The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap

Steven Krauwer

Utrecht Institute of Linguistics / ELSNET
Trans 10, 3512JK Utrecht, The Netherlands
steven.krauwer@elsnet.org

Abstract: The problem addressed by this paper is the fact that many languages, especially the languages of little or no commercial interest are lagging behind with respect to the development and the use of language and speech technology. The definition, adoption and implementation of a standard Basic Language Resource Kit (BLARK) for all languages, irrespective of their size or importance, should help to create better starting conditions for research, education and development in language and speech technology. A co-ordinated effort in this direction should help reducing the burden of arriving at a definition of a minimal set of resources required to do any work at all for each individual language, it should facilitate porting of insights and expertise between languages, and it should help ensuring interoperability and interconnectivity.

1. Introduction

In this paper we will first give a brief explanation of ELSNET. In the following section we will give an outline of ELSNET's roadmapping activities, especially with respect to language resources. We will then describe the problem we want to address and investigate how a co-ordinated approach to the definition and creation of language resources may help improving the situation. In the next section we will develop the BLARK concept, followed by a section where we illustrate the concept and the way it could be instantiated on the basis of the first BLARK proposal that was produced by the Dutch Language Union for the Dutch Language [1]. We will end with some concluding remarks.

2. ELSNET

ELSNET, the European Network in Human Language Technologies, was first created in 1991, as one of the first so-called European Networks of Excellence, funded by the European Commission under the ESPRIT programme. The network unites some 145 organisations in 29 countries, all active in language and speech technology, both in research and in development or integration, mostly based in Europe. Ca 60% of the members are academic institutes, the remaining 40% are private companies.
ELSNET's main objectives are (i) bringing together the language and speech communities, (ii) bringing together academia and industry, and (iii) facilitating R&D in language and speech technology. Its main action lines are training, information dissemination, resources & evaluation, and roadmapping. Its main instruments are publications on the web, mailing lists, summer schools, courses, workshops, and a quarterly paper newsletter. More information about ELSNET, its activities, its members, and application for membership or associate membership can be found on its website http://www.elsnet.org.

3. The ELSNET Roadmap

3.1. What do we mean by a roadmap

Technology roadmapping has become a popular activity in all subareas of technology over the last ten years. There exists no single universal definition of what a roadmap is and how it should constructed or presented. In the ELSNET view of the world a roadmap is a broadly supported vision of where the field is moving, not only from the research perspective, but also (and according to some even more importantly) from the market perspective.
The roadmap identifies future developments, main challenges, intermediate milestones and their interrelationships. The milestones serve as intermediate goals and may help us in measuring progress -- or reconsidering our longer term goals when the results do not correspond to the expectations, or if (by internal or external factors) the playing filed has changed.

A roadmap, provided it is broadly supported by the field, will help researchers, educators, developers, service providers, and funders deciding where to concentrate their efforts in order to give a maximal push to the development of the field.
A few words of caution are in place here. First of all it should be noted that a roadmap is neither a prediction nor a commitment, but rather an expectation. Secondly, a roadmap is necessarily very dynamic in the sense that e.g. changes in funding policies may have a severe impact, both positive and negative. Third, it is very dynamic in the sense that external factors, such as evolving technologies and markets, political crises, social factors may lead to dramatic changes.
Roadmaps will have to be continuously updated in order not to lose their value within months after publication.

3.2.The ELSNET approach

The ELSNET approach to roadmapping is based on a cyclic process, where we start consulting a specific subcommunity by means of workshops or expert meetings. We then try to extract relevant challenges, milestones and their interdependencies, timescales and other relevant factors, which we then include in our roadmap. The results are published on the web, and the members of our community are invited to provide feedback. We adjust the roadmap on the basis of the feedback provided, and then start a next cycle, on a different topic area. Gradually we hope to complete the picture, but as we said above: the exercise has to be repeated on a regular basis in order for the roadmap not to become obsolete.
The roadmap itself takes the shape of a collection of objects (challenges, milestones) in different categories (at this moment resources, technologies and applications), with a number of extra information items (such as expected year of completion, definition, background documentation, etc) attached to it. For presentation on the web we have adopted a graphical metaphor, which displays the objects on a 3-lane motorway, going from now to the future, with 1 year intervals. Users can zoom in on the picture, display properties and interrelationships, and they can comment on the roadmap as it is. To see the graphical representation see http://elsnet.dfki.de. For an overview of our resources activities thus far visit http://www.elsnet.org/roadmap.nl.

