

Language Technologies for Less-Resourced-Languages LRL Workshop Series at the Language and Technology Conferences (LTC) from 2009 to 2019

Zygmunt Vetulani¹, Khalid Choukri², Joseph Mariani³ and Patrick Paroubek⁴

Adam Mickiewicz University in Poznań¹, ELRA-ELDA², LIMSI/CNRS^{3,4}
ul. Uniwersytetu Poznańskiego 4, 61-614 Poznań, Poland¹, 9 Rue des Cordelières, 75013 Paris, France², Campus
universitaire bât 507, Rue du Belvédère, F - 91405 Orsay cedex^{3,4}
vetulani@amu.edu.pl, choukri@elda.org, joseph.mariani@limsi.fr, pap@limsi.fr

Abstract

When LTC started in Poznań, Poland, as LT Awareness Days (1995), Polish was still a "less-resourced-language". This report presents the LRL Workshop Series organized since 2009 as integral part of LTC. We present as *raison d'être* of LRL to contribute to a "roadmap towards supplying LR and LT for all languages". We go one by one through all LRLs (until 2019) to present themes suggested by organizers, affiliation countries of the authors, as well as the concerned languages. We note positive phenomena such as appearance of countries and languages so far very rare at the international LT conferences.

Keywords: language resources, less-resourced-languages, language technologies, LTC

Résumé

La première conférence LTC, tenue à Poznań (Pologne) en 1995, portait le nom "L&T Awareness Days", et à cette époque, le polonais faisait partie des langues peu dotées. Nous présentons ici la série des ateliers sur les langues peu dotées (LRL Workshop Series) qui font partie intégrale de la conférence LTC depuis 2009. LRL a pour sa raison d'être d'apporter une contribution à la feuille de route pour doter toutes les langues de ressources et technologies du langage. Nous parcourons un à un les ateliers LRLs jusqu'en 2019 pour rappeler les thèmes proposés par les organisateurs, les pays d'affiliation des auteurs et les langues concernées. Nous notons avec satisfaction l'émergence à LTC de pays et de langues rarement vues à l'international.

1. Introduction

The Language and Technology Conferences (LTC) started in Poznań, Poland, in 1995. The first event of this series were Language and Technology Awareness Days co-organized by Adam Mickiewicz University, Poznań and the European Commission – DG XIII. At those days the Polish language was clearly a "less-resourced language" with, however, a solid traditional linguistic description based on strong logical and mathematical background. Similar situation was in other central and East European countries. During the next 15 years the golden age of Language Technologies continued and resulted in reduction of the initial technological gap between these countries. This was due to the intensive international collaboration within joint projects and individual mobility of experts and researchers. The series of LTC events, supported by ELRA/ELDA, FlareNet and Meta-Net, were part of the global effort to foster integration of the LT community in Europe and abroad, for all languages

Since 2005 until now the LTCs were organized every two years as an international forum, open for both academia and language industry – these two communities together contributing to the development and dissemination of language technologies. The first global-scale event of 2005 mobilized over 100 contributors from Europe, Asia and North America. It was dedicated to the memory of Maurice Gross and Antonio Zampolli – two visionary personalities among the first to understand the necessity to

bring together the two communities to favor the emerging language industries.¹

But with dynamic development of language technologies and language industry, as well as amplification of globalization trends, the new threats of technological exclusion become more present in a world of fierce economic competition because of the world population increase and dwindling natural resources.

2. LRL Workshops

2.1 Kick-off

In 2009, we decided to attract more attention of the LT community to the case of the technologically "under-resourced" countries and menaced by technological, and successively, cultural exclusion. This idea gave birth to the Less-Resourced Languages Workshops (LRL) as being an integral part² of the LTC conferences. The first of them, co-chaired by Khalid Choukri, Joseph Mariani and Zygmunt Vetulani, was entitled "Getting Less-Resourced Languages on Board!" The rationale of this workshop was the following:

¹ We adhere to this idea with the full title of LTC which is: "Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics".

² "Integral part" means that LRL is open for all participants registered directly to all other workshops and tracks of the LTC and *vice versa*. Also acceptance procedures for LRL submissions are strictly the same as for all LTC attenders (see

"Language Technologies (LT) provide an essential support to the challenge of Multilingualism. In order to develop them, it is necessary to have access to Language Resources (LR) and to assess LT performances. To this regard, the situation is very different across the different language. Little or sparse data exist for languages in countries or regions where limited efforts have been devoted to such issues in the past, also known as Less-Resourced Languages (LRL). The workshop aims at reporting the needs, at presenting achievements and at proposing solutions for the future, both in terms of LR and of LT evaluation, especially in the European, Euro-Mediterranean and regional frameworks. This will allow to identify the factors that have an impact on a potential and shared roadmap towards supplying LR and LT for all languages." (www.ltc.amu.edu.pl/2009/).

