META-NET
META-NORD
www.meta-nord.eu

# Newsletter

**Number 2, 2012, August-October**

## The White Papers Dissemination Campaign

*The European Day of Languages, September 26th was the perfect occasion for spreading the most recent and comprehensive information regarding LT (language technology) situation in European countries*

That day citizens of all European countries could read online or in the traditional media articles based on the press release "At Least 21 European Languages in Danger of Digital Extinction". The press release prepared by the META-NET coordinators presented the results of a new study on the assessment of language technology support for 30 European languages. The results of the study conducted by more than 200 experts from all these languages were far from being positive. According to the study about 21 European languages have weak language technology support, meaning they are in danger of digital extinction.

The main aim of the White Paper Series was to map the situation of language technology in the European languages. The study shows that the situation regarding language technologies and tools, as well as the support of the political institutions given to LT and the scope of LT industry players differs from one European language to another. There is a need thus for strategic language technology development plans that have as their cornerstone this diversity of challenges and opportunities.

The success of the press release informing about the White Paper Series, leading on the European Day of Languages to an avalanche of articles written in all the different languages on our continent, was the result of the seamless coordination of the activities by the Communication Working Group of META-NET. All in all, there were 450 responses to the press release in the printed, online and audio-visual media. The META-NORD project played an important role in the press campaign by raising the interest of the national and local media towards the White Paper Series. As a result there were approximately 70 responses to the press release in the media of the META-NORD project countries. Among these responses there were also several radio and television interviews.

You can read more about the findings of the White Paper Series in the press release included in this issue of the META-NORD newsletter. The 31 volumes of the META-NET White Paper Series covering 30 European languages can be downloaded from www.meta-net.eu/whitepapers/overview.

Dear reader,

This second issue of META-NORD newsletter is dedicated to events that took place in the META-NORD project countries during the last few months. During this time there was intense work on META-NET's strategic plans as well as on ensuring the success of a Europe-wide media campaign.

The Language White Paper Series dissemination campaign organized centrally by META-NET was a major event of this period. META-NORD has played an active role in the campaign and managed to raise awareness of the alarming situation of language technology in the Baltic and Nordic countries. You can find in this issue META-NET's Language White Paper related press release and articles about the campaign.

Another recent important event for the META-NORD project was the metadata upload of the second batch of language resources and tools to META-SHARE. The article "The second upload of language resource metadata" on page 4 gives a brief overview of this achievement.

On page 5 Andrejs Vasiljevs reports on the successful Latvian national workshop "Language, Technologies and the Future of Europe", in which participated also influential politicians and international guest from different European countries.

The last pages of this issue are dedicated to the success stories of two META-NORD consortium members. The Norwegian META-NORD team describes its fruitful collaboration with the business sector, while our colleagues from the University of Copenhagen present the advantages of the upgraded Danish lexical database. If you, dear reader, happen to have a success story you would like to share in the following issue, do not hesitate to contact us by email at the address: dovile.bieleviciute@lki.lt.

Enjoy your reading!

*Imre Bartis, University of Helsinki*
*Dovilė Bielevičiūtė, Institute of the Lithuanian Language*

# *At Least 21 European Languages in Danger of Digital Extinction*

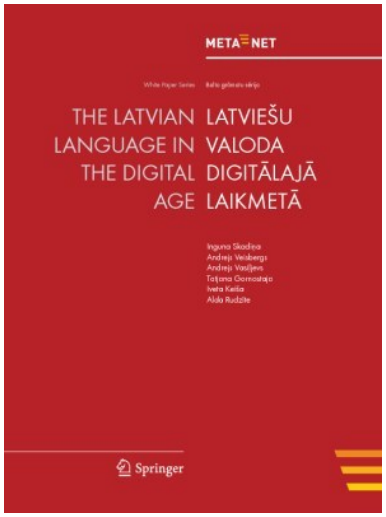
*Press release on the European Day of Languages...*

**Good News and Bad News on the European Day of Languages**

Most European languages are unlikely to survive in the digital age, a new study by Europe's leading Language Technology experts warns. Assessing the level of support through language technology for 30 of the approximately 80 European languages, the experts conclude that digital support for 21 of the 30 languages investigated is "non-existent" or "weak" at best. The study "Europe's Languages in the Digital Age" was carried out by META-NET, a European network of excellence that consists of 60 research centres in 34 countries, working on the technological foundations of multilingual Europe.