3.3.The resources roadmap

At this moment the roadmap activities cover a number of topic areas, such as speech technology, machine translation, multimodality and knowledge management. Not all of the results have been integrated yet, but we expect to complete this before the end of 2003. A subcomponent of our roadmap that goes across all topic areas is dedicated to language resources. Because of their generic and underpinning nature they constitute a separate lane of our motorway. The collection of data and milestones for the resources roadmap has been initiated in close collaboration with the ENABLER network, which is an EU funded network of national resources projects. The website is http://www.enabler-network.org. In this paper we will focus on the resources part of the roadmap, and more specifically on what we feel should be the first milestone for every single language.

4. The problem

European citizens will be living in one

- economic space
- cultural space
- monetary space (eventually)
- information space
- political space (within certain limits)
- touristic space
- entertainment space
- …

But we are NOT living in one linguistic space, as can be illustrated by the following facts about the EU alone:

- 15 member states, with 11 official languages (plus quite a few 'unofficial languages')
- 10 new member states with (at least) 10 new official languages
- 3 applicant countries with at least 3 extra languages

And Europe has 17 other countries with lots of other languages -- but even so it is still in a much better position than the rest of the world, as the following figures from the Ethnologue website [http://www.ethnologue.org] show:

- Europe: 230 languages
- The Americas: 1013 languages
- The Pacific: 1311 languages
- Africa: 2058 languages
- Asia: 2197 languages

The question is: what will happen to all these languages and to the position of their speakers as the global information society is gradually expanded.

I see at least three possible scenarios:

1. *A few big languages end up dominating the scene, smaller languages will gradually disappear*
2. *A few big languages end up dominating the scene, and although the smaller languages are preserved, their speakers will be marginalized*
3. *Language and speech technology will be used to ensure participation of all Europeans in the European space on an equal footing, irrespective of their language*

Without much hesitation I would choose the third scenario as my favourite, not just because I happen to be professionally active in language and speech technology, but more importantly because I happen to be a speaker of one of the smaller languages in Europe, and I don't want to be excluded from participation in the global information society because I happen to be born with the wrong native language. A direct consequence of this choice is that we have to make language and speech technology for all languages. It is clear that we need not be worried about the commercially important languages, such as English, French and German. There is a huge potential and wealthy market, so that the big market players will take care of these languages. But who will provide the proper technologies for the other languages if the market doesn't do it? Smaller languages have smaller national economies, and hence less opportunities for private or public funding for the production of technologies. All languages are equally difficult, but their financial conditions can differ dramatically. This is not a problem we can solve here and now, but there is one thing we, as a research community, can do and that is to contribute to the creation of the best possible starting conditions for language and speech technology to take off: the creation of human resources (educate language and speech researchers and engineers, design proper curricula) and of language resources (corpora, dictionaries, parsers, tools, etc, and an infrastructure around them).
We will focus on the latter point: language resources. Language resources exist for many languages. More for some and less for others – but when can one say that a language is properly covered from a resources point of view? There is no broadly accepted definition of what counts as sufficient. In the next section we will outline how we intend to arrive at such a definition by introducing the concept of a BLARK: a Basic Language Resource Kit.

5. The BLARK

5.1.Description of the concept

We define the Basic Language Resource Kit (abbreviated BLARK) as the minimal set of language resources that is necessary to do any precompetitive research and education at all. The definition is in principle intended to be language independent, but as specific languages may come with different requirements, instantiations of the BLARK may vary in some respects from language to language.
A BLARK comprises many different things, such as

- written language corpora
- spoken language corpora
- mono- and bilingual dictionaries
- terminology collections
- grammars
- modules (e.g. taggers, morphological analysers, parsers, speech recognisers, text-to-speech)
- annotation standards and tools
- corpus exploration and exploitation tools
- bilingual corpora
- etc

The list is far from exhaustive but serves to illustrate the scope of the BLARK. In addition it should comprise an infrastructure for the management, maintenance and distribution of the resources. A BLARK should not be seen as a static object: over time it may gradually evolve as new technologies and application areas emerge, with new requirements in terms of resources. The idea was first launched in the ELRA Newsletter in 1998 [2].