Ten years later this rationale still remains valid.

This half-a-day workshop, open to the whole LTC audience, was organized into three parts: invited talks, presentation of technical papers, and panel discussion to summarize the workshop. The first one was given by Briony Williams (Bangor Univ., Wales and ISCA) who presented a talk "Less-Resourced Languages and Language Resources: Lessons learned from the Celtic languages of Great Britain and Ireland." The second invited lecturer was Aleksandra Wesolowska (EC, DG – Information society and media) who addressed an essential question to the concerned audience: "New start for European language technology. Are you ready?". The second part consisted in presentation of 8 technical papers (32 authors) presented by teams from Bulgaria, France, Germany, Ghana, Greece, Norway, Slovak Republic, Spain, (with contributions on Amharic, Basque, Bulgarian, Catalan, Ga, Galician, Luxembourgish, Romani). The panel discussion with the motto "*Linguists need technologists, technologists need linguists, societies and cultures need both to survive*" concluded the workshop with a number of observations that were collected in the final workshop report³:

- a strong political will to consider the language dimension and sufficient funds are necessary,
- this must go with the awareness that Language Technologies and Language Resources are essential to the development of society,
- there should be specialists in the processing of any given language, reaching a critical mass, and young researchers should be trained.
- 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 ones, if possible as Open Source),
 - metadata, annotation schemes, standards,
 - development platforms,

³ Unpublished Report on the Special joint LTC-FLaReNet session « Getting Less-Resourced Languages On-Board! »; LTC'09 Conference Poznan, 6-8 November 2009 by Joseph Mariani, Khalid Choukri and Zygmunt Vetulani.

- evaluation means (adapted to the language specificities, such as for Machine Translation of morphologically-rich languages),
- 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 favored.
- 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, and the identification of gaps and roadmaps should be aimed at.
- 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 and Sustainability.

2.2 Next meetings

LRL 2011

The 2nd LRL (A JOINT LTC-ELRA-FLaReNet-META_NET EVENT) (2011) was subtitled: "Addressing the Gaps in Language Resources and Technologies".

In the Call for papers the workshop is defined as follows: "*The workshop will draw on the inventories of all language technologies and resources that are presently being carried out, such as the ones conducted by FLaReNet, ELRA or META-NET (e.g. LRE Map, Program Surveys, Language Matrixes, Language Gaps, META-SHARE infrastructure). These are now available and help better understand the current landscape and work out the possible solutions, for each individual language and technology. The idea is to discuss availability, quality, maturity, sustainability, and gaps of the LR and LT for a number of languages and technologies.*" (www.ltc.amu.edu.pl/2011/)

This time 15 papers were presented by 30 authors from laboratories in France, India, Ireland, Italy, Japan, Norway, Spain and Switzerland. The contributions addressed the following languages (Basque, Chinese, Irish, Indian languages (12), Khmer, Luxembourgish, Magahi, Punjabi, Quechua, Vietnamese).

LRL 2013

The 3rd LRL WORKSHOP focused on new technologies appearing as a challenge for less resourced languages, as its full name was "A Joint LTC-ELRA-FLaReNet-META-NET Workshop on Less-Resourced Languages: Less Resourced Languages, new technologies, new challenges and opportunities".

The theme for this even was defined in the following way: *"Many less resourced languages (LRL) that are thriving to get a place in the digital space and that could profit of the new opportunities offered by the Internet and digital devices will seriously face digital extinction if they are not supported by Language Technologies. Language Technologies (LTs, i.e. spelling and grammar checkers, electronic dictionaries, localized interfaces, voice dictations, audio transcriptions and subtitling, as well as multimedia/multimodal search engines, language translators or information extraction tools) are essential instruments to secure usability of less resourced languages within the digital world, thus ensuring those languages equal opportunities and raising their profile in the eyes of natives but also non-natives from the younger, digitally-oriented generation. However, there are many challenges to be faced to equip less resourced languages with LTs (from basic to advanced): a substantial delay in development of basic technologies, a lack of cooperation among languages communities, a chronic shortage of funding (in particular for minority languages not officially recognized, yet often the most vital ones over the Internet) and the limited economic value placed over LTs for minority languages by the market rules. At this critical time, this workshop seeks to continue the debate as to what new technologies have to offer less resourced languages, and how the research community might seek to overcome the challenges and exploit the opportunities"* (www.ltc.amu.edu.pl/2013/).

