The LREC Workshops Map

Roberto Bartolini, Sara Goggi, Monica Monachini, Gabriella Pardelli
Istituto di Linguistica Computazionale del CNR "Antonio Zampolli"
Via Giuseppe Moruzzi 1 – 56124 Pisa, Italy
{roberto.bartolini, sara.goggi, monica.monachini, gabriella pardelli}@ilc.cnr.it

Abstract

The aim of this work is to present an overview of the research presented at the LREC workshops over the years 1998-2016 with the aim to shed light on the community represented by workshop participants in terms of country of origin, type of affiliation, gender.
There has been also an effort towards the identification of the major topics dealt with as well as of the terminological variations noticed in this time span. Data has been retrieved from the portal of the *European Language Resources Association* (ELRA) which organizes the conference and the resulting corpus made up of workshops titles and of the related presentations has then been processed using a term extraction tool developed at ILC-CNR.

Keywords: corpus creation, terminology, LREC

1. Introduction

Over the years the increasing availability of online open access documentation has permitted to gather information for various types of analysis. For this work we decided to retrieve the information about all workshops organized as satellite events of the *International Conference on Language Resources and Evaluation* (LREC) in its ten editions (1998-2016): data has been retrieved from the portal of the *European Language Resources Association* (ELRA) which organizes the conference.
The aim is to monitor the research in the field of Language Resources (LRs) and Language Technology (LT) presented at LREC workshops and describe the thematic and terminological trend over these 18 years. In the introductory pages to the Proceedings of LREC 2000, Antonio Zampolli wrote: "The workshops, whose Proceedings are published in separate volumes, have contributed greatly to the scientific relevance of the Conference, both for the quality of the work and the choice of topics." (Zampolli, 2000).
During the ten editions of the conference its satellite workshops have witnessed the participation of 6039 researchers from both academy and industry, coming from all continents. This heterogeneity results in a authorship which is composite for country, affiliation, sector and also gender; as for the latter, the female representation is relevant, as well as the percentage of single authors.

2. Some Remarks on Topics

The topics dealt with and the papers related to these topics are the focus of the work: we created a corpus (WS-CORPUS) made of the titles of all the workshops and their related papers presented at LREC since 1998. In ten editions of the conference 206 workshops have been organized around the most varied subjects, some of them as single events but many ended up constituting a "series" over the years.

The workshop on Sign Language, for example, has been organized at LREC since 2004 by the same researchers; while a workshop on the so-called "minority/less-resourced/under-resourced" languages[1] is always present at LREC but the organizing teams have been different over this time span: in the first Granada edition a workshop was organized by researchers working on the development of resources for the indigenous minority languages of Europe ("Language Resources for European Minority Languages"); since then Workshops on this topic have always been organized at LREC but it is worth mentioning – and investigating - the terminological shift from the term "minority languages" used in 1998, 2000 and 2004 to "less-resourced languages" in the three following editions and finally to the term "under-resourced language" which establishes itself in 2014 and is used again in 2016. Research in this field range from the creation of LRs (in the first two editions but again in 2010), to the development of the needed language technologies up to machine translation (in two editions) for these specific languages.
In Granada was also organized a workshop titled "Speech Database Development for Central and Eastern European Languages" with the aim of promoting speech technologies in Eastern Europe.
This theme will appear again a decade later at LREC 2010 where a workshop titled «Exploitation of Multilingual Resources and Tools for Central and (South) Eastern European Languages» was organized: to be noted the additional geographical specification (*South*) for a better definition of the linguistic area and the substitution of the outdated term *database* with *resources*.
«The same holds for well established or emerging linguistic knowledge representation frameworks, which can only benefit from embedding components for Central, Eastern and Southern European languages» (Stelios Piperidis et alii, 2010).

[1] From the English Oxford Living Dictionaries: "A language spoken by a minority group, if different from that of the majority. Origin 1920s; earliest use found in American Journal of International Law".
https://en.oxforddictionaries.com/definition/minority_language

Still focusing on languages, at LREC 2012 in Istanbul the first workshop on Indian languages is organized («WILDRE –the first 'Workshop on Indian Language Data: Resources and Evaluation»); and WILDRE will take place in the next two editions as well, in Reykjavik and Portoroz.
A special focus on Arabic and more generally on Semitic languages dates back to the 2002 edition in Las Palmas and covers most of following editions where also two tutorials took place (in 2006 and 2008).
Specifically for Arabic, the last two LREC editions the same group of researchers organized the OSACT workshop («Open/Free–Source Arabic Corpora and Processing Tools»), adding the social media theme to the 2016 edition.
Invitations for submission on the topic related to speech technology and African languages converge in the «Second Workshop on African Language Technology (AfLaT)» presented at LREC 2010 in Malta while already in 2006 a first call had been launched for developing a network aimed at cooperation in the development of resources and tools for the African languages («Networking the development of language resources for African languages»).

