Dwarfs Sitting on Giants' Shoulders: How LTs for Regional and Minority Languages Can Benefit from Piggybacking on Major Languages

Claudia Soria

Istituto di Linguistica Computazionale "A. Zampolli", Consiglio Nazionale delle Ricerche
Via Moruzzi 1, 56124 Pisa, Italy

[claudia.soria@ilc.cnr.it]

Joseph Mariani

Institute for Multilingual and Multimedia Information (IMMI), LIMSI-CNRS
B.P. 133, 91403 Orsay, France

[Joseph.Mariani@limsi.fr]

Carlo Zoli

Smallcodes S.r.L.
Via del Campuccio 118, 50125 Firenze, Italy

[carlo.zoli@smallcodes.com]

Abstract

LTs (language technologies) are necessary instruments for all languages, especially for those aiming at conquering a space in digital devices. Languages that are not seriously equipped with LTs face digital extinction in the long run. Many challenges are to be faced to equip minority languages with LTs (from basic to advanced): an almost complete lack of knowledge about available resources and technologies; substantial delays in development of basic technologies; lack of cooperation among minority languages communities; a chronic shortage of funding (in particular for minority languages not officially recognized, which are often the most vital ones on the Internet); and the limited economic value allotted to LTs for minority languages by digital market rules. In this paper we show how these challenges can be overcome, and how coordinated and standardized cooperation among all interested stakeholders can lead to better knowledge and awareness of the breadth and depth of available technologies.

Résumé

Les technologies langagières sont des instruments indispensables pour toute langue, mais surtout pour celles qui visent à se procurer un espace dans les appareils numériques. Des langues qui ne sont pas bien équipées des technologies langagières sont confrontées, à long terme, à une disparition numérique. De nombreux défis sont à adresser pour équiper les langues minoritaires des technologies langagières, telles que (rangés d'élémentaires à avancées): l'absence presque catégorique de connaissance des ressources et technologies disponibles; des retards considérables dans le développement des technologies de base; un manque de coopération parmi les communautés linguistiques minoritaires; l'insuffisance chronique de fonds (particulièrement pour les langues minoritaires qui ne sont pas officiellement reconnues, bien qu'elles soient parmi les plus importantes sur l'internet); et la valeur économique limitée attribuée aux technologies langagières pour les langues minoritaires de la part des règles du marché numérique. Dans cet article, nous montrons comment ces défis peuvent être surmontés, et comment une coopération coordonnée et standardisée peut entraîner une meilleure connaissance et conscience de l'étendue et de la profondeur des technologies disponibles.

Introduction

The digital revolution we are witnessing has opened up an entirely new context of uses for small (minority, regional and/or endangered) languages. Small languages are now using social media, YouTube, text messaging and various technologies to expand their voice and expand their presence. More and more linguistic groups turn to the web as a powerful instrument for preservation and revitalization of their languages. Thousands of people blog and "tweet" in their mother tongue and use Facebook in their own languages. The web has become a fantastic opportunity for minority languages, as it offers publishing opportunities at virtually no cost, and under no control. The web is giving back to minority languages the space that was denied to them by other media. Virtual online communities are doing – with modern means – what was done in the past by language activists who organized rallies and meetings. They are reclaiming a voice and a space for their languages.

Facebook, for example, is a wonderful venue for endangered and minority languages communities, especially since it makes those languages attractive in the eyes of the younger generation. Social media have become very powerful for language revitalization. From Facebook to voice recognition to interactive learning tools, both small and large communities, private and public, can subvert the tools of daily modern life to keep a language alive.

The new digital tools offer a way back from the brink for a lot of languages that seemed doomed just a few years ago. There are numerous examples of how new media and digital technologies can help in saving

moribund or endangered languages: North American tribes use social media to re-engage their young; there is an iPhone app to teach new students the pronunciation of Tuvan words (an indigenous tongue spoken by nomadic peoples in Siberia and Mongolia); an app for Tusaalanga Inuktitut is being developed as a resource for learning several Inuktitut dialects.

