Meeting of the Chairpersons of the Committees on Education, Science and Culture and the Committees on the Development of Information Society "United in Diversity: Political and Social Development Aspects of EU Languages and Cultures"

Your Excellencies,
Ladies and gentlemen,

I am very honoured to be allowed to address you on this very special occasion.

I am standing here speaking to you,

- not in the language of our hosts, Lithuanian,
- not in the language of my country, which is Danish,
- not in my mother tongue, which is German,

but in the most powerful *lingua franca* of our time: English.

Having a universal language that can be used and understood by everyone is a great gift and an enormous advantage. It speeds up direct communication, it is an easy and practical way of giving information - and not least – English is a great language.

However, the use of English or any other language as a *lingua franca,* though seemingly time and cost saving, comes at price - and a high price if we do not take care.

Every language is closely connected to the people who use it, their environment, their culture, and most of all their common history.

The way we view the world, our cognition, our understanding, is deeply rooted in the structure of our language itself.

Languages are closely linked to our identity, both as individuals and as members of a social group or a whole country.

Most of our cultural heritage is encoded in our languages. Words, phrases, idiomatic expressions and proverbs are usually best understood with reference to events in history, great literature or the words of great men and women.

And what happened in the past is also a vivid part of our present. It even influences the language of our technological world today. Who, for instance, has any idea how the the name of the *Bluetooth* technology came about? It is a piece of technology that is in use almost everywhere in the world.

For instance, the devices you have in your pocket are enabled by *Bluetooth* to connect at a close range. Probably, only Danes have fun when they hear the name *Bluetooth*. It refers to the famous Danish Viking King who led Denmark into Christianity. That is why the logo of *Bluetooth* is written in one of the oldest writing systems in the northern hemisphere - runes.

Every translator and interpreter knows about the intriguing differences between languages. During their training and through their experience they have learned how to handle the problems caused by these differences between the languages, and this is why a good translation is a work of art.

Learning a language makes you aware that things can be seen from more than one perspective. And the ability to see the world from a different angle lies at the core of innovative power.

I do not believe that it is a coincidence that the boost of inventions and ideas that started in the 19th century and continued way into the 20th century started exactly at the same time as Latin ceased to be the common language of

science. From then on, each scientist started to describe the world in his own language and from his own local innovative perspective.

Learning a language, however, also makes you aware that people are different, and that they hold different views of the world, and it makes you aware that these differences can be overcome. And it gives you the pleasure to fully appreciate the variety of cultural experiences that a country and their people can offer.

This is true for human relations and for politics but not least for trade and commerce where it has been shown time and again that the ability to approach a customer in his own language most certainly is the primary key to success.

I hope by now that I have convinced you to agree with me that our languages, no matter how small and rare they may be, are our treasures and that the loss we would suffer if we gave them up is immeasurable.

There are good reasons to preserve them, not only as historical constructs preserving the heritage of our countries but also as vivid, creative and intelligent means to express what we think and how we see the world today.

Today, every modern means of communication has to face language issues. The internet, our computers, our mobile phones, our music players, our TV, our cars - even our refrigerators are beginning to speak to us and to accept spoken input instead of using buttons that have to be pressed. And every user of these devices expects, of course, that this could be done in his or her own language.

So – how many of you can speak in your own language to your mobile phone?

- I can't.

And how many of you can speak in your own language to your mobile phone and be properly understood?

- Even fewer!

Today, machine translation services are available on almost all devices, and the number of language pairs that are offered for automatic translation is increasing even as we speak.

Today, information systems adapt automatically not only to the choices that we make, but also to the content of the words we write in emails, on *Facebook* or on *Twitter*.

Private and public institutions in all countries are shifting to servicing the citizens on the internet rather than relying on surface mail and personal contacts. They too are faced with language issues as they must accommodate linguistic minorities within the country as well as immigrant languages in order to ensure the basic human rights of their citizens.

We all know that the costs for maintaining spoken and written services in the bodies of the European Union in all languages are immense.

Voices are heard from time to time that it would be much easier to switch everything into English. All the problems of translation and interpretation would simply go away and lots of money would be saved.

I hope that I have made it clear in the beginning of my talk that this is a poor vision. It would mean that everything would converge towards one *lingua franca* and that diversity and creativity would decrease. We would be using one language for all the important stuff and for our communication devices, and another for casual communication at home.

Already now, languages with few speakers are poorly served, because big companies such as *Apple* and *Microsoft* do not consider them to be attractive markets. I recently heard from the Home Rule of Greenland that they had developed a spell checker for Greenlandic and offered it to *Microsoft* as a plug-in for *Word* – completely free of charge. But *Microsoft* was not even interested.

I have a vision that one day I will be able to speak the language of my country, however small it might be, and all of you out there, regardless how small your languages might be, will receive a perfect interpretation through the headset of your mobile phone.

I have a vision that companies all over Europe will be able to access emerging markets in other parts of the world without having to worry about language barriers, and that in all countries new innovative language companies will thrive in providing new language services or adapting existing ones for their own language.

I hope that one day information can flow freely between languages to the benefit of us all, ensuring better understanding between people.