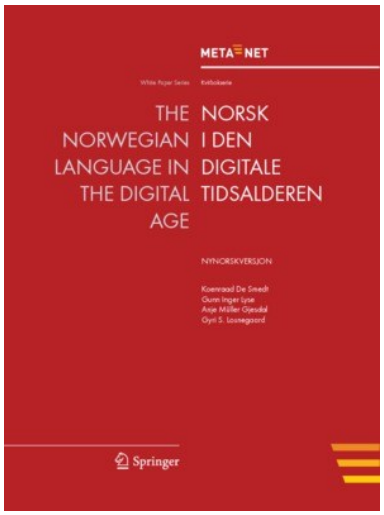
Europe must take action to prepare its languages for the digital age. They are a precious component of our cultural heritage and, as such, they deserve future-proofing. The European Day of Languages on September 26 recognises the importance of fostering and developing the rich linguistic and cultural heritage of our continent. The META-NET study shows that, in the digital age, multilingual Europe and its linguistic heritage are facing challenges but also many possibilities and opportunities.

The study, prepared by more than 200 experts and documented in 30 volumes of the META-NET White Paper Series (available both online and in print), assessed language technology support for each language in four different areas: automatic translation, speech interaction, text analysis and the availability of language resources. A total of 21 of the 30 languages (70%) were placed in the lowest category, "support is weak or non-existent" for at least one area by the experts. Several languages, for example, Icelandic, Latvian, Lithuanian and Maltese, receive this lowest score in all four areas. On the other end of the spectrum, while no language was considered to have "excellent support", only English was assessed as having "good support", followed by languages such as Dutch, French, German, Italian and Spanish with "moderate support". Languages such as Basque, Bulgarian, Catalan, Greek, Hungarian and Polish exhibit "fragmentary support", placing them also in the set of high-risk languages.





"The results of our study are most alarming. The majority of European languages are severely under-resourced and some are almost completely neglected. In this sense, many of our languages are not yet future-proof.", says Prof. Hans Uszkoreit, coordinator of META-NET, scientific director at DFKI (German Research Center for Artificial Intelligence) and, together with Dr. Georg Rehm (DFKI), co-editor of the study. Dr. Georg Rehm adds: "There are dramatic differences in language technology support between the various European languages and technology areas. The gap between 'big' and 'small' languages still keeps widening. We have to make sure that we equip all smaller and under-resourced languages




with the needed base technologies, otherwise these languages are doomed to digital extinction."

The field of language technology produces software that can process spoken or written human language. Well-known examples of language technology software include spell and grammar checkers, interactive personal assistants on smartphones (such as Siri on the iPhone), dialogue systems that work over the phone, automatic translation systems, web search engines, and synthetic voices used in car navigation systems. Today language technology systems primarily rely on statistical methods that require incredibly large amounts of written or spoken data. Especially for languages with relatively few speakers it is difficult to acquire the needed mass of data. Furthermore, statistical language technology systems have inherent limits in their quality, as can be seen, for example, in the often amusing incorrect translations produced by online machine translation systems.

Europe has succeeded in removing almost all borders between its countries. One border still exists, however, and it seems to be impenetrable: the invisible border of language barriers is one that hinders the free flow of knowledge and information. It also harms the long-term goal of establishing a single digital market because it hinders the free flow of goods, products, and services.

**Continued on the next page**

While language technology has the potential to get rid of language barriers through modern machine translation systems, the results of the META-NET study clearly show that many European languages are not yet ready. There are significant gaps in technology due to the English-language focus of most R&D, a lack of commitment and financial resources, and also a lack of a clear research and technology vision. A coordinated, large-scale effort has to be made in Europe to create the missing technologies as well as transfer technology to the majority of languages. There are strong reasons for approaching this immense challenge in a community effort involving the EU, its member states and associated countries, as well as industry: the high per-capita financial burden for smaller language communities; the needed transfer of technologies between languages; the lack of interoperability of resources, tools, and services; and the fact that linguistic borders often do not coincide with political borders.