5.2.What makes the BLARK special?

The underlying idea is to make a common generic BLARK definition, applicable in principle to all languages, based on the collective experience and expertise gained with many different languages by the members of the language and speech technology community at large. This common definition will save time and effort (no reinvention of wheels), it will allow for porting of knowledge between languages, it will ensure interoperability and interconnectivity (especially for multilingual or cross-lingual application areas), and it will help making realistic estimates of costs and efforts required to produce them. In addition a broadly supported common definition may be used as an external reference point in discussions with funding agencies about the best way to create a good starting point for language and speech technology, both in academic and industrial (precompetitive) research and academic and professional training.
In order to make a BLARK maximally impactful it should be freely accessible and usable, preferably on the basis of an open source arrangement.

5.3.How to use it

The target audience of the BLARK is researchers (both in academia and in industry), and educators. It is used to train students, to serve as material for research experiments and application pilots. Commercial companies should be free to use the BLARK for the development of commercial products, but in general it is unlikely that BLARK components will be usable for commercial applications as they are. Therefore it is of crucial importance that the BLARK comes with tools for the production and annotation of new corpora, and that all modules and resources are available in source format, so that industrial developers can freely adapt them to the specific requirements of their applications (e.g. domain, footprint, application environment).

5.4.How to arrive at it

At this moment ELSNET and its sister project ENABLER, that has just ended, are in the process of producing an initial BLARK definition. ELSNET will continue this activity in close collaboration with the participants in the ENABLER project, and with others who want to contribute or who are interested in adopting the BLARK for their own language. Once a first definition is in place, each language should try to make an inventory of which BLARK components are already available for their language, and which ones are missing. The amount of missing components may vary dramatically from language to language, as some of the major languages such as English may already be fully covered, whereas others may have to start from scratch. Once the gaps have been identified, priorities should be assigned to the components to be produced, so that a realistic plan for the gradual completion of the BLARK becomes feasible.

6. The BLARK proposal for the Dutch language

6.1.Background

The Dutch language is spoken in the Netherlands and in the Flemish part of Belgium by some 20 million speakers. In comparison with English, German and French it is one of the smaller languages in Europe. A special intergovernmental organisation, the Dutch Language Union (in Dutch Nederlandse Taalunie, abbreviated NTU) was created by the Dutch and Flemish governments to take care of the Dutch language. NTU was the first party to adopt the BLARK concept back in 1998, and they were the first to produce a rather detailed definition of the BLARK, both with respect to the linguistic components and with respect to the priorities for its completion and with respect to proposals for a management, maintenance and distribution structure. Below we will sketch the methodology adopted by the NTU team and their findings -- not because many non-Dutch speakers will be interested in the resources situation for this language, but rather because the underlying methodology was very systematic and can easily be followed by others, and because the resulting BLARK can be seen as a first proposal for the language independent BLARK definition.

6.2.The Dutch BLARK

An excellent summary of the process and the results of the Dutch BLARK exercise can be found in an article by Binnenpoorte et al [1] in the proceedings of the LREC 2002 workshop "*Towards a Roadmap for Multimodal Language resources and Evaluation*" organized by ELSNET. The authors have kindly allowed us to copy the matrices summarizing the results.
The process comprised three stages: (i) definition of the BLARK, (ii) making an inventory of what is available, and (iii) assigning priorities. Starting point in the definition were 8 classes of applications: computer assisted language learning, access control, speech input, speech output, dialogue systems, document production, information access and translation. For each of them it was established which modules would be needed to make them (e.g. morphological analysis, speech synthesis), and for each of these modules it was analysed which language data (e.g. data sets, descriptions) they would require, as well as their relative importance. The matrix in table 1 at the end of the paper shows the results (+ = *relevant*; ++ = *important*). On the basis of this matrix one can determine which components serve most applications, and which data are most needed for most applications, i.e. which elements should be part of the BLARK. We briefly summarize them here.

