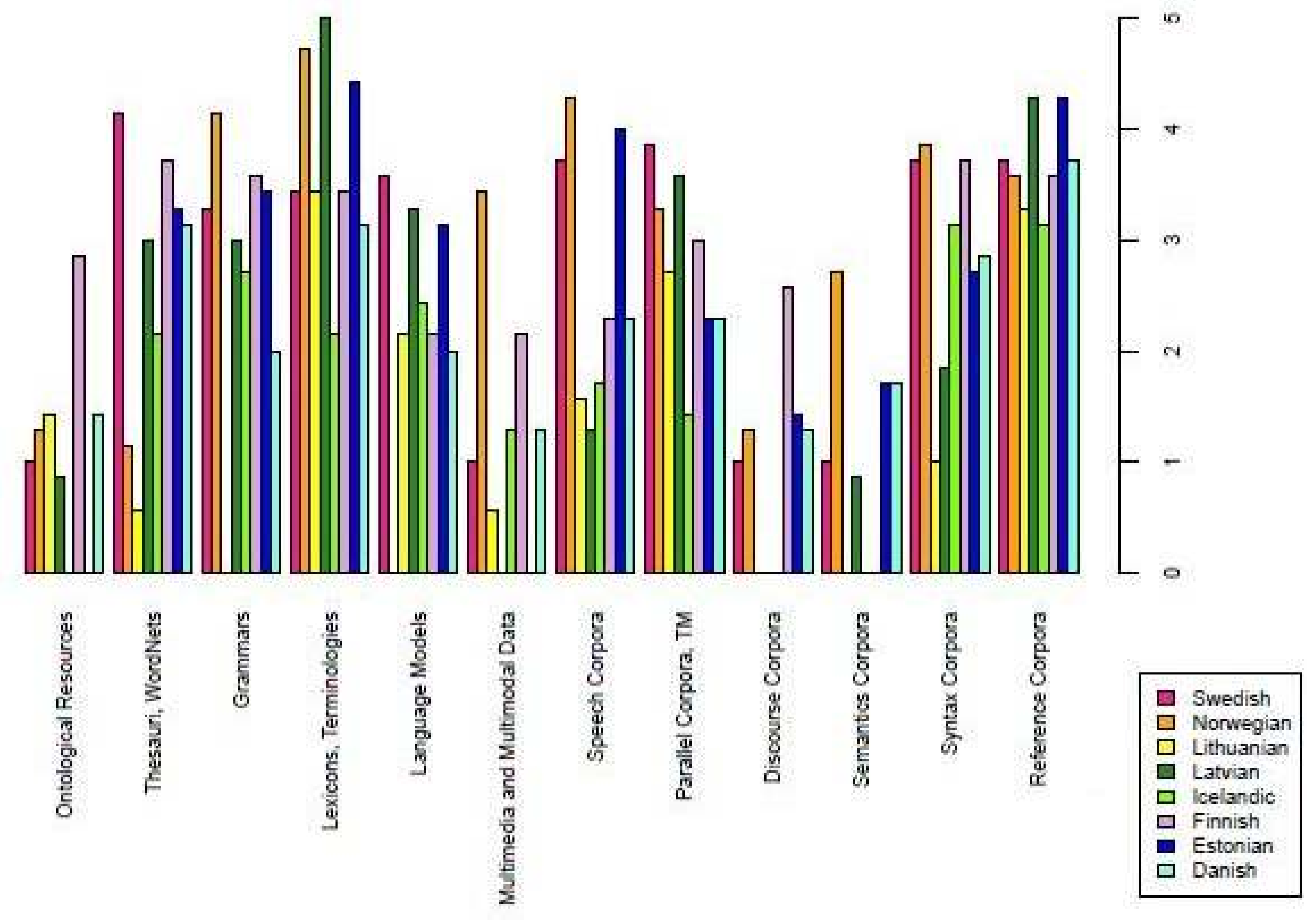
Types of resources

1. Reference Corpora
2. Syntax Corpora (treebanks)
3. Semantics Corpora
4. Discourse Corpora
5. Parallel Corpora, Translation Memories
6. Speech Corpora (raw and annotated)
7. Multimedia and Multimodal data (text data combined with audio/video)
8. Language Models
9. Lexicons, Terminologies
10. Grammars
11. Thesauri, WordNets
12. Ontological Resources for World Knowledge (e.g. upper models, linked data)

META-NET

META-NORD




5
4
3
2
1
0
Ontological Resources
Thesauri, WordNets
Grammars
Lexicons, Terminologies
Language Models
Multimedia and Multimodal Data
Speech Corpora
Parallel Corpora, TM
Discourse Corpora
Semantics Corpora
Syntax Corpora
Reference Corpora
Swedish
Norwegian
Lithuanian
Latvian
Icelandic
Finnish
Estonian
Danish

