The META-NORD language reports

Koenraad De Smedt
University of Bergen
Bergen, Norway
desmedt@uib.no

Eiríkur Rögnvaldsson
Háskóli Íslands
Reykjavik, Iceland
eirikur@hi.is

Abstract

As part of the META-NORD project, the state of affairs in language technology in the Nordic and Baltic countries is being described in a set of eight reports. Each language report describes the situation of a language community and the position of the language service and language technology industry for that language. This position paper presents our methodology and preliminary findings. The final reports will be published in the META-NET series of white papers for all main languages of Europe.

1 Background

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 in order to build language enabled applications.[1] As part of this effort, the project is compiling overviews of the language service and language technology industry for all the languages targeted by the project. These languages include the main official languages spoken in the Nordic and Baltic geographical area: Danish, Estonian, Finnish, Icelandic, Latvian, Lithuanian, Norwegian and Swedish.

For most of these languages, there have been some previous surveying efforts during the past few decades, mostly in preparation of R&D programmes in language technology or for the establishment of language resources infrastructures. These overviews have had different aims and methodologies and their findings are therefore not fully comparable. In some countries, such as Norway, Sweden and Iceland, plan documents and their overviews of the state of the art have often been tied to official language policy and government propositions, whereas in other countries, such as Denmark, government branches dealing with technology and development have also contributed with stimuli towards plans and surveys.

It is not the first time that a surveying effort is launched across the whole of Northern Europe. In the aftermath of the language technology research programme financed by the Nordic Council of Ministers (2000–2005), a comprehensive report was written, known as *Vismansrapporten* (Lindén et al., 2006). This report presents an analysis of needs, opportunities and policies, identifies key areas, estimates magnitudes of R&D funding, indicates obstacles, notably aspects of rights and licensing, and presents a vision for a future embedding of language technology in the Nordic and Baltic society. *Vismansrapporten* is likely the first wide-ranging overview of the situation of language technology in this area. It was compiled by a careful analysis of documents and research budgets, as well as by a questionaire which was sent out to a large number of experts in the area, and includes literal quotes from the expert's answers to open questions.

While the usefulness of *Vismansrapporten* is recognized, the situation of language technology needs and solutions, and the constellation of technology consumers and providers, is rapidly changing, so that a new effort, five years later, is justified. As an indication of the changed situation, consider that fact that access to social media has boomed during the past five years, and in Norway, access to media content from mobile devices tripled from the beginning of 2009 to the end of 2010.[2] Also, new industrial players (especially SMEs) have emerged during the past five years, producing an increased need for contact be-

[1] See elsewhere in this volume for a more extensive overview of general aims and structure of the META-NORD project.

[2] Source: http://medienorge.uib.no/

tween industry and academia. In the same period, the Nordic Language Councils have successfully established a closer cooperation between countries about language technology through seminars and other communication, but they have not published systematic status reports.

The META-NORD reports are written as a series of separate publications for each language, but they are closely coordinated in their structure. Their data includes numerical estimates of a large number of technological aspects, compiled on the basis of the same framework that is used in the whole META-NET network.[3]

2 Aim and audience

The META-NORD reports aim at raising 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. The function of the reports is to serve as the ground for planning cooperation between the participating countries, and for identifying strengths and weaknesses to be addressed. The target audiences are therefore mainly nonexpert readers such as politicians and journalists, national funding bodies, research councils, language councils, private companies in the technology sector, and also universities and research institutions.

Each report, which is about thirty to forty pages long, is brought out in the respective language under discussion as well as in English. Similar reports are prepared by the other partner projects participating in META-NET in order to cover the main languages of Europe. It is expected that the publication of the whole series of papers in the English version will have considerable impact across Europe and may affect the conception of future language technology R&D programmes.

3 Report structure

For each of the languages, an analysis of the language community has been conducted and the role of the language in the respective country/language community is described. The language technology research community and the language service and language 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, which may differ from country to country, are outlined.