For language technology the following elements were identified:

- robust text pre-processing
- morphological analysis
- syntactic analysis
- semantic analysis
- monolingual lexicon

- written Dutch tree-bank
- evaluation benchmarks

For speech technology:
- automatic speech recognition
- speech synthesis
- calculation of confidence measures
- identification tools
- tools for (semi-)automatic annotation of speech corpora
- speech corpora for specific applications
- multi-modal speech corpora
- multi-media speech corpora
- multi-lingual speech corpora
- evaluation benchmarks

When the list of modules and data was completed, an inventory was made in order to determine their availability. As availability is not really a binary distinction (materials may exist, but may not be freely usable, or they may not have the desired quality or coverage) a ten point scale was used to describe availability status. The list is shown in table 2.
On the basis of a comparison of the definition of what was most needed (the BLARK) and the availability analysis a priority list was made and used as the starting point for a plan to complete the BLARK for the Dutch language.

7. Concluding remarks

In the preceding sections we have given a brief overview of ELSNET's roadmap actions, and we have emphasized that especially for the smaller or economically less attractive languages the creation of a solid starting point in terns of language resources is of the utmost importance as the very first milestone along the road. Whereas for the bigger languages a healthy resources situation will be created by the market parties anyway, the smaller languages can benefit a lot if they agree to join a collaborative effort aimed at the definition of a uniform Basic Language Resource Kit for all languages. ELSNET will (together with the former ENABLER partners) continue to move towards the definition of the BLARK, but it should be clear that this effort can only be successful if it can rely on the support from the various language communities in Europe. Even if ELSNET will not be in a position to give financial support to the actual creation of a BLARK for a specific language, we will aim at bringing together the various communities and at establishing an organisational framework that will allow us to create and dynamically update the BLARK definition, and to support all activities related to the actual implementation of BLARKs for specific languages.

8. References and acknowledgements

[1] Diana Binnenpoorte, Catia Cucchiarini, Elisabeth D'Halleweyn, Janienke Sturm and Folkert de Vriend, "*Towards a roadmap for Human Language Technologies: Dutch-Flemish experience*", Proceedings of the workshop "Towards a Roadmap for Multimodal Language Resources and Evaluation" at LREC 2002, Las Palmas, Canary Islands, June 2002. [Also on http://www.elsnet.org/dox/lrec2002-binnenpoorte.pdf]

[2] Steven Krauwer, "*ELSNET and ELRA: Common past, common future*", ELRA Newsletter, Vol 3 nr 2, May 1998. [Also on http://www.elsnet.org/dox/blark.html]

ELSNET is supported by the European Commission's Human Language Technologies Programme