### Language Technology: Background

Language technology already supports us in everyday tasks, such as writing e-mails or buying tickets. We benefit from language technology when searching for and translating web pages, using a word processor's spell and grammar checking features, operating our car's entertainment system or our mobile phone with spoken commands, getting recommendations in an online store, or following the instructions spoken by a mobile navigation app. In the near future, we will be able to talk to computer programs as well as machines and appliances, including the long-awaited service robots that will soon enter our homes and work places.

Wherever we are, when we need information or help, we will simply ask for it. Removing the communication barrier between people and technology will change our world.

Language technology is generally acknowledged today as one of the key growth areas in information technology. Large international corporations such as Google, Microsoft, IBM, and Nuance have invested substantially in this area. In Europe, hundreds of small and medium enterprises have specialised in certain language technology applications or services. Language technology allows people to collaborate, learn, do business, and share knowledge across language borders and independently of their computer skills.

### The META-NET White Paper Series

The META-NET White Paper series "Europe's Languages in the Digital Age" reports on the state of 30 European languages with respect to Language Technology and explains the most urgent risks and chances. The series covers all official EU Member State languages and several other languages spoken in Europe. While there have been a number of valuable and comprehensive scientific studies on certain aspects of languages and technology, until now there has been no generally understandable compendium that presents the main findings and challenges for each language with regard to a technology-supported multilingual Europe. The META-NET White Paper Series fills this gap. META-NET can now show why most languages face serious problems and pinpoint the most threatening gaps. In total, more than 200 authors and contributors helped preparing the Language White Papers.

The white papers were written for the following European languages: Basque, Bulgarian, Catalan, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, Galician, German, Greek, Hungarian, Icelandic, Irish, Italian, Latvian, Lithuanian, Maltese, Norwegian (bokmål and nynorsk), Polish, Portuguese, Romanian, Serbian, Slovak, Slovene, Spanish, and Swedish. Each Language White Paper is written in the language it reports upon and includes a complete English translation.

### About META-NET and META

META-NET, a Network of Excellence consisting of 60 research centres from 34 countries, is dedicated to building the technological foundations of a multilingual European information society. META-NET is co-funded by the European Commission through a total of four projects.

META-NET is forging META, the Multilingual Europe Technology Alliance. More than 600 organisations from 55 countries, including research centres, universities, small and medium companies as well as several big enterprises, have already joined this open technology alliance.

# *A successful media campaign in Iceland*

*The media campaign that was launched to promote the publication of the META-NET White Papers was a huge success in Iceland with 11 news stories on the subject.*

The national television covered the story in their evening news on September 26, the Day of European Languages, including an interview with META-NORD project leader in Iceland, Eiríkur Rögnvaldsson. Rögnvaldsson was also interviewed on two different programs on Iceland's second largest radio station, Bylgjan as well as by both daily newspapers in the country. The largest newspaper, Fréttablaðið, which is delivered free of charge into almost every home in the country, ran an editorial on the White Papers in addition to publishing an article by Kristinn Halldór Einarsson, the president of the Association for the blind, on the White Papers and the importance of speech-to-text systems for the blind. The second largest newspaper, Morgunblaðið, also ran a story on the White Paper Series. In addition to all this the White Papers were discussed on three different websites; that of the Ministry of Education, Science and Culture, the University of Iceland, and the online student newspaper at the University of Iceland.

The successful media campaign has ensured that the message of the White Papers has been delivered. It has reached the ears of those that have the power and the means to change things. The next step is to follow up with a coordinated effort and a focused plan towards the reachable goal that the Icelandic language be used in all aspects of the digital world.