The structure of the language reports for all the META-NET languages is the same. They have three main sections. The first section, which is common to all the reports and written by experts from the DFKI (Deutsches Forschungszentrum für Künstliche Intelligenz) is entitled “A Risk for our Languages — A Challenge for Language Technology”, and is intended to explain the opportunities and challenges for language technology in the modern information society.

The remainder of each report is different for each language and written by experts on that language. It contains subsections on 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, and the role of the language on the Internet.

The reports further contain an important section on language technology support for the language in question. It contains subsections on the core application areas of language and speech technology, such as language checking, web search, speech interaction, machine translation, etc. and describes the situation in the language with respect to the application areas. Furthermore, there are language particular subsections on language technology in education and language technology programs in the country in question. The language particular parts of this section are written by experts on each language.

The reports present a detailed table with ratings 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 rate the following 13 types of tools and 12 types of resources according to these criteria for their language:

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

[3] META-NET is a Network of Excellence of which META-NORD forms a part; `http://www.meta-net.eu/`

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)
14. Reference Corpora
15. Syntax Corpora (treebanks)
16. Semantics Corpora
17. Discourse Corpora
18. Parallel Corpora, Translation Memories
19. Speech Corpora (raw and annotated)
20. Multimedia and Multimodal data (text data combined with audio/video)
21. Language Models
22. Lexicons, Terminologies
23. Grammars
24. Thesauri, WordNets
25. Ontological Resources for World Knowledge (e.g. upper models, linked data)

A preliminary results are summarized as barplots in the Appendix, where the mean value for all criteria (each rated on a scale from 0 to 6) is given for each language and each tool or resource type. The data are not finalized for all languages, as more input from experts for some language is still expected. Also, it must be taken into account that all values are based on estimates.

The results indicate that 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. Furthermore, 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.

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. 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 quite limited scope. The means for all languages together (final tables) indicate that quantity and availability may be a greater concern than quality; this need is the very *raison d´être* of the META-NORD project.

4 Discussion and conclusion

The closely parallel methodology for writing the META-NORD language reports, in coordination with all of META-NET, secures the representation of the Nordic and Baltic languages in a Europe-wide series of white papers on the status of language technology in all main national language communities.