Having worked in the field of language technology for 27 years (since 1986), I know that we are getting closer to this vision. Being in charge of an official institution for language and language policy of a small country, I also know that leaving the care for our languages to market forces is absolutely counterproductive. It has not and will not ensure diversity and multilingual coverage. It will only ensure survival of the biggest.

If we want to have the technology that can overcome language barriers for our countries we have to make sure that we have our own resources and our own expertise, and that we stimulate language industry nationally as well as internationally.

In order to find out how this can be done, one has to understand one thing about language technology – it develops only if it has access to language resources such as electronic documents and recorded speech. And lots of it.

Modern language technology systems are trained on millions of words and sounds of a language, and from that they extract language models which are then applied to process new text or spoken input. This is both an enormous advantage and a huge limitation. It means that one can use the same type of technology for many different languages. But it also means that the systems only know the language that they were trained on, and that they will perform poorly if they are to cope with a new domain or with a language where the access to digital documents and spoken resources are limited.

This is why we see that *Google Translate* sometimes provides surprisingly good translations in the general domain for language pairs with big presence on the internet, whereas the translations become more and more ridiculous when we are dealing with more specialised domains or language pairs where the internet presence is more scarce.

So, the best way of speeding up the development of language technology for any language is to make sure that researchers and developers have free access to lots and lots of digitalised texts and sound recordings of that language. But only few politicians and government officials know this.

Let me give you an example from my own country.

The municipality of Odense, the birthplace of Hans Christian Andersen, decided last year to invest about 8 million Euros in a speech-to-text system that would enable more than 2000 employees in the public administration to speak to their computer instead of typing reports and letters to the citizens. Together with one of the largest international providers in the field, they presented a business case that was adopted by the city council.

When the system was delivered, it turned out that it failed to recognise most of the domain-specific expressions in the field of employment, health care, child care and other important areas. After only a few weeks most of the employees had given up on the system because it took more time to correct the errors than to type everything from scratch.

What had gone wrong?

First of all, the system had not been trained on texts and speech samples from the domains that were relevant for the municipality. Such texts and recordings were not readily available and therefore, the company thought it was too expensive to train the system properly.

Secondly, the company had no expertise in Danish, and the municipality had not consulted any Danish language technology experts, who could have told them that the project was endangered from the very start.

Of course the company was very sorry and promised to improve the system. For this they needed access to lots of texts from the municipality's archives in the various domains as well as more recorded speech – which was much harder to get. But worst of all, the company would then keep the resources and all the processing that had gone into it, making it practically impossible for the municipality to change to another provider or to launch a bidding round in the future.

This is pretty bad, but the story gets even worse, since 11 other municipalities in Denmark are currently in the same process risking the same failure.

The positive side to the story is that the municipalities are now learning from Odense's mistakes. They have joined forces and are now together with language technology experts at Copenhagen Business School about to

provide a large repository of text material and sound recordings that can be used to train speech recognitions systems for all the domains they need. They are also co-operating on developing contracts and terms for companies that want to provide speech recognition software. One of the central terms is that the speech and text data used continue to be the property of the municipalities, thus enabling the municipalities to reuse the data for any other application they may need in the future.

What we see here is a growing awareness that every language community needs to be in charge of its own language resources. They are too expensive and too precious to be left at the mercy of private companies who fail to maintain them and to make them comply to open standards.

Danish municipalities are about to create a pool of relevant language data that they would maintain and develop and that they would offer free of charge to any developer who would want to create language technology for Danish.

I hope this will be a significant turning point, because in the area of language technology, Denmark is lagging behind. Recent benchmarks made by the Meta-Net-project have placed Danish on the same level as Bulgarian, Estonian, Greek, Hungarian, Polish, Serbian, Slovak and Slovene languages. All these languages have only fragmentary support for speech processing.

In other countries, such as the Netherlands, Norway, Sweden and Iceland, large language repositories are being compiled as a part of recently adopted national strategies for language technology. They contain written and spoken language as well as parallel texts of translated language, dictionaries and terminology. But for many languages the pace of development is still very slow. European projects such as Clarin and Meta-Net are striving eagerly to bring these issues to the attentions of local governments.

They argue that free access to data repositories containing various language resources is a necessary infrastructure of the information society, but only few governments have recognised this fact.

This year the European Parliament has adopted an update of the so-called PSI Directive, a directive on the reuse of public sector information. It states that documents and metadata are to be made available for reuse under open standards and using machine readable formats. This is a great step forward, and now libraries, museums, and archives are also covered by the directive.

During the next two years the directive will be integrated into the legislation of various EU-countries and it would be a wise step for governments to take the needs of language technology into account when working with the directive and to think about, for instance, how the rules of copyright can be made compatible with the needs to develop language technology.

As the world changes, so do our languages, because they are adapting to our need to express ourselves in a changing world. And so the process of collecting and accessing language data has to be continuous and flexible. It is an integrated and important part of our information society and it needs a proper infrastructure that is constantly maintained and improved, just as we maintain and develop our energy supply and the roads we drive on.

Only if we keep using our languages in all domains, if we keep translating, speaking and writing in our own languages wherever we can, our languages will be the powerful and wonderful tools that we need in order to understand and express our views about the world.

Language technology is there to help us do this, and we can make it work if we are ready to provide the necessary resources.

And language is the greatest resource of them all – the more we use it, the more it grows.