«Getting Less-Resourced Languages on-Board !» Workshop Report

Joseph Mariani (LIMSI-CNRS & IMMI)
Khalid Choukri (ELRA & ELDA)
Zygmunt Vetulani (Adam Mickiewicz University, Poznan)

Getting LRL on-Board WS

- Satellite event of LTC'09
  - Poznan, Nov. 6, 2009
- Co-organized
  - J. Mariani, K. Choukri and Z. Vetulani
- Report available on FLaReNet Web site
  - http://www.flarenet.eu/sites/default/files/LTC'09_LRL_Report.pdf

Program Committee (18)

- Nuria Bel (Univ. Pompeu Fabra, Spain)
- Gerhard Budin (Univ. Vienna, Austria)
- Nicoletta Calzolari (ILC, Italy)
- Daffyd Gibbon (Univ. Bielefeld, Germany)
- Marko Grobelnik (J. Stefan Institute, Slovenia)
- Jan Hajic (Charles Univ., Czech Republic)
- Alfred Majewicz (UAM, Poland)
- Asunción Moreno (UPC, Spain)
- Jan Odijk (Univ. Utrecht, The Netherlands)
- Nicholas Ostler (Foundation for Endengered Languages, UK)
- Stellios Piperidis(ILSP, Greece)
- Gabor Proszeky (Morphologic, Hungary)
- Mohsen Rashwan (Cairo Univ., Egypt)
- Kepa Sarasola Gabiola (Univ. del Pas Vasco, Spain)
- Marko Tadić (Croatian Academy of Sciences and Arts, Croatia)
- Dan Tufiş (RACAI, Romania)
- Cristina Vertan (Univ. Hamburg, Germany)
- Briony Williams (Bangor Univ., UK)

Theme

- LT are essential to support the challenge of Multilingualism in order to :
  - Preserve languages and cultures
  - Allow for communication among humans speaking different languages
- Necessary to have access to LR and to LT evaluation for developing a LT in a given language
- Situation very different across different languages
  - Little or sparse data exist for languages for which limited efforts have been devoted to such issues in the past
  - « Less-Resourced Languages (LRL) »

Theme

- Aims of the workshop:
  - reporting the needs,
  - presenting achievements and on-going plans,
  - proposing solutions for the future,
    - both in terms of LR and of LT evaluation,
    - especially in European, Euro-Mediterranean and regional frameworks.
- Objective
  - identify key factors that have an impact on a shared roadmap towards supplying LR and LT for all languages.

Topics

- Experience in the specification, production, validation and distribution of LR for Less-Resourced Languages
- Experience in the evaluation of LT for Less-Resourced Languages
- Infrastructures for making available LR and LT in Less-Resourced Languages
- Alternative approaches (comparable corpora, pivot languages, language-family clustering, portability…)

Program

- **13h45-15h30 Part 1 (Chair: *K. Choukri*)**
  - 13h45-13h50 Introduction (*J. Mariani*)
  - 13h50-14h15 Keynote Talk : *B. Williams*, "Less-resourced languages and language resources: lessons learned from the Celtic languages of Britain and Ireland"
  - 14h15-14h30 Surprise Guest Talk: *Girish Nath Jha* Indian Languages Corpora Initiative (ILCI-TDIL)
  - 14h30-14h50 *M. Rusko, S. Darjaa, M. Trnka*, Speech synthesis as a first step towards speech technologies in Romani
  - 14h50-15h10 *M. Yifiru Tachbelie, S. Teferra Abate, W. Menzel,* Morpheme-Based Language Modeling for Amharic Speech Recognition
  - 15h10-15h30 *N. D. Snoeren, M. Adda-Decker*, Pronunciation and Writing Variants in Luxembourgish: The Case of Mobile N-Deletion in Large Corpora
- 15h30-16h00 (Real) Break

Program

- **16h00-17h25 Part 2 (Chair: *J. Mariani*)**
  - 16h00-16h20 *I. Alegria, M. Aranzabe, X. Arregi, X. Artola, A. Diaz de Ilarraza, A. Mayor, K. Sarasola*, Valuable Language Resources and Applications Supporting the Use of Basque
  - 16h20-16h40 *M. Farrus, G. Iglesias, C. Henriquez, M. Poch, R. Munoz, N. Ezeiza, E. R. Banga, J. B. Marino*, An experience on statistical machine translation between Spanish and the regional languages of Spain
  - 16h40-17h00 *D. Beermann P. Mihaylov*, Interlinear Glossing On-Line
  - 17h00-17h20 *L. Hellan, M. Esther, K. Dakubu*, A methodology for enhancing argument structure specification
  - 17h20-17h25 Conclusion (*K. Choukri*)
- 17h25-17h35 (Technical) Break

Program

- **17h35-19h00 Part 3 (Chair: D. Gibbon)**
  - 17h35-17h50 Invited Talk: *Aleksandra Wesolowska (EC)*, New start for European language technology. Are you ready?
  - 17h50-19h00 Panel Session “Getting less-resourced languages on board: which languages, which resources, which reasons, which methods, which funding ? “
    - **Moderator: *D. Gibbon***, Panelists: *Aleksandra Wesolowska*, *Marc Kemps-Snijders* (repr. Clarin), *Khalid Choukri* (repr. Flarenet and ELRA), *Alfred Majewicz* (Poland), *Girish Nath Jha* (India), *Briony Williams* (UK)

Conclusions

- Generally speaking, a strong political will to consider the language dimension (not only lip-service) and enough funds are necessary.
- This must go with the awareness that Language Technologies and Language Resources are important.
- There should be specialists in the processing of that language, reaching a critical mass, and young researchers should be trained.

Conclusions

- An infrastructure must exist, including:
  - a writing system/a transcription code/an agreed orthography,
  - Language Resources (sufficient in quantity and quality),
  - tools (especially language independent (based on statistical training) ones, if possible as Open Source),
  - metadata, annotation schemes, standards,
  - development platforms,
  - evaluation means (adapted to the language specificities (such as for Machine Translation of morphologically-rich languages)).

Conclusions

- The effort should be devoted in the long-term, resulting in a necessary strong foundation.
- Dialects variants and sociolinguistics should also be taken into account.
- Addressing only the short-term development of a specific product or service for that language (as a kind of simple toy), should be avoided, whereas demonstrating applications based on a strong foundation should be favoured.

Conclusions

- When a majority language also exists, both should be studied together, and it would save time and efforts to consider a family of languages all together.
- Bootstrapping approaches facilitate the coverage of a language.
- Cooperation among countries or programs would greatly help by providing the less advanced ones with examples and Best Practices, such as the definition of a commonly agreed basic set of Language Resources which have already been proven necessary to correctly produce the corresponding technologies for a given language, with an estimate of the cost, and the identification of gaps and roadmaps should be aimed at.

Conclusions

- The related costs could be shared between the corresponding countries or regions, and international bodies (such as the EC), which could also ensure a proper coordination.
- Master keywords should be
  - *Interoperability (*which means coordination)
  - and *Sustainability (*which means a strong foundation*)*