Strengthening the smaller languages in Europe

Steven Krauwer

Utrecht University / ELSNET
Trans 10, 3512 JK Utrecht, The Netherlands
Steven.Krauwer@let.uu.nl

Abstract

The problem we want to address is that - even if the complexity of a language is independent of the number of speakers - industrial developers of language and speech technology (and, unfortunately, hence the EU R&D programmes) focus their efforts on the major languages, because of their economic potential. This is something we cannot change, but in our talk we will discuss what the smaller language communities can do to create optimal conditions for the development of their own language and speech technologies..

Krepitev manjših evropskih jezikov

Čeprav je kompleksnost jezika neodvisna od števila govorcev, industrijski razvijalci jezikovnih in govornih tehnologij zaradi ekonomskega potenciala osredotočajo svoja prizadevanja na velikih jezike, kar na žalost posledično velja tudi za raziskovalno-razvojne programe EU. Tega ne moremo spemeniti, vendar bomo v prispevku preučili, kaj lahko govorci manjših jeziki storijo, da vzpostavijo optimalne pogoje za razvoj lastnih jezikovnih in govornih tehnologij.

1. Introduction

This paper is based on earlier presentations given at the Baltic HLT Conference in Tallinn in April 2005 and at the SALTMIL Workshop at LREC in Genoa in May 2006. The purpose of the paper is not to present new research results, but rather to draw attention to the fate of the smaller languages in Europe and to discuss what we can do to improve the conditions for the smaller languages. In this paper I will use the term smaller languages to refer to languages with limited technological support. There have recently been many discussions on various mailing lists about the most appropriate and politically correct term for this[i], but in the absence of a satisfactory solution I'll stick to smaller for the time being.

2. Roles of Language

Our language is our most important instrument for communication with others. Where in the past the circle of others would normally remain limited to people living in our direct environment (neighbourhood, city, country) the creation and expansion of the EU have made us all member of a much larger community, where more than 60 languages are being used for communication between citizens, 20 of which have the status of official languages. Contrary to the situation in the past we all have to face the fact that most of our fellow EU citizens do not speak or understand our language. This affects a number of aspects of our daily and professional life, and we should ask ourselves to what extent this may cause problems or disadvantages for some of us, and – more importantly – how language and speech technology could help to overcome the problems.

Politically we see that more and more of our local policies are determined by EU legislation coming from Brussels. Although the decision procedures are democratic and every member state gets its chances to participate in the discussions leading to legislative measures and is allowed to use its own language at all formal sessions one may wonder whether everybody's voice is heard equally well during this process and the preparatory stages, where informal discussions may be held in one of the major working languages. At this moment language and speech technology is used by the EU to support professional translators and interpreters, and to provide quick and dirty translations of internal documents between some languages. In spite of these efforts there is no guarantee that all EU legislators are playing on an equal playing field as far as language is concerned.

Economically we can now observe that Europe has become our home market, and the world at large our foreign market. In order to be able to sell products and services both on our home and our foreign market we will always have to cross language barriers. In many countries users of services expect to be addressed in their own language, and very often national legislation requires user manuals to be provided in the national language.

From a cultural point of view we have now become part of the European culture. From an integration point of view it is desirable that our cultural heritage is accessible to our fellow EU citizens, and that they have access to ours. Unfortunately much of this heritage is based on or described in language, which constitutes a major obstacle for mutual cultural exchanges. At the same time we should realize that our language is not only an instrument to convey information about our culture: language is an inalienable part of our cultural identity, and needs to be preserved and protected in the same way we protect buildings, paintings and literature.

Our society and economy become more and more information driven. Unfortunately most information is encoded in language, which means that having electronic access to information is a necessary but not a sufficient condition for having full access to our information society.

Individuals from all member states have become European citizens and can now move freely around in Europe, but one can wonder what it means to be a European citizen if one cannot communicate with most fellow EU citizens. Taking away political frontiers is one step, taking away the language barriers is a natural next step.

3. Where does that leave us?

The ability to cross language barriers is essential for the integration of Europe and for further economic development of the EU as a whole. This is more pressing for small language communities than for the larger ones. One can easily live in an English, French or German speaking region without ever realizing that there exist people who speak a different language. All books are translated, movies are dubbed, and if president Putin, Chirac, Bush or the Pope open their mouth on television a voice-over will take over from him within half a second, unless he happens to speak the local language. If one is living in a smaller language community one is constantly confronted with other languages and with language barriers that have to be crossed.