There is no doubt about the absolute importance of the role of Information and Communication Technologies (ICTs) for small languages. They will show young people that their community language is up and running, and that it can express all kinds of concepts. They will show that any language is fully apt and suited for computing and modern life, not only big languages such as English, French and Spanish. Thus, ICTs are powerful mechanisms for building pride in the language. Internet and, in general, ICTs will help spread awareness about the problem of language endangerment and preservation, in a capillary*** way. ICT is not to be seen as a threat, but rather, as an opportunity for a small language to extend its voice and to reach a global audience. Small languages can and must profit from the incredible opportunity offered by current technologies.

However, **small languages need to be given the voice**, in technological terms. There are several challenges, ranging from digital divide and connectivity access, problems in terms of scripts and their digital encoding, lack of terminology, etc. to availability and development of language technologies. In this paper we concentrate on language technologies: their availability, development, and how to cope with the limitations imposed by a market that sees very little economic value in them and therefore hinders their development.

Language Technologies

LTs[21] (language technologies: spelling and grammar checkers, electronic dictionaries, localized interfaces, as well as search engines, automatic speech recognition and synthesis, language translators and information extraction tools) are a necessary instrument in securing the usability of minority languages over the web, thus ensuring equal digital opportunities for those languages and raising their profile in the eyes of the younger, digitally-oriented generation.

If we accept that modern ICTs are indeed an opportunity for small languages, we must recognize that on the other hand they constitute a big challenge, as they require fast development of high quality Language Technologies to keep up with the pace of technological development. In other words, ICTs will help minority languages to gaining a place in the digital space as long as good and effective language technologies are developed and integrated into ICTs: if a language is not adequately supported by language technologies, its use over the Internet and through digital devices becomes cumbersome, communication is difficult, and usability is dramatically affected. Development of LTs is an important – in fact, critical – part of language preservation and revitalization.

If we want to save and preserve minority languages, we must necessarily let these lesser-used languages have access to the tools and resources of the same technological level as those of "bigger" languages. The moment is now: if we don't act quickly and effectively now, if carefully planned and focused intervention is not immediately carried out, it might be too late. Many challenges are to be faced to equip minority languages with LTs. These are, from basic to advanced: the almost complete lack of knowledge about available resources and technologies; the substantial delay in development of basic technologies; the lack of cooperation among minority languages communities; the chronic shortage of funding (in particular for minority languages not officially recognized, which are often the most vital ones over the Internet); and the limited economic value allotted to LTs for minority languages by digital market rules. How can these challenges be overcome? This can only be done by sharing expertise, experience, and costs among minorities, and adopting a minimum set of strategies. To avoid fragmentation, and useless dispersal of human and financial resources, the rapid development of awareness about the importance of a collaborative approach to the development of LTs for small languages – one where small languages can benefit from the experience and technological development already reached by major languages – is of utmost importance. Using a metaphor, we define this approach as "dwarfs sitting on the shoulders of giants".

In the rest of this paper we will detail some practical recommendations concerning the development of LRTs for small languages, with a focus on the specific constraints and concerns posed by minority languages.

An Agenda for LRTs for Small Languages

Connect and Cooperate

More than anything else, joining forces and building a compact community is essential for a rapid and effective development of LTs for small languages. All involved players, regardless of the particular discipline they belong to, would take enormous advantage from networking, collaborating and coordinating efforts. An international forum acting as an aggregator and facilitator of information sharing and discussion – bringing together leading experts of research institutions, academia, companies, consortia, associations, and individual language activists – would be beneficial for promoting and sustaining international cooperation, facilitating at the same time the creation of a compact community around LTs. Ideally, this forum should be aimed at a) providing information about

[21] In this paper we use the term Language Resources (LR) to refer to data (e.g. lexicons, thesauri, corpora, grammars) and Language Technologies (LT) to refer to software applications (e.g., orthography checkers, automatic summarizers, document indexing tools, machine translators, information retrieval tools, etc.). When referring to both sets, the term Language Resources and Technologies (LRT) is used instead.

already available resources and technologies, b) promoting cooperation among developers and users, and c) offering a place for reciprocal help and sustainment.