*Kristín M. Jóhannsdóttir,*
*University of Iceland*

# *The second upload of language resource metadata*

*META-NORD's second upload of language resource metadata was a success, due to both the commitment of the consortium and the great improvements to the META-SHARE software.*

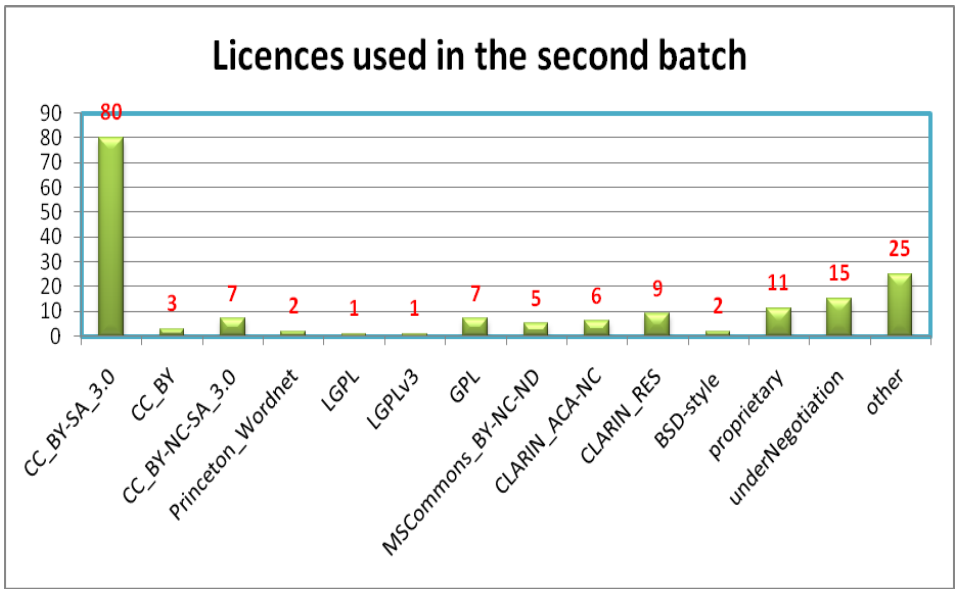
All in all metadata for 174 LR (language resources) were provided, as compared to the 127 envisioned in autumn 2011. Also the number of META-SHARE nodes increased from three to four with the release of the University of Tartu's node on 28 June 2012 (http://metashare.ut.ee/). The other nodes are Tilde's (http://metashare21.tilde.lv/), the University of Gothenburg's (http://spraakbanken.gu.se/metashare//) and the University of Helsinki's (http://metashare.csc.fi/).

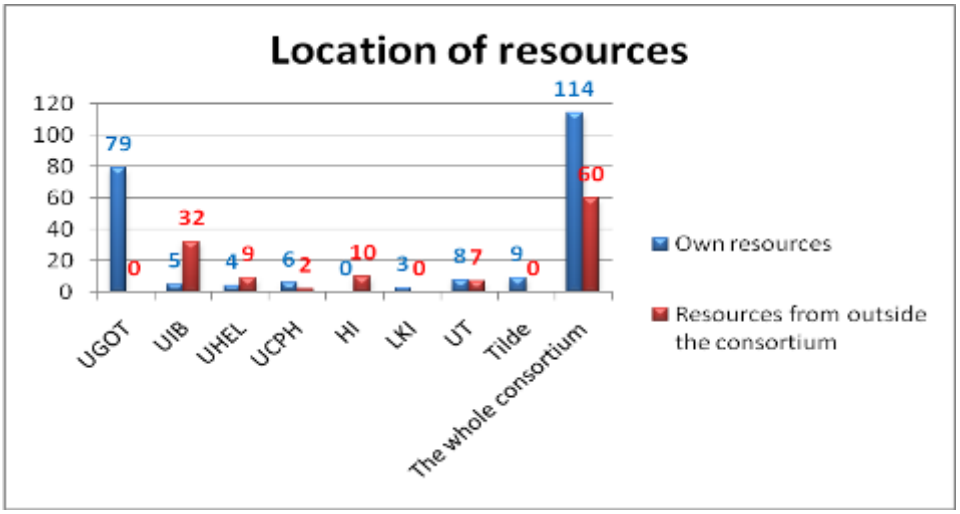
Out of the 174 LRs provided for the second upload 114 were resources of the consortium, while 60 were from outside the consortium. Most of the LR provided for the second upload were corpora (111). The other types of resources were lexical resources (6), lexical conceptual (29), corpus-ngrams (9), speech (3), tools (10), lexicons (5) and ontology (1).

