Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Forschungsbereich Sprachtechnologie Leiter:
Prof. Dr. Hans Uszkoreit
Wissenschaftlicher Direktor am DFKI

DFKI GmbH
Projektbüro Berlin
Alt-Moabit 91c
10559 Berlin

Telefon: +49 (0)30 23895-1811
Telefax: +49 (0)30 23895-1810
E-Mail: office@meta-net.eu
Hans.Uszkoreit@dfki.de
Internet: www.dfki.de

09.11.2012

META-NET – Open letter to Europe's national institutions for language and other public organisations with large text resources

Dear colleague,

The European Network of Excellence META-NET consisting of 60 top-level research centres in 34 countries and co-funded by the European Commission through several projects, has established a collaborative infrastructure for sharing language resources for research. This infrastructure, META-SHARE, is currently being tested and will be publicly launched in January 2013.

Language resources are an indispensable precondition for progress in language technologies. This progress is urgently needed for the preservation of the linguistic diversity of our multilingual European society and for the creation of future generations of IT, which will be capable of dealing with human language and knowledge. For details on the need for more research on language technologies and on the role of language resources for this research, please refer to the appendix.

META-NET is currently conducting a drive for better support of research through language resources. As the creator and/or owner of valuable language resources, we ask you whether you would consider to making these resources available to the research community through META-SHARE.

We would also like to invite you to respond to the following questions:

- Would you support making your resources available to the scientific community for research purposes?
- Would you be able to make your resources available to research organisations through META-SHARE if META-NET gives you the guarantee that data will only be released to parties who agree to certain restrictions (such as solely using them for research or solely using them for non-commercial research)?
- Should you be already engaged in an effort to make your data available for research purposes, can we as META-NET help in any way?
- Do your agreements with the data providers permit any research use of the data?
- If they do, do they also allow you to pass on the data (under certain conditions) to other organisations for research purposes?
- If not, have you tried anything to make the data sets available to the scientific community for research purposes? Is it possible to check the agreements with the data providers again?
- What are the main stumbling blocks? We would be happy to provide free legal help in the form of the expertise of our legal consultants!
- From your point of view, what could or needs to be done to provide full access to your data sets for non-commercial research purposes?
- Could there even be a chance to provide full access to your data sets for commercial research purposes? What would need to be done?

Please take the time to answer these questions – ideally until November 19, 2012 – so that we can broaden the base of available research resources and also better understand which unsolved problems need to be attacked in order to ultimately make all important existing resources available to the research community. Thank you!