Later on, this network of players can evolve into a facility (under the form of a repository or web portal) for discovering, accessing and sharing actual data and tools. A few are already available (e.g., META-SHARE[22], ELRA UC[23], CLARIN[24]), and the suitability of these to the needs of small languages should be evaluated before embarking on a new enterprise.

At the same time, such an infrastructure could and should be backed up by international cooperation initiatives. Cooperation among countries and programs is essential to drive the field forward in a coordinated way and avoid duplication of efforts and fragmentation. It is crucial to discuss future policies and priorities for the field of LRTs for small languages on a global scale. This is true both when we try to highlight future directions of research and – even more so – when we analyze which infrastructural actions are needed. The growth of the field must be complemented by a common effort that looks for synergies and overcomes fragmentation.

Use Standards

Use of standards is the key to interoperability of resources, as they allow resource sharing, re-usability, maintainability and long-term preservation. The availability of data represented in a standardized format is an essential prerequisite for successful exploitation of the already available data, making it possible, for instance, to merge different corpora or lexicons, or to build multilingual lexicons from two monolingual ones. So, standards are the key to resource reusability, which in turn is one of the pillars of an effective development strategy for LTs for small languages.

It is essential that small language resource developers consistently use and adopt standards for representation and encoding of language resources. Yet this is an aspect often overlooked by developers, who are not willing to spend their often limited resources in an effort whose utility they can't see. This triggers a vicious cycle: LT developers do not use standards because they find it a useless burden; no standardized data are available; no proof of usefulness of standards can be shown. Still, the work of several scholars (see, for instance, Bel et al., 2011; Tokunaga et al., 2011) clearly shows the benefits of adopting standardized representation formats for deriving brand new resources, in particular for lesser used languages (Enguehard & Mangeot, 2013). Lexical Markup Framework for the representation of lexicons, TEI for representation of annotated corpora and, for instance, TBX for terminologies, are only some of the standardized representation formats that can be adopted and fully enjoy the support of the LRT community (for specific details, see Monachini et al., 2011).

The community needs to act compactly and join forces to drive forward the use of existing and emerging standards, at least in the areas where there is some degree of consensus (e.g., external descriptive metadata, meta-models, part-of-speech (POS) and morpho-syntactic information, etc.). By the way, the only way to ensure useful feedback to improve and make progress is to use these standards on a regular basis. It will be even more important to enforce and promote the use of standards at all stages, from basic standardization for less-resourced languages (such as orthography normalization, transcription of oral data, etc.) to more complex areas (such as syntax, semantics, etc.). Providers of language resources, on their side, should look for standards and best practices that best fit the LRs to be produced, already at the early stages of design/specifications; adhere to relevant standards and best practices; produce LRs that are easily amenable to reuse (e.g., adopt formats that allow easy reuse). It is to be said, however, that facilitating mechanisms should be put in place to make it easy for resource developers to use standards, or to convert resources from proprietary formats to standardized ones. This could be offered as a community service by an infrastructural organization.

Document Resources and Technologies

Accurate and reliable documentation of Language Resources is an indisputable need. Instead, as of today, LRs are still often poorly documented or not documented at all. Documentation of resources is another aspect of use of standardized descriptions, and it suffers from the same problems described above. Resource and technology developers do not have the time and will to describe and document resources, and when they are over with their project they move on to the next one. However, documentation is the gateway to LR discovery. Ensuring that Language Resources are discoverable is the first step towards promoting a market for LRTs for small languages.

In those rare cases where documentation is indeed available, it is often not easy to find, or is done according to proprietary formats and terminology. This makes it difficult to search for and compare available resources.

When producing a LR, time and manpower should be allocated to documentation from the start, and documentation should always be provided (or links to it) when giving access to a LR: every release of a Language Resource should be accompanied by provision of the corresponding documentation. This should become part of everyday culture in LRT development, so as it is the acknowledgment for the funding or support received. In every language resource production project, part of the funding should be allocated to documentation and dissemination activities.

[22] http://www.meta-share.eu

[23] http://universal.elra.info

[24] http://www.clarin.eu

Moreover, it should be made sure that appropriate metadata are consistently adopted for describing LRs: standard metadata and unique resource identifiers should be used wherever possible. Therefore, definition and adoption of standardized metadata must be the first priority and first step for all LRT providers.