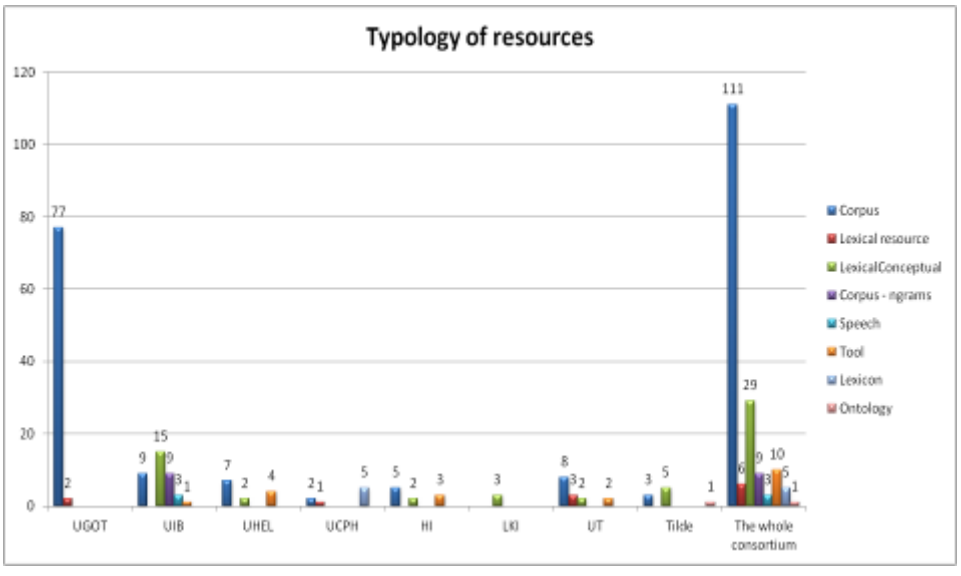
As illustrated in the diagram "Licences used in the second batch" , from the 174 resources of the second batch only 12 have licences that are still under negotiation, while the licences of the others are not an open question anymore. The most popular category of licences is CC_BY-SA_3.0, while also CC_BY, CC_BY-NC-SA_3.0 and GPL are well represented. This proves that the interest towards sharing resources in the spirit of open data is strong in the META-NORD network.

The META-NORD consortium hopes that META-SHARE will be the basis of a strong and constantly growing community keen on sharing both language resources and their respective metadata.

*Imre Bartis, University of Helsinki*



# *Riga Conference on Language, Technologies and the Future of Europe*

*A major META-NET event was organized in Riga on September 21, 2012 – the international conference Language, Technologies and the Future of Europe*

From left to right: Andrejs Vasiļjevs, Kimmo Rossi, Stelios Piperidis, Hans Uszkoreit

The conference brought together leading European and Baltic specialists and government representatives to gain an understanding and to create a vision of the future role of national languages in Europe.

LANGUAGE TECHNOLOGIES and the FUTURE of EUROPE
RIGA 2012

Valdis Dombrovskis, the Prime Minister of Latvia, in his video address to the conference participants stressed that the diversity of cultures, traditions and languages is one of the most important treasures of Europe. At the same time, language diversity creates problems to commerce and communication across Europe. Language technologies play a crucial role to overcome the language barriers. For smaller languages like Latvian keeping up with the ever increasing pace of time and technological development is crucial.

In her address Žaneta Jaunzeme-Grende, Minister of Culture of Latvia, emphasized the unique economic value of our language and culture, its role in our global competitive differentiation. Lauma Sīka, Acting State Secretary of the Ministry of Education and Science of Latvia, talked about the application of language technologies in education, communication, in helping people with disabilities. Language technologies can serve the needs of creative industries and be part of the forthcoming BIRTI research infrastructure.