Kind regards

Prof. Dr. Hans Uszkoreit
META-NET Coordinator

Dr. Georg Rehm
META-NET Network Manager

Appendix 1: List of recipients of this open letter

- Austria: Austrian Federal Ministry for Education, Arts and Culture, Muriel Warga-Fallenböck
- Austria: Österreichisches Sprachen-Kompetenz-Zentrum, Ulrike Haslinger
- Belgium: Fédération Wallonie-Bruxelles (Wallonia Brussels Federation), Martine Garsou
- Belgium: Nederlandse Taalunie (Durch Language Union), Johan Van Hoorde
- Bulgaria: Институт за български език (Institute for Bulgarian Language, Bulgarian Academy of Sciences), Mariyana Tsibranska
- Czech Republic: Ústav Českého národního korpusu Filozofické fakulty Univerzity Karlovy (Inst. of the Czech National Corpus, Faculty of Arts, Charles University), František Čermák
- Croatia: Institute of Croatian Language and Linguistics, Željko Jozić
- Croatia: Institute of Linguistics, University of Zagreb, Marko Tadić
- Denmark: Dansk Sprognævn, Sabine Kirchmeier-Andersen
- Denmark: Det Danske Sprog- og Litteraturselskab, Lars Trap-Jensen
- Estonia: Eesti Keele Instituut, Urmas Sutrop
- Estonia: Eesti keelenõkogu, Birute Klaas
- Finland: CSC — IT Center for Science Ltd, Kimmo Koski
- Finland: Kotimaisten kielten keskus, Institutet för de inhemska språken, Pirkko Nuolujärvi
- France: Délégation générale à la langue française et aux langues de France, Xavier North
- Germany: Institut für Deutsche Sprache, Ludwig M. Eichinger
- Germany: Deutsche Akademie für Sprache und Dichtung, Peter Eisenberg
- Greece: Institute for Language and Speech Processing, Yannis Ioannidis
- Greece: ΚΕΝΤΡΟ ΕΛΛΗΝΙΚΗΣ ΓΛΩΣΣΑΣ (ΚΕΓ) (Kentro Ellenikis Glossas), Maria Theodoropoulou
- Hungary: Magyar Tudományos Akadémia Nyelvtudományi Intézet, István Kenesei
- Iceland: Íslensk málnefnd, Gudrún Kvaran
- Iceland: Árni Magnússon Institute, Guðrún Nordal
- Ireland: Foras na Gaeilge, Seán O' Cearnaigh
- Italy: Accademia della Crusca, Nicoletta Maraschio
- Italy: CNR Opera del Vocabolario Italiano, Pietro Beltrami
- Latvia: Latviešu valodas institūts, Ina Druviete
- Latvia: Latviešu valodas aģentūra, Jānis Valdmanis
- Latvia: University of Latvia, Inst. of Math. and Computer Science, Rihards Balodis-Bolužs
- Lithuania: Lietuvių kalbos institutas, Jolanta Zabarskaitė
- Lithuania: Vytautas Magnus University, Centre of Computational Linguistics, Andrius Utka
- Luxembourg: Institut Grand-Ducal, Guy Berg
- Luxembourg: Conseil permanent de la lanuge luxembourgeoise, Ralph Fichtner
- Malta: Il-Kunsill Nazzjonali tal-Ilsien Malti, Manwel Mifsud
- Norway: Språkrådet, Arnfinn Muruvik Vonen
- Netherlands: Nederlandse Taalunie, Johan Van Hoorde
- Poland: Rada Języka Polskiego przy Prezydium Polskiej Akademii Nauk, Anna Dąbrowska
- Poland: Polish Academy of Sciences, Institute of Computer Science, Jacek Koronacki
- Portugal: Camões Instituto, I.P., Ana Paula Laborinho
- Romania: Academia Româna, Institutul de Lingvistica, Marius Sala
- Serbia: University of Belgrade, Faculty of Mathematics, Duško Vitas
- Slovakia: Ľudovít Štúr Institute of Linguistics, Pavol Zigo
- Slovenia: Služba za slovenski jezik, Ministrstvo za izobraževanje, znanost, kulturo in šport, dr. Simona Bergoč
- Spain: Real Academia Española, José Manuel Blecua
- Sweden: Språkrådet, Jennie Spetz
- Sweden: Språkbanken & Centre for Language Technology: Lars Borin
- United Kingdom: British Council, Michael Carrier
- United Kingdom: Oxford English Dictionary: John Simpson
- European Patent Office: Benoît Battistelli, Oswald Schröder

Appendix 2: Background information

In the past decades, Europe has successfully managed to eradicate many barriers, most notably the financial barrier with the introduction of a mutual currency and also political barriers that prevented free movement among its citizens with the Schengen agreement. One specific barrier still remains, however: the language barrier has a significant effect on Europe and is hindering the free flow of knowledge, thought, debates, goods, products, ideas and innovations. The language barrier seems an almost insurmountable stumbling block with regard to the goal of establishing a single digital market and providing Europe's citizens the means to communicate with one another.

In its founding documents, the European Union reserved a special status for its many different languages, it realised that our languages are an important part of our culture and identity that must be preserved. META-NET is a European network of Excellence that aims to bring about a truly multilingual Europe, supported by sophisticated language technologies, especially crosslingual and multilingual technologies such as document search across many different languages and automatic translation. No one is able to learn either all or a significant subset of Europe's more than 80 languages. This is why modern, precise, sophisticated, high-quality language technologies are urgently needed to tear down the language borders that significantly hinder Europe.