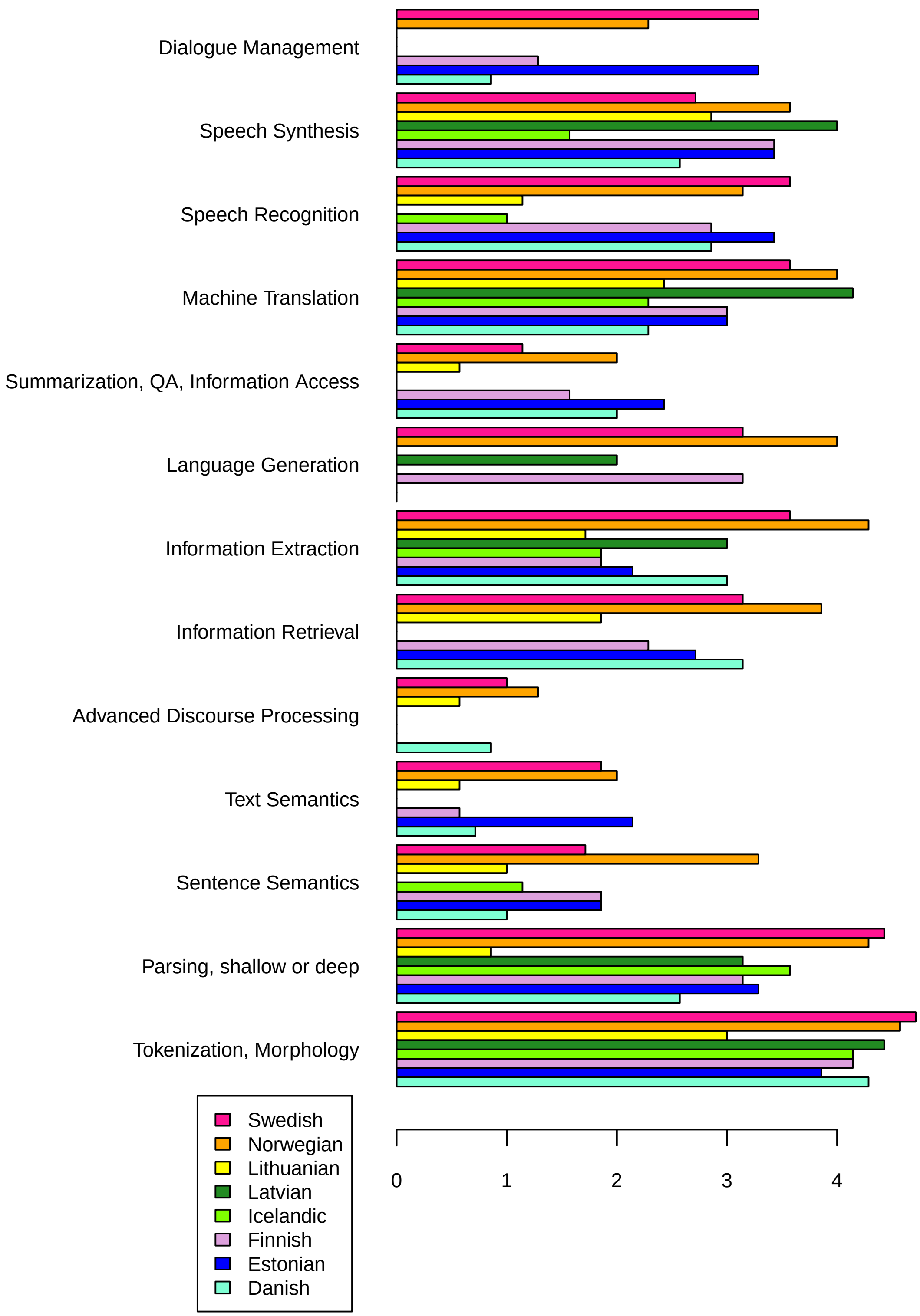
A shortcoming of the current effort is that META-NORD is focusing only on the eight main languages in its geographic area, while minority languages are not explicitly addressed. This means that the smaller Nordic languages Greenlandic, Faroese, Kven and Sami are mentioned only in passing. Also, Russian is not included, even if Northwestern Russia is a part of Northern Europe and Slavic languages are important minority languages in the Baltic countries.