Documentation should be as exhaustive as possible, and include information about data format and data content, the production context, and existing possible applications.

Recommended metadata sets are CLARIN Component MetaData Infrastructure (CMDI)[25] and META-SHARE ones (Gavrilidou et al., 2012), which offer the additional advantage of being mutually compatible and modular, so allowing for minimal to maximal depth of description of resources and technologies. These metadata sets offer documentation templates for each resource type. Adoption of already available metadata sets makes a step forward to interoperability of metadata sets, which is another important feature: different sub-communities, data distribution centers, archiving institutions and projects, and other providers tend to use their own, non-interoperable metadata sets to describe their data, often at different levels of granularity, depending on who does it.

Again, we recommend the setting up of a dedicated community service to this end, so as to help spreading use and awareness of how to cut down on costs and time effort. For the specific purpose of LRT documentation, LRT developers and users could adhere to community-wide initiatives such as the LRE Map (Calzolari et al., 2010; Calzolari et al., 2012). The LRE Map is a community-based documentation initiative by which massive documentation of existing resources is achieved in a limited time frame and with limited effort, with the additional advantage that all resources are documented in a uniform and standard-compliant way. Besides being an instrument for enhancing availability of information about resources, either new or already existing ones, it is a measuring tool for monitoring various dimensions of resources across places and times, thus helping to highlight evolutionary trends in language resource use and related language technology development by cataloguing not only language resources in a narrow sense (i.e., language data), but also tools, standards, and annotation guidelines.

Be "Green": Reuse and Recycle

A new mentality should be enforced that discourages creation of new resources from scratch wherever already existing resources can be found for a given language and/or application. We should encourage a "green" attitude of re-use and re-purposing via a recycling culture that reuses development methods, existing tools, and translation/transliteration tools, etc. Well-known examples include the work of Padó & Lapata (2009), Bentivogli & Pianta (2005), Hwa et al. (2005), and Smith & Eisner (2006).

The reuse of existing approaches and resources for the specific purpose of time and cost-effective development of LRT for small languages is well demonstrated, for instance, by the work of Scannell (2006; 2012), or very recently by Sherrer & Sagot (2013). These methods illustrate an emerging and promising approach that exploits already existing resources for major languages to derive, often semi-automatically, brand-new resources for small languages. This approach is just one example of innovative production methods that massively involve automatic procedures to reduce human intervention to a minimum and consequently cut the production costs. Although the quality of so-derived resources is often questioned, it is worth remembering that in most cases, an imperfect product is far better than its complete absence, and therefore automatic or semi-automatic approaches to the development of LRTs for smaller languages should be encouraged and enforced.

Crowdsource Your Resources

Given the high cost of language resource production, and given the fact that in many cases it is impossible to avoid the manual construction of resources (e.g., if accurate models are requested or if there is to be reliable evaluation), it is worth considering the power of social/collaborative media to build resources, especially for those languages where there are no language resources built by experts as of yet. Collaborative and Web 2.0 methods for data collection and annotation seem particularly well-suited for collecting the data needed for the development of Language Technology applications for smaller and minority languages. Given the chronic lack of funding, the collaborative accumulation and creation of data appears to be the best and most practicable way to achieve better and faster language coverage and in purely economic terms could well deliver a higher return on investment than expected. Also, it is a good way to approach small populations of speakers who live in remote countries, or are disseminated in a diaspora all over the world.

There are several experiments in crowd-sourcing data collection and NLP tasks (Chen & Dolan, 2011), and most of them look promising. For instance, it has been estimated that Mechanical Turk translation is 10 to 60 times less expensive than professional translation (Callison-Bruch & Dredze, 2010). Moreover, small language communities usually show strong motivation and personal investment, making them ideal candidates for carrying out experiments in collaborative creation and annotation of resources.