Tables

Modules	Data									Applications							
	monoling lex	multilin lex	thesauri	anno corp	unanno corp	speech corp	multi ling corp	multi mod corp	multi media cor	CALL	access control	speech input	speech output	dialog systems	doc prod	info access	transla-tion
Language Technology																	
Grapheme-phon. conv	++			++						+			++	++	+	+	
Token detection	++			+	++					+		+		+	+	+	+
Sent boundary detection	+			++	++					+		++	++	+	++	++	++
Name recognition	+	+	+	++	++	++				+		++	++	+	++	++	++
Spelling correction										+							
Lemmatising	++			++	+					+		+	+	+	+	+	+
Morphological analysis	++			++	+					+		+	++	+	++	++	++
Morphological synthesis	++			++	+					+			++	+	++		++
Word sort disambig.	++			++	+					+		++	+	++	++	++	++
Parsers and grammars	++			++						+		++	++	++	++	++	++
Shallow parsing	++			++	++					+		++	++	++	++	++	++
Constituent recognition	++			++	+					+		++	++	++	++	++	++
Semantic analysis	++		++	++				++	++	+		++	++	++		++	++
Referent resolution	+		++	++	+					+		++		++	++	++	++
Word meaning disambig.	+		++	++	+					+		++	+	+	+	++	++
Pragmatic analysis	+		+	++				++	++	+		++	++	++		+	++
Text generation	++		++	++				++	++	+			++	++	++		++
Lang. dep. translation		++	++	++			++			+						++	++
Speech Technology																	
Complete speech recog.	++	+		++	+	++	+	++	++	++	++	++		++	++	++	++
Acoustic models	++	+		++	+	++	+	+	+	++	+	++		++	+	+	+
Language models	+			++	+	+	+	+	+	++	+	++		++	++	++	++
Pronunciation lexicon	++	+		+		++	+	+	+	++	+	++	+	++	+	++	++
Robust speech recog.	+			+	+	+	+	+	++	+	+	++		++	+	+	+
Non-native speech recog.	+	++		+		++	++	+	+	++	+	+		+		+	+
Speaker adaptation	+			+	+	++	+	+	++	+	+	++		+	+	++	+
Lexicon adaptation	++	+		+		++	+	+	+	++	+	++	+	++	+	++	++
Prosody recognition	+	+		++	+	++	+	+	+	++	+	++		++	++	++	++
Complete speech synth.	++	+		+		+		+		+			++	++	+	+	++
Allophone synthesis	+	+		+		+		+		+			+		+	+	+
Di-phone synthesis	++	+		+		+		+		+			++	++	+	+	+
Unit selection	++	+		+		+		+		+			++	++	+	+	+
Prosody prediction for Text-to-Speech	++	+		+		+		+	+	++			++	++		+	++
Aut. phon. transcription	++	++		+	+	++	+	+	+	++	+	+	+	+	+	+	+
Aut. phon. segmentation	++	++		+	+	++	+	+	+	++	+	+	+	+	+	+	+
Phoneme alignment	+	+		+		++	+	+	+	++	+	+		+			+
Distance calc. phonemes	+	+		+		++	+	+	+	++	+	+		+			+
Speaker identification	+			++	++	++	+	++	+	+	++	+		+		+	+
Speaker verification	+			++	++	++	+	++		+	++	+		+		+	+
Speaker tracking	+			++		++			++	+	++	+		+	+	+	+
Language identification	+	++		+	+	++	++	+	+	+	+	+		+		+	+
Dialect identification	+	++		+	+	++	++	+	+	+	+	+		+		+	+
Confidence measures	+			+	+	++	+	++	+	++	++	++		++	+	+	+
Utterance verification	+			+	+	++	+	+	+	+	+	++		++	+	+	+

Table 1 *Overview of the importance of data for modules and the importance of modules for applications.*

Modules	**Availability**
Grapheme-phoneme conversion	8
Token detection	9
Sentence boundary detection	3
Name recognition	4
Spelling correction	3
Lemmatising	9
Morphological analysis	7
Morphological synthesis	9
Word sort disambiguation	7
Parsers and grammars	3
Shallow parsing	2
Constituent recognition	5
Semantic analysis	3
Referent resolution	2
Word meaning disambiguation	2
Pragmatic analysis	1
Text generation	3
Language dependent translation	3
Complete speech recognition	4
Acoustic models	8
Language models	3
Pronunciation lexicon	5
Robust speech recognition	2
Non-native speech recognition	2
Speaker adaptation	2
Lexicon adaptation	2
Prosody recognition	2
Complete speech synthesis	6
Allophone synthesis	7
Di-phone synthesis	6
Unit selection	1
Prosody prediction for Text-to-Speech	3
Autom. phonetic transcription	3
Autom. phonetic segmentation	5
Phoneme alignment	8
Distance calculation of phonemes	8
Speaker identification	2
Speaker verification	2
Speaker tracking	2
Language identification	2
Dialect identification	2
Confidence measures	2
Utterance verification	2
Data	
Unannotated corpora	9
Annotated corpora	5
Speech corpora	4
Multi lingual corpora	3
Multi modal corpora	1
Multi media corpora	1
Test corpora	1
Monolingual lexicons	8
Multilingual lexicons	6
Thesaurus	4

Table 2 *Availability of modules and data*