Traditionally we have three methods to help us to cross language barriers: human translators for written language, human interpreters for spoken language, and (last but not least) learning a foreign language. The first two methods are valid and effective in some situations, but not always applicable in day-to-day communication. The third method can be very helpful in certain situations, but language learning (especially after school age) requires a long-term investment, and there are limits to the number of languages one can learn in a lifetime.

It is not obvious how this situation can ever be improved without the help of technology. In this paper we will discuss how technology can be used to overcome (or at least) reduce the language barriers, and more specifically how we can make sure that both bigger and smaller languages can benefit from the new technologies.

4. The role of language and speech technology

Over the years the EU has invested massively in the development of language and speech technology, and many dedicated R&D programmes have had a significant impact on its advancement, including applications oriented towards solving the multilinguality problem. Even though the 6th Framework Programme, now running towards its end, does not have specific language and speech technology oriented action lines, many of the present projects address language issues.

Unfortunately the strong industrial bias of recent EU programmes has led to a situation where the major part of the funding for language and speech technology goes to the major languages. This is not surprising, as industrial players will prefer to invest in the development and deployment of technologies for larger markets. As a consequence there has been only marginal support for the development of language and speech technology for the language communities that do not constitute profitable markets. As the development cost of such technologies is independent of the number of speakers of a language ("all languages are equally difficult") this has created a very unbalanced situation.

5. What can we do to improve the situation at the EU political level?

At the EU political level it is important that the speakers of smaller languages don't accept that their languages (and the speakers themselves) be marginalized in Europe. It is well-known that the cost (both in time and in money) of multilinguality for the EU is enormous (€ 1123 million in 2005[ii]), and that it will be hard to resist the temptation to reduce the number of official working languages to just a couple. One may be forced to resort to such or similar pragmatic solutions, but representatives of the smaller (or maybe rather commercially not attractive) languages should under all circumstances try to avoid that such pragmatic solutions put them in a disadvantaged position in comparison with those who will be able to use their native languages on all occasions.

It is mandatory to keep the multilinguality problem on the EU agenda as a top priority, and a common responsibility. In this context one should keep in mind that the biggest potential enemy is the so-called subsidiarity principle. There is nothing wrong with the principle as such ("don't treat anything at the EU level that could be treated at the national level"), but in past discussions with EU officials this same principle has been used to explain why the EU could not possibly provide financial support for the technological development of smaller languages, as a language is primarily the responsibility of the national government. This attitude does not only do injustice to the fact that multilinguality is primarily a European problem (as opposed to a collection of national problems), but it also does not seem to be completely consistent with the fact that effectively most of the EU funds for language and speech technology are used to support a few major languages (some of which are supported by strong economies and would actually not need any EU support at all).

One would hope that the coming 7th Framework Programme will recognize the language dimension of Europe, and will address support for language and speech technology development explicitly, irrespective of the economic potential of individual languages or EU world leadership ambitions.

6. What can be done at the national level?

6.1. Human resources

As speakers of smaller languages we have to face the facts: if we don't take care of our languages no one will do it – or Microsoft (provided they judge the potential market interesting enough to make the investment).

In order to properly develop language and speech technology for one's own language (both from a monolingual and from a multilingual point of view) a number of preparations are necessary. First of all language and speech technology have to find their way to higher education curricula. Traditionally language technologists tend to come from a linguistics background, whereas speech technologists have an engineering background. Very few of them have received an education directly aimed at language or speech technology, and there is very little integration between the two. Researchers in more recently emerging areas (multimodality, interfaces, knowledge engineering) have even to a larger extent been obliged to educate themselves, as no standard curricula exist for these fields. Reflection on future curricula seems desirable; in order to be able to offer the next generation of researchers and developers in these (interdisciplinary) fields a better-tailored package of knowledge and skills. In this context we would like to point to initiatives such as the European Masters in Language and Speech Programme[iii], and the European Masters Program in

Language and Communication Technologies[iv], both aiming at defining (and continuously updating) a masters curricula in language and speech technology. The EU Tempus programme offers special mobility grants that can be used to collaborate in the creation of new curricula[v].