However, the use of crowd-sourcing raises ethical, sociological and practical issues for the community. It is not yet clearly understood, for example, whether all types of LRs can be obtained collaboratively by using naïve annotators; more research is therefore needed on both the technical (e.g., accurately comparing the quality and content of resources built collaboratively

[25] http://www.clarin.eu/node/3219

and those built by experts) and ethical aspects of crowd-sourcing; see, for instance, Zaidan & Callison-Burch (2011) about mechanisms for increasing the quality of crowd-sourced data.

Be Open

Use of open-source software and adoption of licenses allowing for data reuse, modification and redistribution (such as Creative Commons Attribution-ShareAlike 3.0[26]) is another essential prerequisite for fostering the creation of an industry of LRT for small languages. This approach is slowly being adopted by the community of LRT for major languages, but needs to be embraced as soon as possible by developers and users of LRT for smaller languages. Reluctance to give open access to precious data that were often painfully collected is still common and understandable, but once the data will enter the distribution cycle, the advantages will soon become apparent.

This requires a cultural change in attitude, as well as serious training in Intellectual Property Rights (IPR) issues. Lack of knowledge and awareness in the issues involved in clearing IPR rights is a serious hamper to further development of the sector of LRT for small languages. Unfortunately, we do not yet have a sufficient grasp of the trans-border legal issues to support enhanced resource sharing and legally protect LRs against improper reuse, copying, modification etc. The Berne Convention for the Protection of Library and Artistic Works extends copyright protection to creators in countries other than their own, but enforcement is still a national issue and is therefore implemented in different ways. In addition to this, the availability and use of huge quantities of web data as useful resources creates a novel situation that raises further legal problems. On the one hand IPRs (especially authorship) need to be protected; but on the other they tend to restrict accessibility to and usability of language resources. The current trend is towards a culture of free/open use with less protective holders' rights. Creative Commons, for example, is one of the most widely used license models for language resources (see Google, Wikipedia, Whitehouse.gov, Public Library of Science, and Flickr). From a practical point of view, producers of language resources should try to clear IPR at the early stages of production, ensuring that re-use is permitted.

In those cases where resources are developed with public funding, they should always be made publicly available either free of charge or at a small distribution cost. For mixed-funded initiatives (private/public), it should be ensured that there is an agreement to make resources available at fair market conditions right from the start.

It is necessary to elaborate specific, simple and harmonized licensing solutions for data resources: the community should avoid one-size-fits-all solutions. There is a large number of licensing schemes already in use today; while some are backed by strong players (ELRA, LDC, open source communities such as Creative Commons, GNU General Public License, etc.), others have been drafted bilaterally and in some cases by the legal departments of data providers. It is crucial that such licensing is harmonized and even standardized. Licensing schemes need to be simplified through broad-based solutions for both R&D and industry. Electronic licensing (e-licenses) should be adopted and current distribution models to new media (web, mobile devices, etc.) should be accepted.

Share and Sustain

Sharing resources, both data and tools, has become a viable solution towards encouraging open data, and the LRT community is strongly investing in facilities for the discovery and use of resources by federated members. These facilities, such as the META-SHARE infrastructure[27], could represent an optimal intermediate solution to respond to the need for data variety, ease of retrieval, better data description and community-wide access, while at the same time assisting in clearing the intricate issues associated with IPR. Repositories, catalogues and other storage facilities are helpful for securing accessibility if LRs over the long term. This is another very important requirement and LR producers should consider relying on specialized archiving centers such as DANS[28] or the Language Archive[29] (TLA) to make sure their data are properly archived and preserved. These centers offer future-proof archiving for resources, and by delivering a standardized service which backgrounds all the technical and legal implications, represent an optimal solution for individual researchers or smaller groups. We strongly advise against adopting self-made solutions: the issue of long-term preservation of resources is a very delicate and intricate one, and professional service should be preferred over independent solutions[30].

Cooperate to Focused Development

In order to avoid reduplication of efforts, a cooperative approach should be pursued, and the community should participate in order to first document available and already existing resources, and second, to follow a strategy in the development of the missing ones. To this end, at least two instruments are necessary: first, a way to monitor the field of LRT for small languages, in the form of an updated catalogue. Knowing what is already available for any language is of foremost importance to avoid scattered, fragmented and redundant development of resources. Second, a roadmap to sensible development of needed and missing resources: once a

[26] http://creativecommons.org/licenses/by-sa/3.0/

[27] www.meta-share.eu

[28] www.dans.knaw.nl

[29] http://tla.mpi.nl/

[30] See also DANS (2010)

clear picture of already available resources is achieved, it is important to know in which direction to orient research and development.