Harolds Celms, Director of the EU Publications Office, and Uldis Priede from DGT Translation of the European Commission shared their experience addressing challenges of providing information in 23 official EU languages. Future instruments of European Commission to support language technology developments such as Horizon 2020 and Connecting Europe Facilities were presented by Kimmo Rossi from DG CONNECT.

Hans Uszkoreit provided a strategic view on the research and development of language technologies in Europe, outlined the strategic research agenda for the next decade and described the development of META-NET. Stelios Piperidis presented the META-SHARE platform and encouraged participants to share their language resources for the use of both researchers and developers.

**Continued on the next page**

Žaneta Jaunzeme-Grende, Minister of Culture of the Republic of Latvia

The key findings of the newly published study European Languages in the Digital Age (Language Whitepapers) were presented by its authors. Andrejs Veisbers and Inguna Skadiņa presented the Latvian Whitepaper, Kadri Vider the Estonian, and Daiva Vaišnienė the Lithuanian Whitepaper. Although there are different developments in all the Baltic countries, still these languages have less technological support and linguistic resources comparing to larger European languages. Targeted activities on both national and EU level are urgently needed to fill these gaps.

Armands Magone presented national activities in Latvia such as the Language Shore initiative that has already provided some impressive first results and attracted an interest from the global partners like Microsoft Corporation.

Kadri Vider told how Estonian research has benefited from continuous support from the long term national program in language technologies.

Lithuania has also started a large scale program Lithuanian Language in Information Society as presented by Jolanta Zabarskaitė. Several global initiatives to support language diversity and practical tools such as Microsoft Translation Hub were demonstrated by Ēriks Eglītis from Microsoft Corporation.

Presentations evoked involving discussions during the breaks

Andrejs Vasiļjevs (Tilde), Normunds Grūzītis (IMCS, University of Latvia) and Dace Baumgarte (Datorzinību Centrs) showed examples of research results and practical applications for the Latvian language.

The conference concluded with an official launch of the innovative machine translation platform LETSMT.COM which was supported by the European Commission ICT-PSP Programme.

You can watch video recordings of the conference speeches on the website ltrigaconference2012.com.

Andrejs Vasiļjevs, Tilde SIA

# *META-NORD consortium meeting in Iceland*

*The second META-NORD project consortium meeting in 2012 was organized in Reykjavik, Iceland, on August 21-22*

Project partners after the meeting

During the meeting project partners presented their progress with regards to each Work package, as well as their future plans. The discussions focused on main achievements, plans regarding the maintenance of the already existing META-SHARE nodes and installing new ones, the preparation for the third upload of language resources, as well as sustainability strategy. One of the hottest topics was the organization of the upcoming national workshops and the META-NET Language White Paper Series dissemination campaign. The consortium also decided about the coordination of the dissemination activities.

# *Upgraded Danish lexical database STO*

*Enhanced user perspectives by upgrading the Danish lexical database STO to Lexical Markup Framework in META-NORD*

The Danish lexical database STO has now been upgraded to LMF. This update has made the information of the lexicon readily accessible because the LMF format makes it much simpler for a potential user to understand it.

The morphological part of the lexicon is currently being used by various companies and institutions, and an on-line user interface that allows the user to search for the different word forms of the lemma has been used a lot for teaching. Now that we can offer STO in LMF format, there is no doubt that the morphological part of the lexicon will be even more attractive for users in the future.

UCPH aims at a better exploitation of all the different parts of the lexicon including syntax and semantics. This is another good reason for upgrading STO to Lexical Markup Framework (LMF) in the context of META-NORD. In addition, the UCPH team is planning to develop a morphological analyzer/generator that makes use of all the morphological information in STO-LMF.

The syntactic information will also be much easier to comprehend in an LMF format but it will still require a certain linguistic knowledge to be able to exploit it. With STO-LMF as an easier format to grasp, we expect that inspiring examples of use of the syntax can now be compiled and put on the STO-LMF website.

