



Language and speech technology: threats and opportunities

Steven Krauwer
Utrecht University / CLARIN

Possible perspectives



Two options:

- Big language perspective (with many speakers and a strong economy)
- Small language perspective (relatively few speakers or weak economy)

My perspective: small languages

- Big languages will always manage, but what are the threats and opportunities for small languages?

Threat 1: globalization and market drive



From the ICT 2009 Conference in Lyon:

- Future playing field for ICT is global
- ICT should be driven by the market

What does this mean for language and speech technology in Hungary?

Opportunities (1)



- EU: More attention for Social Sciences and Humanities (e.g. CLARIN, HERA, Marie-Curie)
- Hungarian may be hard to sell but expertise with more complex languages than English can be exportable and exploitable (cf Finnish successes in morphology)
- Invest and participate in the development of new services that are not language dependent

Threat 2: marginalization



- Small languages risk to be marginalized
- Same applies to their speakers, both at the political and the economic level

Opportunities (2)



- EU: decision to maintain equality of languages and to accept the consequences
- Reflected by ICT Call 4 (Language based interaction, now open), with strong focus on crossing language boundaries for all EU languages (see CORDIS site)
- Information days 14-15 Jan, Luxembourg
More info on: <http://www.elsnet.org/ict>
- ICT proposers day 22 Jan, Budapest
See http://ec.europa.eu/information_society/events/budapest_2009/index_en.htm

Threat 3: limited access to information and knowledge



- Most information and knowledge is NOT written in Hungarian
- ICT of the future will be knowledge based
- Political frontiers may be disappearing but Cyberspace is still full of language barriers

Opportunities (3)



- EU: commitment to Single European Information Space; reflected by ICT-PSP Call 3 (expected Jan 2009):
 - machine translation for the multilingual web
 - multilingual web content management
 - standards & best practices for the multilingual web
- Information day 26 Jan, Brussels
- More info: see link to ICT-PSP on <http://www.elsnet.org/ict>

Threat 4: Loss of cultural identity



- Language is an important component of cultural identity
- Large part of cultural heritage is described in language but often not easily accessible (hidden in different archives, sometimes even in different languages)
- Part of our cultural identity is based on how others see us and our culture, provided they can cross the language barrier

Opportunities (4)



Steven Krauer
Budapest
02-12- 2008

www.clarin.eu

- EU: ESFRI Roadmap for Research Infrastructures addresses also social sciences and humanities (SSH)
- Includes CLARIN, aimed at creating a federation of digital language based archives plus access through language and speech technology tools for SSH research in all EU countries for all languages
- Hungary one of the leading players (HASRIL)

Threat 5: Fragmentation



- Scientific research tends to be fragmented, not just at the European level, but also at the national level
- Fragmented research reduces competitiveness both nationally and internationally
- Linguistics is a very bad example of this (compare authorship of conference proceedings in e.g. physics and linguistics)

Opportunities (5)



- EU: European Research Area for coordination of research (e.g. HERA for coordination of humanities research)
- At the national level: creation of national platforms to bring researchers / developers / deployers together to create critical mass and coordinate efforts to increase quality and competitiveness
- CLARIN, as a pan-European infrastructure, will also explicitly address defragmentation at national and European level

Threat 6: Big players will take over



- If we don't protect ourselves and our language big companies like Microsoft may end up defining our language (through spelling checkers, grammar checkers and authoring tools) ...
- ... and Google may buy all our young talents in language and speech technology

Opportunities (6)



- National initiatives needed to strengthen own language and ensure availability of tools and applications, stand-alone or integrated in products of global players
- National education initiatives needed to ensure sustainable availability of excellent staff
- National initiatives needed to ensure availability of essential language resources (BLARK) for research and education

Threat 7: Absence of a 'Big Threat'



Big Threats are a great trigger of funding:

- Climate change
- Pollution
- Energy
- Health
- Aging of population

Opportunities (7)



- We DO have a threat: “Social Climate Change” caused by mobility and migration.
- Language and speech technology can help e.g.
 - to overcome language barriers
 - to detect changes in attitudes, e.g. by monitoring the media

What a platform could do



- Bring together the main players (researchers, developers, deployers) and offer instruments for collaboration in order to strengthen the field, both in research and in development
- Agree on standards to improve sharability and interoperability
- Agree on basic resources required
- Agree on curricula aimed at excellence
- Agree on strategic priorities

Concluding remarks



- I have sketched threats and opportunities from the perspective of a small language
- My overall impression is that Hungary has always been and still is a strong international player:
 - Morphologic 1991
 - ELSNET Summer School 1996
 - Eurospeech 1999
 - EACL 2003
 - Key player in CLARIN
- I see the creation of a national platform as an important strategic move and I am looking forward to continuing our collaboration