The first objective can be easily addressed by using cataloguing tools such as the above-mentioned LRE Map. Besides being used for documenting resources, the LRE Map can be exploited for searching for available resources and inquiring about related information (e.g., degree of availability, contact details, potential uses or applications). Its contents easily lend themselves to deriving pictures of available LRTs according to language, modality or application purposes. These pictures, so-called "Language Matrices", represent a powerful discovery tool for monitoring the field and yield an updated picture of the language resources available for the various languages, modalities, or applications, provided that up-to-date and trustworthy content is delivered.

The Language Matrices can in turn be used as a guide towards implementation of Basic Language Resource Kits or BLaRKs (Krauwer, 2003). A BLaRK is the minimal set of language resources that is necessary to do any precompetitive research and education. It lists, for a given language and for several different language technologies applications, the data and software modules that represent a prerequisite for those technologies. Although, in principle, this is a language-independent concept, its instantiation heavily depends on the specific requirements of individual language. We can think of a BLaRK as a LRT "checklist": we this list in one hand an updated catalogue of the available resources in the other, it becomes possible to effectively make a development plan, prioritized according to the different needs of different languages, for endowing less resourced languages with a minimal "basic digital survival kit".

BLaRKs should be supported and developed for all languages and, at least, main applications (Machine Translation, Information Retrieval, and Question Answering, to mention a few). In this direction, first the BLaRK concept needs to be worked out in detail, so that it can be embodied as a standard, and possibly planned revision sessions should be set, as it is intrinsically a dynamic notion that changes in time with the change in technology development in the different countries. Second, regular BLaRK surveys must be conducted to produce a clear picture of technology trends, and establish (and regularly update) a roadmap covering all aspects of LTs. Third, resource production should be funded on the basis of BLaRK-like criteria, that is, giving priority to the development of "missing" resource types for each language.

Conclusions

If small languages miss the opportunity offered to them by modern digital ICT by failing to develop adequate language technologies, the divide with the more resourced language will get wider, and smaller languages will definitely cease to exist in the digital space. This would represent yet another loss in linguistic diversity, and a big harm to the profile of smaller languages, which will appear to the eye of the younger generation as failure to be modern. To avoid this, a compact community is needed to share experience and know-how. Smaller language communities can push their languages into the digital space, but they need to join forces and follow the trail of more resourced languages, profiting from the tools, resources, and experience already developed and accumulated.

References

Bel, N., Padrò, M. & Necsulescu, S. (2011). A Method Towards the Fully Automatic Merging of Lexical Resources. In *Proceedings of the Language Resources, Technology and Services in the Sharing Paradigm Workshop at IJCNLP 2011* (pp. 8-15), Chiang Mai, Thailand, 12 November 2011.

Bentivogli, L. & Pianta, E. (2005). Exploiting Parallel Texts in the Creation of Multilingual Semantically Annotated Resources: The MultiSemCor Corpus. *Natural Language Engineering* 11(3), 247-261.

Callison-Burch, C., & Dredze, M. (2010). Creating Speech and Language Data With Amazon's Mechanical Turk. In *Proceedings of NAACL-2010 Workshop on Creating Speech and Language Data With Amazon's Mechanical Turk* (pp. 1-12). Los Angeles, CA.

Chen, D. & Dolan, W. (2011). Building a Persistent Workforce on Mechanical Turk for Multilingual Data Collection. In *Proceedings of the 3rd Human Computation Workshop*. San Francisco, CA.

Calzolari N., Soria C., Del Gratta R., Goggi S., Quochi V., Russo I., Choukri K., Mariani J. & Piperidis S. (2010). The LREC Map of Language Resources and Technologies. In *Proceedings of LREC 2010* (pp. 949-956). La Valletta, Malta, 17-23 May 2010.