*Sussi Anni Olsen,*
*University of Kopenhagen*

**Background information**

STO is a Danish lexical database with about 80,000 entries with morphological information, 43,000 entries with syntactic information, and about 10,000 entries with rich semantic information. STO was finished in 2005 and is updated regularly according to the official Danish orthography. The data is stored in a relational database and has until now been exported for users in a comma-separated flat format for morphology and a self-defined XML-format for syntax.

# *Reaching the business community in Norway*

*On September 20 the META-NORD team at the University of Bergen was invited to give a presentation of the project at the head offices of Tansa Systems in Oslo*

Tansa is a good example of the increasing number of small and medium enterprises that are developing language technology all over Europe.

META-NORD meets Tansa Systems. From left to right: Morten Krøtø (CEO), Koenraad De Smedt, Gunn Inger Lyse, Henriette E. Berntsen, Viggo Kristensen, Kjetil Haug

Tansa's focus is a sophisticated proofing system that is used by leading publishing companies globally (for more information on this see http://www.tansasystems.com/).

Although Tansa currently has a good resource basis, they see a need for improved access to terminology resources for Norwegian, as well as to tools for text analysis. Hopefully META-NORD and CLARINO will increase the availability and visibility of resources needed by the LT industry.

META-NORD's outreach effort to the business community has been followed by a two-day workshop in Oslo, October 15-16 (http://ematch.eu/page/spraktekn-2012) where representatives from the LT industry, researchers and policy makers have met to discuss innovation in language technology.

*Anje Müller Gjesdal,*
*University of Bergen*

member of META

## *Join META!*

***We invite institutions related with or interested in the development of LT to become a part of the Multilingual Europe Technology Alliance – META.***

**Members**
**In the Nordic and Baltic countries**

Country; Organization; Location; Representative

**Denmark**
Ankiro ApS; Copenhagen; Bo Vincents
Copenhagen Business School; Copenhagen; Peter Juel Henrichsen
Dansk Sprognævn; Copenhagen; Sabine Kirchmeier-Andersen
GramTrams; Viby J; Eckhard Bick
Ordbogen; Odense; Peter Revsbech
RAP; København, Son; Keld Simonsen
Society for Danish Language and Literature; Copenhagen; Jørg Asmussen
Technical University of Denmark; Lyngby; Jan Larsen
The National Museum of Denmark; Copenhagen; Birgit Rønne
The Royal Library; Copenhagen; Anders Conrad
University of Copenhagen, Centre for Language Technology; Copenhagen; Bente Maegaard, Bolette Sandford Pedersen
University of Southern Denmark, Faculty of Humanities; Kolding; Johannes Wagner
Wind Kommunikation; Copenhagen; Jørgen Christian Wind Nielsen

**Estonia**
University of Tartu, Institute of Computer Science; Tartu; Tiit Roosmaa

**Finland**
Aalto University, Computational Cognitive Systems Research Group; Aalto; Timo Honkela
CSC, the Finnish IT Center for Science; Espoo; Antti Pursula
Lingsoft Inc; Helsinki; Juhani Reiman
Sunda Systems Oy; Helsinki; Markku Kiiski
The Research Institute for the Language of Finland; Helsinki; Toni Suutari
University of Helsinki; Helsinki; Martti Vainio
University of Helsinki, Department of General Linguistics; Helsinki; Kimmo Koskenniemi
University of Joensuu; Joensuu; Jussi Niemi
University of Oulu; Oulu; Marketta Harju-Autti
University of Tampere; Tampere; Eero Sormunen

**Iceland**
CLARA; Reykjavik; Jon Edvald Vignisson
HEILAHEILL; Reykjavik; Þórir Steingrímsson
Heyrnarhjálp; Reykjavik; Sigurjón Einarsson
IceStat; Reykjavik; Snorri Gudmundsson
Máltæknisetur (Icelandic Centre for Language Technology); Reykjavik; Eiríkur Rögnvaldsson
Reykjavik University; Björn Þór Jónsson
Talþjálfun Reykjavíkur; Reykjavik; Þóra Sæunn Úlfsdóttir
The Árni Magnússon Institute for Icelandic Studies; Reykjavik; Sigrún Helgadóttir
University of Iceland, School of Humanities; Reykjavik; Eiríkur Rögnvaldsson