When building up local expertise with respect to the national language it is important to keep in mind that even if every language is unique, many problems may manifest themselves in several (often related) languages, and may have been solved there. Even if these solutions might not be directly applicable to one's own language, it is often easier to port the solutions than to try to solve the problem from scratch. In order for researchers to optimally benefit from this it is very important that they get the opportunity to attend international conferences, workshops or courses. The organization of local (or regional) training courses is a very useful instrument to introduce new technologies that have been developed elsewhere.

6.2. Language resources

Language resources (written and spoken corpora, lexicons, parsers, annotation tools, etc) are essential for the development of language technologies and for the training of students. These resources, whatever their nature, have all in common that they are expensive (in time and money) to create. In order to maximally exploit the resources that have been and will be created their re-usability is a very important feature. Funders of the creation of resources should take great care to ensure that once these resources have been created for a specific purpose (e.g. a project) they can be re-used by future projects. This has different aspects:

(i) from an IPR point of view it should be ensured that resources created through public funding can be re-used by others without any legal constraints, at least for research purposes;

(ii) technically these resources should be created in conformity with existing standards or best practice, in order to ensure optimal interoperability with other tools and resources;

(iii) organisationally it should be ensured that a body is identified that is responsible for the maintenance and further distribution of these resources, in order to guarantee that these precious materials do not get lost when research teams are dissolved or new hard- and software platforms emerge.

6.2.1. The BLARK

Given the emergence of statistical methods in all sub-areas of language and speech technology there is virtually no limit to the amount of resources researchers can use. As the creation of such resources can be a significant financial burden ELSNET, in cooperation with a number of partners, including ELDA (Paris), CST (Copenhagen), CNR-ILC (Pisa), is in the process of developing the BLARK concept. BLARK stands for Basic Language Resource Kit, and it aims at defining the minimal collection of resources that is needed to do any research and (precompetitive) development in language and speech technology at all. In its final form it should comprise a list of necessary components (specified both qualitatively and quantitatively), and the standards (formal or de facto) to be adhered to. We will also aim at including cost estimations for the production of the various components, based on experience. The BLARK concept was first launched in the ELRA Newsletter published in May 1998[vi]. The definition allows for adaptations to specific properties of languages.

The BLARK definition should be used as a common reference point for language communities that want to start their own language and speech technology activities, and that need to make up a priority list of what is needed. Once the definition is available teams can make an inventory of what exists and what is missing.

Initial BLARK definitions have been provided for the Dutch language, by researchers associated with the Dutch Language Union. A first inventory and an identification of priorities has led to a large language and speech technology programme funded by the Dutch government and the regional Flemish government in Belgium.

In the framework of the EU funded NEMLAR project[vii] an initial BLARK definition has been prepared for Arabic, and first steps have been made towards the creation of a BLARK for Arabic. The current version of the definition can be found on the same site. [viii]

6.2.2. The BLARKette

One of the findings of the NEMLAR project with respect to the BLARK was that even if a BLARK should be seen as a modest entry point for the creation of resources for a language it has a tendency to grow quickly, as technology advances and discipline boundaries become more and more vague. There is a certain risk that the definition and creation of a full-blown BLARK may be one or more steps too far for smaller, regional languages in Europe, for which no or very little technological support exists, and for which only modest national or regional funding is made available.

In order to accommodate this problem we have proposed the definition of a scaled down, entry-level version of the BLARK, targeting exclusively the research and (especially) the education community. It should be light and compact, not too demanding in terms of hard and software requirements, cheap, free from IPR issues, and ideally small enough to fit on a CD or DVD. We expect to release a first document, with tentative summary specifications, towards the end of 2006. Check the ELSNET site for news[ix].

7. What can be done internationally at the EU level?

Many countries have a long and well-established tradition of national language and speech technology programmes. Within the framework of the creation of the ERA the EU aims at better coordination between national language and speech technology related programmes. Language and speech technology would be an excellent opportunity for such coordination, because it would facilitate both porting of knowledge and expertise between languages addressing cross-lingual issues.

The EU's 7th Framework Programme will also offer opportunities for language and speech technology oriented research and development. There are indications that language and speech technology, which were completely out of focus in the 6th Framework Programme, will be given a more prominent role, and it is hoped (but not guaranteed) that the smaller (i.e. commercially less significant) languages will receive more attention.