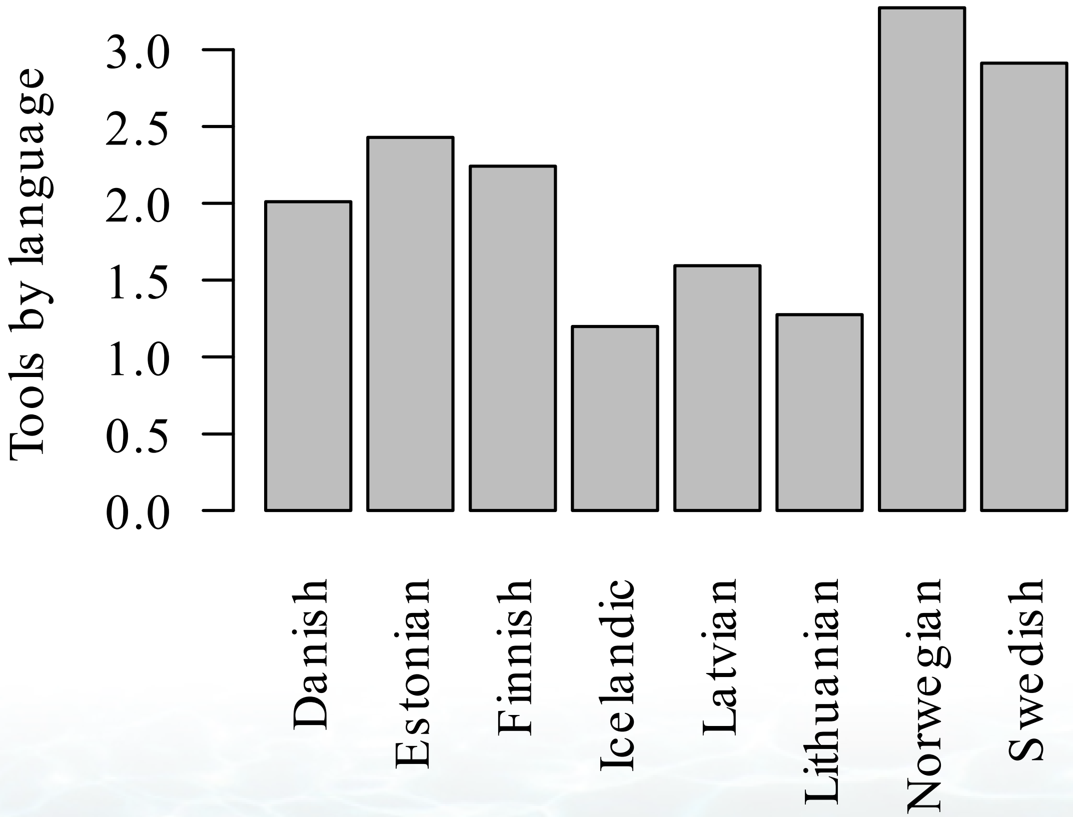
Results - basic tools and resources

- Only with respect to the most basic tools and resources such as tokenizers, PoS taggers, morphological analyzers/generators, syntactic parsers, reference corpora, and lexicons/terminologies, the situation is reasonably good for all the META-NORD languages
- All the languages seem to have some tools for information extraction, machine translation and speech recognition and synthesis, as well as resources like parallel corpora, speech corpora, and grammars, although these tools and resources are rather simple and have a limited functionality for some of the languages