**Latvia**
ACCURAT; Riga; Aivars Berzins
Let´s MT; Riga; Artūrs Vasiļevskis
Tilde; Riga; Andrejs Vasiljevs
University of Latvia, Institute of Mathematics and Computer Science; Riga; Inguna Skadiņa

**Lithuania**
Institute of the Lithuanian Language; Vilnius; Jolanta Zabarskaitė
Vytautas Magnus University, Center of Computational Linguistics; Kaunas; Rūta Marcinkevičienė

**Sweden**
ESTeam AB; Älgarås; Gudrun Magnusdóttir
Interverbum Technology; Linköping; Ioannis Iakovidis
KTH; Stockholm; Joakim Gustafson
KTH Royal Institute of Technology; Stockholm; Rolf Carlson
Linköping University; Linköping; Lars Ahrenberg
Lund University; Lund; Sven Strömqvist
MOLTO; Gothenburg; Emilia Rung
Swedish Institute of Computer Science; Kista; Jussi Karlgren
The Language Council of Sweden; Stockholm; Rickard Domeij
Transmachina AB; Stockholm; Nicholas Cottrell
Umeå University; Umeå; Patrik Svensson
University of Gothenburg, Department of Swedish Language; Gothenburg; Lars Borin
University of Gothenburg, Department of Swedish Language, CLT Dialogue Technology Lab; Gothenburg; Staffan Larsson
Uppsala University; Uppsala; Joakim Nivre

**Norway**
Clapter Create; Stavanger; Tor Einar Enne
Clue Norge ASA; Oslo; Merete Kravik
CognIT a.s; Oslo; Harald Falsen
Comperio; Oslo; Trond Renshusløkken
Computas; Lysaker; David Norheim
Dictatr; Bergen; Thomas Hagen
ESIS Norge AS; Oslo; Robert HP Engels
Include; Bergen; Sverre Andreas Hilditch Holbye
Kaldera språkteknologi AS; Oslo; Lars Nygaard
Lingit; Trondheim; Torbjørn Nordgård
MediaLT; Oslo; Magne Lunde
NAV The Norwegian Labour and Welfare Administration; Oslo; Daniel Scheidegger
NHH Norwegian School of Economics; Bergen; Gisle Andersen
Narvik Universit College; Narvik; Bernt Bremdal
National library of Norway; Oslo; Oddrun Pauline Ohren
Norwegian School of Economics and Business Administration; Bergen; Gisle Andersen
Norwegian University of Science and Technology; Trondheim; Torbjørn Svendsen
Nynodata AS; Bø i Telemark; Bjørn Seljebotn
Opera Software; Oslo; Pål Eivind Nacobsen Nes
SINTEF; Oslo; Diana Santos
Språkrådet; Oslo; Torbjoerg Breivik
Standards Norway; Lysaker; Marit Sæter
Tansa Systems AS; Oslo; Henriette Edvarda Berntsen
Tekstlaboratoriet, Universitetet i Oslo; Oslo; Kristin Hagen
TextUrgy; Bergen; Trond Walker
The Language Council of Norway; Oslo; Torbjoerg Breivik
The National Library of Norway; Oslo; Kristin Bakken
Tobii Technology Norge AS; Bergen; Morten Mjelde
Unifob AS; Bergen; Eli Hagen
Universitetet i Oslo; Oslo; Stephan Oepen
University of Tromsø, Det humanistiske fakultet; Trond Trosterud
University of Bergen, Department of Linguistic; Bergen; Koenraad De Smedt
Western Norway Research Institute; Sogndal; Rajendra Akerkar

The META-NORD project has received funding from the Europan Commission through the ICT PSP Programme, grant agreement no 270899.