This time Claudia Soria (CNR-ILC, Italy) joint the group of LRL co-chairs. The workshop attracted 20 authors from 7 countries (France, India, Italy, Norway, Qatar, Spain, USA), who covered by their research over 14 languages (Arabic, Bengali, Chinese, Hindi, Indian languages, Indochinese language, Italian dialects, Italian German, Malay, Malayalam, Khmer, Occitan, Sardinian, Vietnamese) in 9 presentations.

LRL 2015

The 4th LRL Workshop: "Language Technologies in support of Less-Resourced Languages" was co-chaired by Khalid Choukri, Joseph Mariani, Claudia Soria and Zygumnt Vetulani.

Expectations of the LRL co-chairs were articulated as follows:

"This Workshop is targeting all stakeholders somehow involved in Language Technology for less-resourced languages, either as users, developers, researchers, language activists, policy makers. As such, the Workshop broadly addresses current use and usability of Language Technologies for less-resourced languages. This year, we take the opportunity of celebrating the 20th anniversary of the Language and Technology Conference to analyze the influence of Language Technologies on Less-Resourced Languages over two decades. We will particularly welcome contributions addressing the following issues:

- 1) *LRLs in the digital age - how well are regional/minority/less-resourced languages equipped for the digital age? What is the experience of speakers, what are their opportunities to act in the digital sphere by means of these languages? Do speakers of regional/minority/less-resourced languages experience any kind of "unequal digital opportunity"? What is the impact of LRTs on the use and usability of LRL on digital media and devices?*
- 2) *LRTs for LRL - development of LRTs for LRLs is often linked to purposes other than availability of applications for retrieving information or for enabling communication (e.g. language learning, identity-building or language reclamation): how often*

are LRLs targeted by applications for educational, entertainment, or revitalization purposes?

3) *LRL: charting the field - what do we know about currently available LRTs for LRL? How to draw a comprehensive and accurate picture? Who are the actors to be involved? What is the experience of researchers and developers?*

4) *LRL: rethinking the BLaRK - the BLaRK still proves a useful tool for planning and implementing LT for LRL. How can it be remodeled/rethought in the light of current technological development? How can it be channeled into a coherent development roadmap?"* (www.ltc.amu.edu.pl/2015/)

The 42 authors of 14 research papers affiliated in 10 countries (Canada, Germany, India, Ireland, Italy, Madagascar, Norway, Poland, Switzerland, UK) contributed to this workshop with their papers concerning over 12 languages (African languages, Ancient Greek, Indian languages, Krio, Malagasy, Sanskrit, Scottish Gaelic, Sambalpuri, Swahili, Swiss German, Vietnamese, Welsh).

LRL 2017

The 5th edition of the Joint LTC-ELRA-FLaReNet-META_NET Workshop on Less-Resourced Languages was announced and prepared by Girish Nath Jha (JNU, New Delhi, India) and Claudia Soria (co-chairs). In order to attract and mobilize the attendees themes and motivation the suggested themes were defined in form of questions:

"LRL: charting the field - what do we know about currently available LTs for LRLs? What is the current status of language technologies and use of LRLs in the digital and social media environments? How to draw a comprehensive and accurate picture and create a road map for future? Who are the actors to be involved? What is the experience of researchers and developers?"

LRL: Resource development - how are the LRLs dealing with resource crunch, creation and related issues of standards, IDEs and platforms, funding, usability, sharing etc? What are the perceptions and roles of various stake holders including the governments, industry and language communities? What are the additional challenges posed by multilingual societies? What are the language preservation strategies for LRLs in the digital age? LRL : technology development - challenges in the development of specific enabling technologies for LRLs at language, speech and multi-modal levels. How are these technologies used in areas such as communication, education, entertainment, health, administration, governance, etc?" (www.ltc.amu.edu.pl/2017/)

The 22 authors of 7 research papers affiliated in 5 countries (Canada, Croatian, France, Japan, Poland, UK) contributed to this workshop with papers concerning languages (Awadhi, Ainu, Braj, Embosi/Bantu, French Vietnamese, Georgian).

LRL 2019

The formula of the workshop was proposed by the new team of co-chairs Dorothee Beermann (Norwegian University of Science and Technology), Laurent Besacier (Grenoble Alpes University, France) and Claudia Soria (CNR-ILC, Italy), following the creation of the joint ELRA-ISCA Special Interest Group on Under-resourced Languages (SIGUL).