However, the current pace of progress in this area is way too slow. Historically Europe has always been in a leading position in the field of language especially multilingual technologies. Very large enterprises from the US and also Asia have been developing and deploying successful and popular products and online services that go in the right direction. As a continent with a population of more than 500.000.000, we cannot, however, base our future information and communication technology infrastructure on technologies developed and controlled abroad. What Europe needs is language technology *for* Europe developed *by* Europe.

For this ambitious goal to be realised all stakeholders involved have to team up – otherwise we could as well declare defeat and let other continents and companies take the lead role in this important and increasingly popular new market. This heterogeneous group of stakeholders includes researchers in academia, research centres and industry, the user and provider industries of language technologies, funding agencies, politicians, officials, journalists and the many different language communities. Only if we act together and firmly stand behind this important goal will we be able to establish a truly multilingual Europe that is necessarily supported through language technologies.

Our EC-funded Network of Excellence META-NET prepared a study on the level of support that language technology provides for 30 European languages. The preparation of the study took more than two years and resulted in 30 volumes of our White Paper Series "Europe's Languages in the Digital Age". More than 200 experts from academia and industry contributed as authors to this immense undertaking (PDF versions are available free of charge on the META-NET website). The study clearly shows that at least 21 European Languages are in danger of digital extinction. The corresponding press campaign in late September 2012 proved that Europe is very much passionate about the language topic: our study generated more than 500 mentions in the international press as well as dozens of radio interviews and television reports.

In our follow-up document – META-NET Strategic Research Agenda for Multilingual Europe 2020 – we provide a detailed plan as regards how to bring about a future European information society, which is based on multilingual technologies. Such technologies not only support multilingual Europe but also present huge economic and social opportunities for Europe. The plan centres around three priority research themes: *Translation Cloud*, *Social Intelligence and E-Participation* and *Socially Aware Interactive Assistant*. This document is the result of dozens of discussions with hundreds of experts from academic and industrial research.

An important cornerstone of language technologies are language resources, such as, for example, very large collections of text documents that were annotated with linguistic information. These

corpora are needed to train language models for automatic processes based on statistical machine learning methods.

Our META-NET White Paper Series shows that many European languages are seriously and dangerously under-resourced. However, almost every European country has a national institution for language. In addition, almost every national institution for language has a very large corpus with texts of the respective language. These texts are typically provided by organisations such as publishing houses that transfer texts to the national institution for language on a regular basis. Usually these corpora are used in-house for linguistic research purposes. These corpora are also an extremely important resource for language technologies but full access to the corpora – to the full datasets – is in almost all cases impossible. The reasons stated usually involve legal restrictions that prohibit download or transfer of the data to third parties, as agreed with the data provider.

With our open META-SHARE exchange infrastructure for language resources we have now established a secure distributed network through which we can guarantee that specific language resources are only distributed to those persons who fulfil all requirements. For example, if a certain resource is only to be made available for non-commercial research purposes, the user has to be logged in and explicitly accept the respective terms of service and license agreement. From the technical and legal point of view this process of providing, accessing and using corpora is water-proof and fail-safe. META-TRUST AISBL is a non-profit organisation under Belgian law that acts as META-NET's and META-SHARE's legal entity.

From many discussions with representatives of Europe's national institutions for language we know that many national institutions for language would like to free up their important data collections and to make them available for research. Especially now that several US and Asian companies are in the process of further developing their technologies and leaving Europe's language technology community in their wake, our European language technology research depends on these national corpora. We as META-NET would like to help freeing up and making these data sets available.

Europe needs to take a stand and act together in order to help its individual languages and language technology communities to develop sophisticated technologies. We're confident that we can count on the support of all stakeholders to avoid being overtaken in this field by big multinational enterprises from the US and Asia. We cannot leave the important and highly sensitive issue of the future of our languages and our ability to communicate across language borders through technologies to a small number of global internet companies.

Further information on META-NET, META-SHARE, META-TRUST and the META-NET White Paper Series are available online at http://www.meta-net.eu.