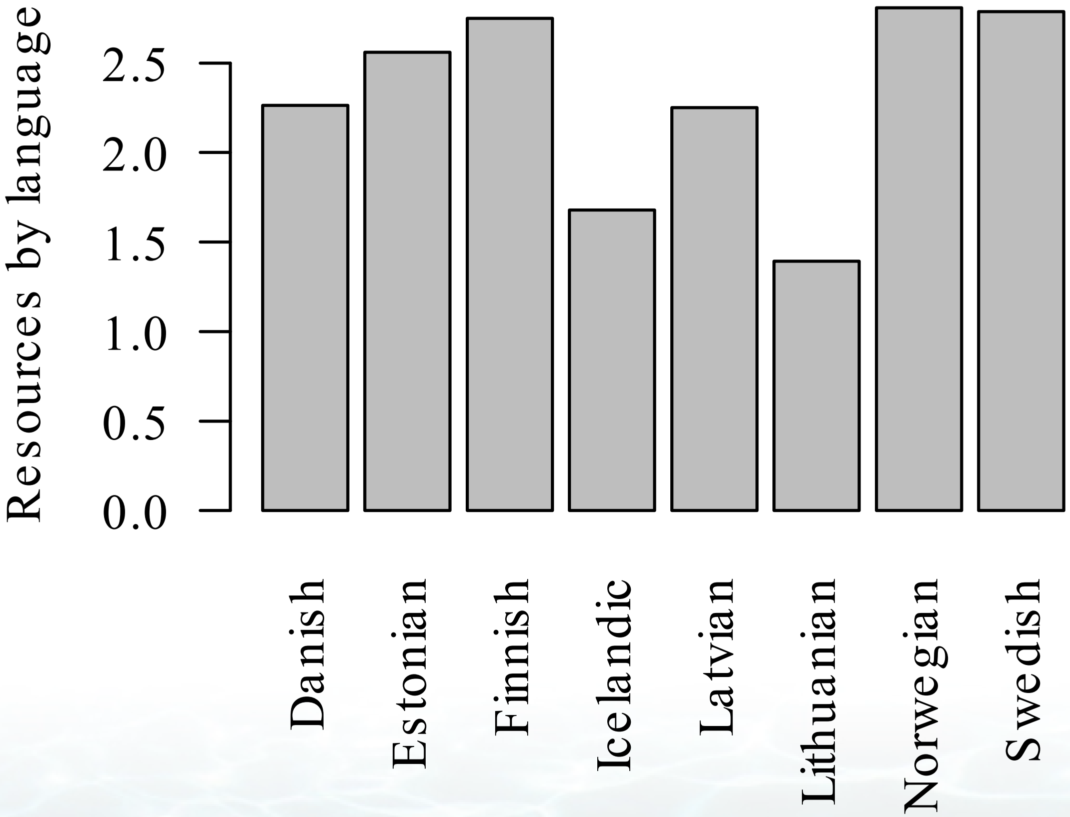
Results - advanced tools and resources

- When it comes to more advanced fields like 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 fields this completely
- For the most advanced tools and resources like discourse processing, dialogue management, semantics and discourse corpora, and ontological resources, most of the languages either have nothing of the kind or their tools and resources have a very limited scope