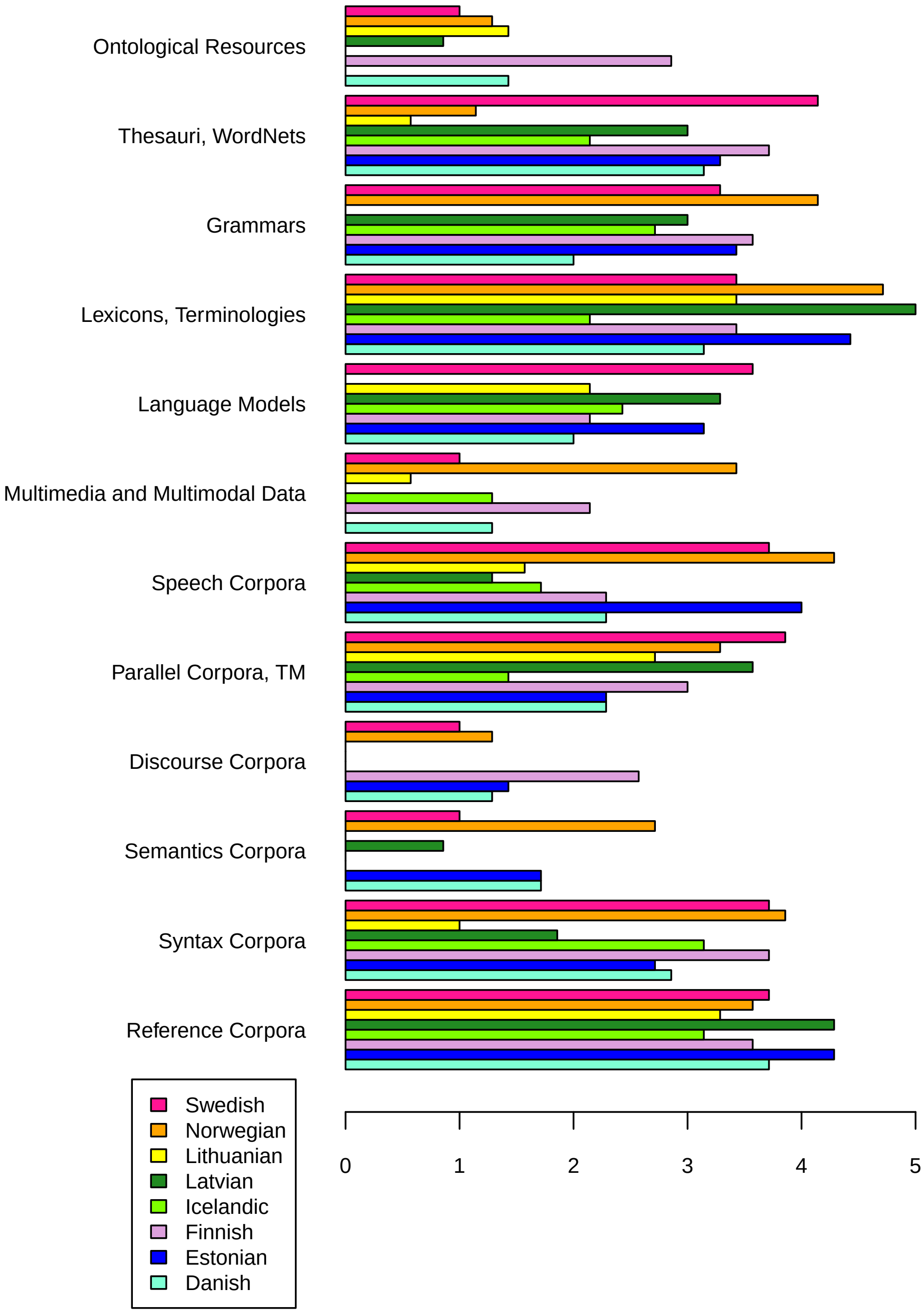
The language reports show that the Nordic and Baltic countries still have a long way to go to realize the vision of making the area a leading region in language technology, which was the aim that *Vismansrapporten* set out for 2016. However, the reports will hopefully enable us to locate our strengths 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.

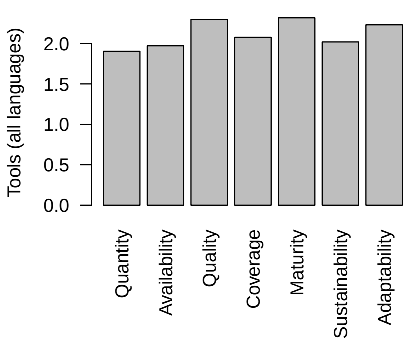
References

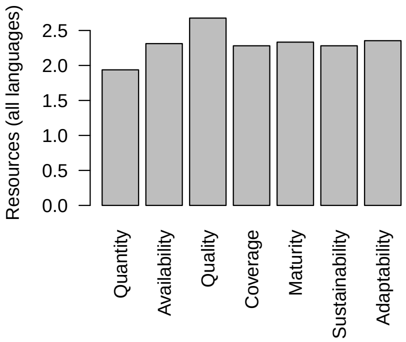
Lindén, Krister, Kimmo Koskenniemi, and Torbjørn Nordgård. 2006. Språkvis — Vismansrapport — Expert Panel Report. The Nordic Countries — A Leading Region in Language Technology. `https://kitwiki.csc.fi/twiki/bin/view/Main/LTExpertPanelBookView`.

Appendix: Barplots of the assessment of the status of tools and resources




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
0
1
2
3
4
5


Tools (all languages)
2.0
1.5
1.0
0.5
0.0
Quantity
Availability
Quality
Coverage
Maturity
Sustainability
Adaptability


Resources (all languages)
2.5
2.0
1.5
1.0
0.5
0.0
Quantity
Availability
Quality
Coverage
Maturity
Sustainability
Adaptability