SaLTMiL workshop on
Creation and use of basic lexical resources for less-resourced languages

Panel discussion: **Less-resourced languages and Language technology. Short- and medium-term objectives**

LREC 2010, Valetta, Malta, 23 May 2010

A wide variety of "less-resourcedness"...

Less-resourced languages (LRL) covered by the papers in this workshop:

- Nganasan (`nio`, 500 middle-aged and elderly speakers)
- Iñupiaq (`ik`, 2,000 speakers)
- Latin (`la`, official in the Roman Catholic church, Wikipedia, periodicals, radio stations)
- Icelandic (`is`, 300,000 speakers, national language)
- Sorani Kurdish (`ckb`, 8,000,000 speakers)

Each LRL has a different way of being less-resourced.

What do we really want to have...?

What do we really want to have in the short or medium term for each less-resourced language (LRL)?

- Having...
  - ...just resources (e.g. lexical resources)...
  - ... vs...
  - ...having real-life applications (spelling or grammar checkers, machine translation systems...)
- Is the BLARK (*Basic LAnguage Resource Kit* , http://www.blark.org) idea still alive?

We should focus on lexical resources (workshop theme).

Strengths

- Some LRL communities have well-educated language activists: how do we motivate them to get involved?
- Growing interest in less-resourced languages in the field of human language technologies:
  - 7 SaLTMiL workshops
  - FLaReNet workshop 2009
  - AfLaT 2010: African Language Technology @LREC 2010
- Growing availability of open content (e.g. Wikipedia)

Challenges 1/3

- standardization (spelling, morphology) of some LRLs
- interoperability of built resources
  - use of standardized formats and representations
  - modularity
- dissemination: how do we make the resources known and available to all involved? (users, developers, researchers)
- networking (a stronger role for SaLTMiL, collaboration with other societies or initiatives: Foundation for Endangered Languages, SIL International, Bisharat etc.)

Challenges 2/3

- how can we effectively harness the wealth of (non-computational) linguistic talent ?
- choice of licensing (free/open source vs. "academic" licenses): effect on availability,
- return on investment?
    - can we expect return on investment for basic resources for endangered or minority languages?
    - linguists may be willing to create resources for a language even where there is no hope of turning a profit ("every language is interesting")
- organizing/motivating the users of the LRL
- simplifying the elicitation and encoding of linguistic knowledge

Challenges 3/3

- Beating Google (many LRLs in Google Translate: cy, eu, ga, gl...):
  - Should LRL users try? Why?
  - Can researchers working in isolation compete with Google ?