Tools by language



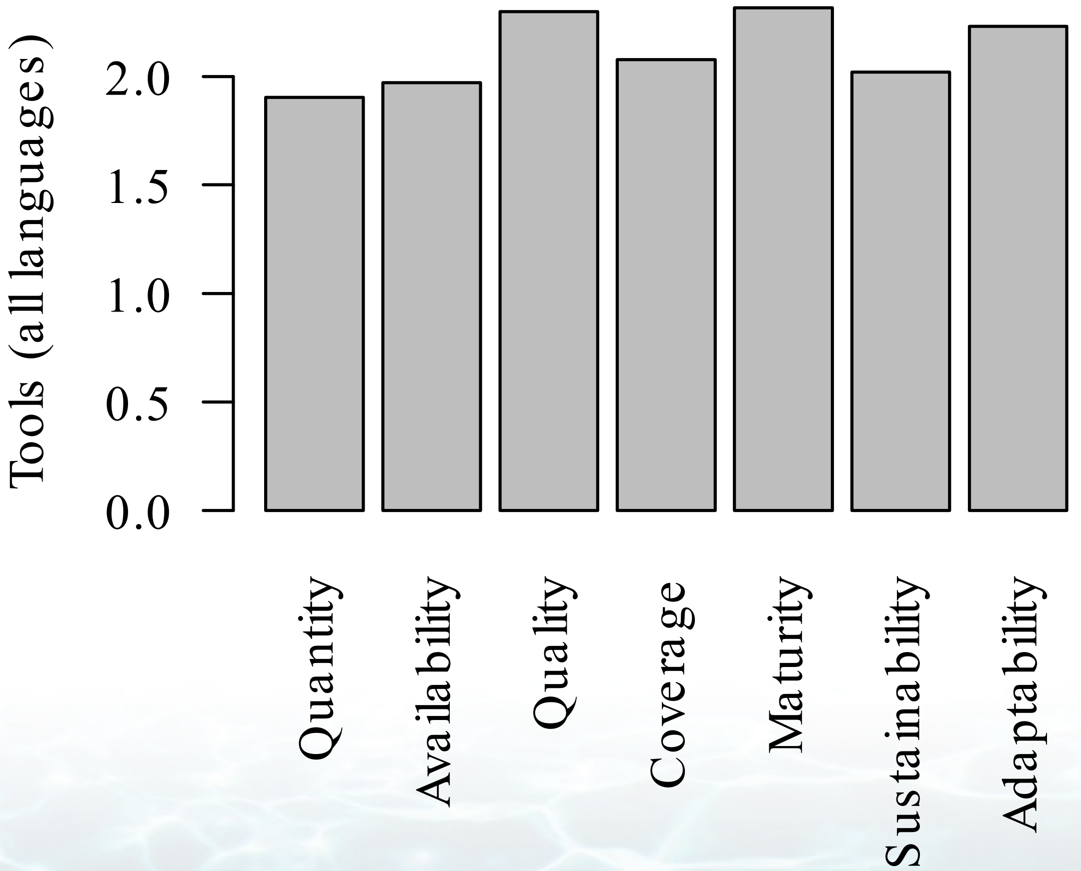
Resources by language



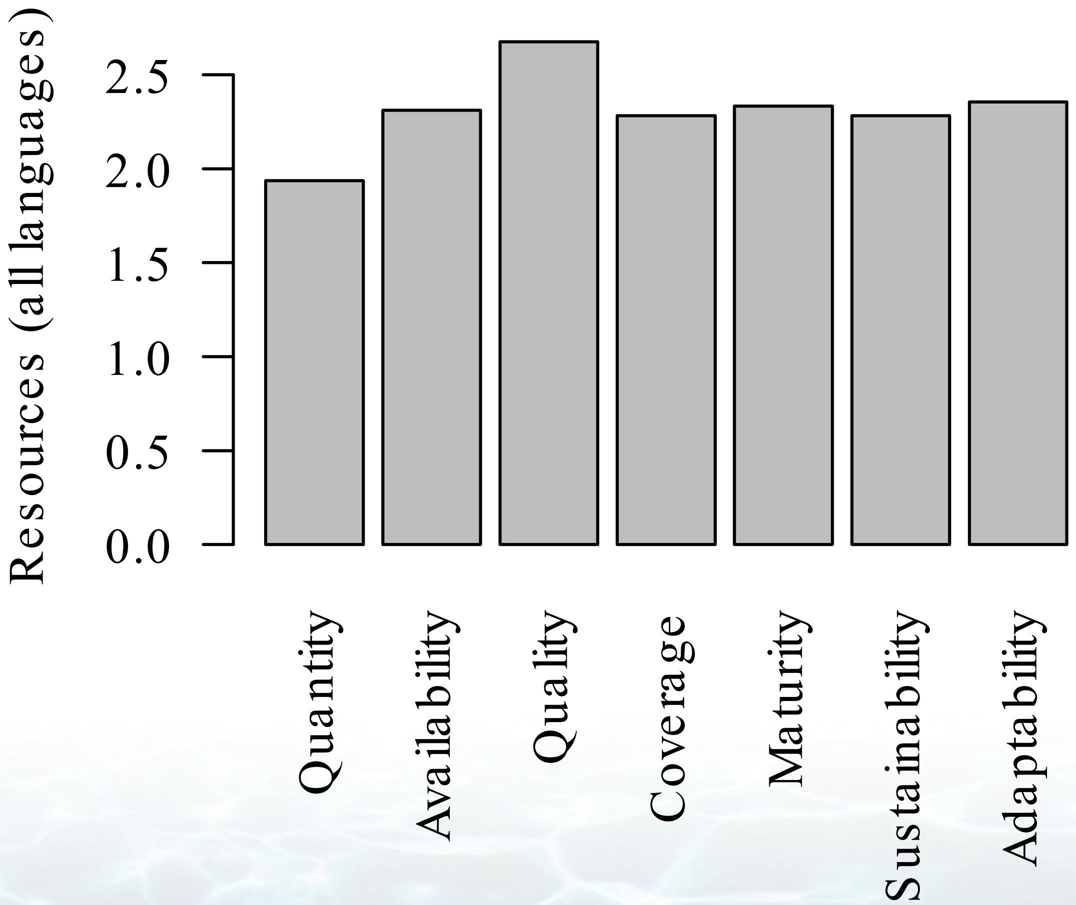
Comparison

- Clear differences between languages
- Danish and Icelandic lower than expected, Norwegian higher than expected
- Calibration needed?

Tools by criterion



Resources by criterion



Main concerns

- The means for all languages together indicate that quantity and availability may be a greater concern than quality (esp. for tools)
- The aim to improve availability though sharing is the very *raison d ´être* of the META-NORD project

Conclusion

- The Nordic and Baltic countries still have a long way to go to realize the vision of making the area a leading region in language technology
- The reports aim at finding our strengths and weaknesses and point to prospective possibilities for fruitful cooperation, in particular sharing of tools and resources, which will considerably strengthen the field in the near future