Another interesting development is the decision by the EU to add Irish to the set of official EU languages, and to give a similar status (although on a self-paying basis, and on the basis of special agreements) to Galician, Catalan-Valencian and Basque and other officially recognized languages in member states.

8. What sort of language and speech technology solutions are we looking for?

It is easy to say that we should resort to language and speech technology in order to get our multilinguality problems out of the way, but how realistic is this? In spite of all the efforts made by the R&D community machine translation (MT) is still not mature enough to be accepted as a generally applicable solution. For the time being the creation of high quality MT systems is still a wonderful research topic, but nothing more than that.

Yet it has to be kept in mind that even state-of-the-art MT can be useful. The obvious example is just finding out what a mail message or a web page in a foreign language is about. I am receiving hundreds of spam messages per day, but sometimes I am really curious what it is that people are trying to sell me from Russia, Korea or China, and a free on-line MT system is good enough to get an idea.

If you buy an MT system like Systran you can get it almost for free, and the quality is moderate (to put it mildly), but if you are prepared to spend a bit (or rather: a lot) more it can be customized to your specific needs, and the quality level improves dramatically. Like in the case the cheap inkjet printers and the expensive cartridges Systran's real business is not the MT system but its customization.

If your company has a professional translation department the introduction of an MT system can easily save you 30% on your translation costs. The raw translation is not good enough for publication, but the total process of making the raw translation and having it edited by a professional translator can become a lot cheaper and faster.

Unfortunately MT companies will normally not be interested in the development of systems for language pairs for which they don't see a large potential market which will guarantee them a significant return on their investments.

For normal citizens MT is not really a useful option to cross language barriers. In order to find good alternatives we have to abandon the idea that one single solution should solve the problem in all situations. Different situations may require different types of solutions, just like in traffic where you can solve the problem that you happen to be in the wrong place by walking, using your bike or car, taking the train or the plane, or just using the phone.

Let me just give a few examples. Many mobile phones or PDAs come with a small camera these days. Why can't I use this to point at the menu in a restaurant in Ljubljana, have it OCR-ed, translated and displayed on the screen in my own language? As a matter of fact I used this example some ten years ago to illustrate my dreams of what future technology might bring us, and only very recently I read that such a facility now exists for Japanese to English!

Why isn't my PowerPoint presentation displayed on two screens in parallel, one in English and one in Slovenian (by way of – possibly imperfect – subtitles)? Why doesn't the manager of my hotel use a multilingual authoring system to present his announcements in my own language? Why can't I use my mobile phone or PDA to have the spoken word spinach translated in Slovenian and displayed on the screen so that I can show the shopkeeper that it is spinach I want?

The morale of this should be clear: even if we don't know how to do full MT yet there are lots of ways to deal with the language problem in different contexts, especially since many contexts offer opportunities to support language communication with additional modalities (combination of spoken and written language, gesturing, facial expressions, video displays, etc).

Apart from that there seems to be a wealth of opportunities in the development of computer assisted learning of languages, not just in class-room settings, but also for adults who want to learn new languages from home, or when sitting in trains, planes or traffic jams.

9. Concluding remarks

I have tried to describe above why multilinguality is a pressing problem, especially for the smaller language communities in Europe. I have also indicated what one could do to in order to keep the problem on the EU's political agenda, what one can do to strengthen one's own local language and speech technology, and what sort of solutions present day language and speech technology can offer. Personally I do not see an immediate danger that our small languages will disappear in the first hundred years or so, but in my view the real danger is that speakers of smaller languages may find themselves more and more marginalized, both economically and politically, if they don't make a serious effort to overcome the language problem. From my own professional point of view the use of language and speech technology is the most promising direction, but at the same time I would like to make it clear that I also sympathize with the EU's efforts in their language action plan to encourage people to learn at least two other EU languages in addition to their native language, and where language and speech technology can become very important instruments to achieve this.

[i] See e.g. MT-list on http://www.mail-archive.com /mt-list@eamt.org/

[ii] See http://europa.eu/languages/en/document/59#8

[iii] See http://www.cstr.ed.ac.uk/euromasters

[iv] See http://lct-master.org/

[v] See http://ec.europa.eu/education/programmes /tempus/index_en.html

[vi] Also published on http://www.elsnet.org/blark.html

[vii] http://www.nemlar.org

[viii] http://www.nemlar.org/Publications/BLARK-final.pdf

[ix] See http://www.elsnet.org