Calzolari, N., Monachini, M. & Quochi, V. (2011). Interoperability Framework: The FLaReNet Action Plan Proposal. In *Proceedings of the Language Resources, Technology and Services in the Sharing Paradigm Workshop at IJCNLP 2011* (pp. 41-49). Chiang Mai, Thailand, 12 November 2011.

Calzolari N., Del Gratta R., Fracopoulo G., Mariani J., Rubino F., Russo I. & Soria C. (2012). The LRE Map. Harmonising Community Description of Resources. In *Proceedings of LREC 2012* (pp. 1084-1089). Istanbul, 23-25 May 2012.

DANS (2010). Preparing Data for Sharing; Guide to Social Science Data Archiving. DANS Data Guide 8. *Data Archiving and Networked Services (DANS)*. Amsterdam: Pallas Publications/ Amsterdam University Press. http://www.dans.knaw.nl/en/content/categorieen/publicaties/dans-data-guide-8

Enguehard, C. & Mangeot, M. (2013). LMF for a Selection of African Languages. In G. Francopoulo

(Ed.), *LMF Lexical Markup Framework*. Wiley-ISTE.

Gavrilidou, M., Labropoulou, P., Desipri, E., Piperidis, S., Papageorgiou, H., Monachini, M., Frontini, F., Declerck, T., Francopoulo, G., Arranz, V. & Mapelli, V. (2012). The META-SHARE Metadata Schema for the Description of Language Resources. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 1090-1097). Istanbul, 23-25 May 2012.

Hwa, R., Resnik, P., Weinberg, A., Cabezas, C., and Kolak, O. (2005). Bootstrapping Parsers via Syntactic Projection Across Parallel Texts. *Natural Language Engineering*, 11(3).

Krauwer S. (2003). The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap. In *Proceedings of the 2003 International Workshop Speech and Computer (SPECOM 2003)* (pp. 8-15).

Calzolari, N., Monachini, M., Quochi, V., Bel, N., Budin, G., Caselli, T., Choukri, K., Francopoulo, G., Hinrichs, E., Krauwer, S., Lemnitzer, L., Mariani, J., Odijk, J., Piperidis, S., Przepiórkowski, A., Romary, L., Schmidt, H., Uszkoreit, H., Wittenburg, P. (2011). *The Standards' Landscape Towards an Interoperability Framework. The FLaReNet proposal*. FLaReNet 2011.

Padó, S. & Lapata, M. (2009). Cross-lingual Annotation Projection of Semantic roles. *Journal of Artificial Intelligence Research*. 36(1): 307-340.

Scannell, K. (2006). Machine Translation for Closely Related Language Pairs. In *Proceedings of the Workshop "Strategies for Developing Machine Translation for Minority Languages" at LREC 2006* (pp. 103-107). Genoa, May 2006.

Scannell, K. (2012). Translating Facebook into Endangered Languages. In *Proceedings of the 16th Foundation for Endangered Languages Conference* (pp. 106-110). Auckland/Aotearoa, New Zealand, 12-15 September 2012.

Scherrer, Y. & Sagot, B. (2013). Étiquetage morphosyntaxique de langues non dotées à partir de ressources pour une langue étymologiquement proche. In *Actes de TALARE 2013 : Traitement Automatique des Langues Régionales de France et d'Europe* (pp. 195—208). Les Sables d'Olonne, France.

Smith, D. A. & Eisner, J. (2006). Quasi-synchronous Grammars: Alignment by Soft Projection of Syntactic Dependencies. In *Proceedings of the Workshop on Statistical Machine Translation* (pp. 23-30). New York, NY.

Tokunaga, T., Lee, S., Sornlertlamvanich, V., Shirai, K., Hsieh, S. & Huang, C. (2013). LMF and Its Implementation in Some Asian Languages. In G. Francopoulo (Ed.), *LMF Lexical Markup Framework*. Wiley-ISTE.

Zaidan, O. & Callison-Burch, C. (2011). Crowdsourcing Translation: Professional Quality from Non-Professionals. In *Proceedings of ACL-2011.*