This LRL workshop was different from all preceding ones. In particular because, several papers of perfect fit with traditional LRL motivations and objectives, were

accepted to other LTC tracks, mostly because their technical nature. In this number were for example papers from Central Asia, but also Georgia, Iran and Oceania. In order to make this overview complete, we will also take these contribution into account. In total we identified 15 papers whose authors (44) are affiliated in 14 countries (Australia, France, Germany, Georgia, India, Iran, Japan, Kazakhstan, Nigeria, Norway, Thailand, Spain, Uzbekistan, Vanuatu). The following languages, commonly classified less-resourced, were object of studies presented at the conference (LRL and other LTC tracks): Alsatian, Georgian, Efate language, Ibibio, Kazakh, Nefsan language, Oceanian languages, Telugu English, Persian, Thai language, Uzbek, Wolof. (www.ltc.amu.edu.pl)

2.3 Some observations

The first LRL Workshop organized in 2009 was an answer to already identified threats of technological and social exclusion on the global scale due to language barriers and scarcity of communication technologies in the world pretending to become a "global village". The response to the first call for papers in 2009 and further LRL meetings, in particular expansion of the LTC/LRL on countries so far rare at international LT events, confirmed interest in this activity (see table 1 below).

LRL 2009-2019	
Affiliation countries of the authors ⁴	29
Number of concerned languages ⁵	46
Number of presented papers	68
Number of authors	170

Table 1: LRL in numbers⁶.

We observed a number of positive phenomena. Among the LRL papers we identified interesting articles addressing existing needs and reports on research concerning languages rare at the international LT conferences (as some African and Oceanian languages, but also dialects and regional languages in Europe as

⁴ Australia, Bulgaria, Canada, Croatia, France, Georgia, Germany, Ghana, Greece, Hungary, India, Iran, Ireland, Italy, Japan, Kazakhstan, Madagascar, Nigeria, Norway, Poland, Qatar, Serbia, Slovak Republic, Spain, Switzerland, Thailand, UK, Uzbekistan, Vanuatu.

⁵ Ainu, Alsatian, Amharic, Ancient Greek, Arabic, Basque, Bengali, Bulgarian, Chinese, Croatian, Early Braj, Embosi(Bantu), Ga language, Georgian, Hindi, Ibibio, Indian Languages, Indonesian, Irish, Kazakh, Khmer, Krio, Kwa languages, Luxemburgish, Magahi, Malagasy, Malay, Malayalam, Nefsan, Occitan, Persian, Pinyin for Taiwanese, Punjabi, Quechua, Romani, Sambalpuri, Sanskrit, Scottish Gaelic, Swahili, Swiss German, Telugu-English, Thai, Uzbek, Vietnamese, Welsh, Wolof.

⁶ Figures in Table 1 are underestimated. As the LRL workshop are fully integrated with the rest of LTC, several contributions on low resourced languages were presented in specialized thematic sessions. For LRL 2009-2017 we report here only contributions directly addressed to the LRL and ignore other, submitted to thematic sessions.

Occitan and Romani, or, last but not least, historical languages as Ainu or Ancient Greek).

One may expect that any given language, as LT works on this language progress, will not be considered "less-resourced" anymore and will stop being discussed at the LRL workshops. Considering, however, the situation where, according UNESCO, at least 2500 languages in the world is considered endangered, the need of that kind of initiatives will not fizzle out soon.

3. Conclusions

Started 10 years ago, the Less Resourced Languages workshop series has seen both its size and its form undergo various changes depending on the evolution of Natural Language Processing. The huge amount of data, computing power and ubiquitous presence of computers in an ever increasing number of activities requiring innovative applications, combined with the growing economic and cultural pressure of today, make the situation more and more critical for many languages that could disappear in a near future. Their survival lies in part in the existence of events like the LRL workshop series. The initiators and past contributors, we should thank for their support and involvement, hoping that new people will join us in the future in our quest to make Less-Resourced Language become "Better-Resourced Languages".

The last day of the LTC 2019, in our informal meeting already after the conference closure session, we discussed the proposal of establishing an international structure for LT to continue the LTC vocation with particular attention to Less-Resourced and Endangered Languages.

4. Acknowledgements

Thanks to all people who were involved in the workshop series – authors, panelists, PC members, reviewers and debaters – for their contribution to the awareness mission of the LRL events.

5. Bibliographical References

- Mariani, J., Paroubek, P., Vetulani, Z. (2009). Report on the Special joint LTC-FLaReNet session « Getting Less-Resourced Languages On-Board!»; LTC'09 Conference Poznan, 6-8 November 2009 (Unpublished)
- Mariani, J., Francopoulo, G., Paroubek, P., Vetulani, Z. (2015). Rediscovering 10 to 20 Years of Discoveries in Language & Technology. In Vetulani, Z. and Mariani, J., editors, Proceedings of 7th Language & Technology Conference, Human Language Technologies as a Challenge for Computer Science and Linguistics, Poznań, Poland, ISBN 978-83-932640-8-7, pages 29-47.
- Language and Technology Conferences web site (1995-2019). <http://www.ltc.amu.edu.pl/>.