3. WS-CORPUS Building and Method

The process of corpus building and analysis is split up in three steps: 1) creation of the corpus by acquiring the titles of all the LREC satellite workshops and their related papers; 2) data cleaning and processing using a NLP tool; 3) terminological analysis and comparison.
Data has been retrieved from the ELRA-ELDA portal where the Proceedings of all the LREC editions - except for the 1998 one – are stored (http://www.elra.info/en/lrec/proceedings/).
The workshops organized over the 18 years are 206 (while tutorials are 46), for an overall total of 2250 presentation titles which build up our small WS-CORPUS.
In the pre-processing phase data have been annotated for having detailed information on the authors' profile: a) year of the workshop, b) title of the workshop, c) title of the paper, d) author(s)'s name and surname, e) author(s) gender, f) author(s)'s affiliation, g) country, h) ISO code.

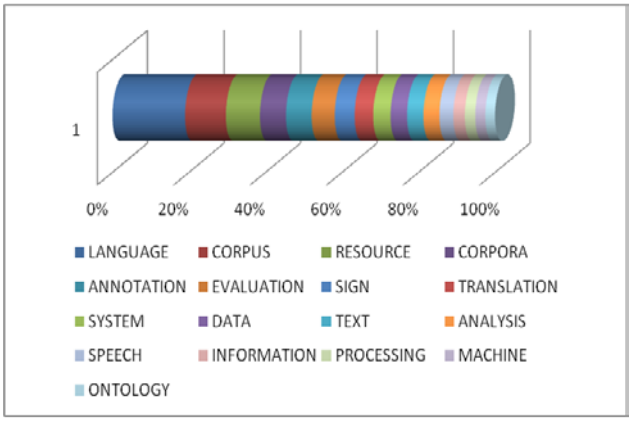
The corpus was then processed using a tool for term extraction (Goggi et al. 2015; 2016): this is a "pipeline" of different tools which extracts lexical knowledge from texts; in short, a rule-based system tool for knowledge extraction and document indexing.
The tool analyzes textual data and its result is an annotated text that allows for terminological extraction of relevant concepts. Within the WS-CORPUS it extracts a list of single (monograms) and multi-word terms (bigrams and trigrams) ordered by frequency with respect to the context.

3.1 Terminological Analysis

In order to correctly comment on the information retrieved from the corpus we should extract the most connotative features of the domain from a terminological point of view; this analysis will allow us to monitor the thematic trends as they appear and disappear from the LREC scene.
The connotative strength of words is not the same for all of them and their frequency is not the only factor to be taken into account for weighting the importance of the single terms in our corpus: as a matter of fact, also terms with just one occurrence within the corpus (hapax) might have a significant value once properly contextualized.

3.2 Terms Frequency

We start with the analysis of the most frequent monograms occurring in the corpus which will give a first terminological overview: starting from the most obvious ones such as *language*, *corpus*, *resource*, *corpora*, *annotation* and following with *evaluation*, *sign*, *translation*, *system*, *data*, *text*, *analysis*; terms with less than a hundred occurrences are *speech*, *information*, *processing*, *machine*, *ontology*. Apart from the general terms like *language*, *resource*, *data* and *text*, some domain-related terms can be retrieved among the ten most frequent ones: *corpus*, *corpora*, *ontology*, *system*, *machine*, *processing*, *annotation*, *evaluation*, *translation*.



Graph 1: Terms Frequency

3.2.1 Languages

The high frequency of the term "language" – rather obvious in itself – is reflected in the total number of spoken languages dealt with in the 200 LREC workshops: there are about sixty languages mentioned in the corpus, some of them with just one occurrence (hapax): <ALGERIAN ARABIC>, <ALSATIAN>, <ARANESE>, <ASTURIAN>, <BASHKIR>, <BELGIAN>, <BENGALI>, <BRETON>, <CREOLE>, <CROATIAN>, <IKOTA>, <JAPANESE> ,<KOMI>, <MAITHILI>, <MALAY> <MARTHI>, <NENET>, <NEPALI>, <OCCITAN>, <SAMI>, <SOMALI>, <SWAHILI>, <TAJIK>, etc.. These hapax might be substantially important because are the expression of an interest in developing LRs, corpora, tools, standards and infrastructures also for the less-studied languages cited above.
Research on a given language could be influenced by the venue where a conference takes place: the most significant example is the edition of Marrakech 2008 where the term "Arabic" reaches the highest number of

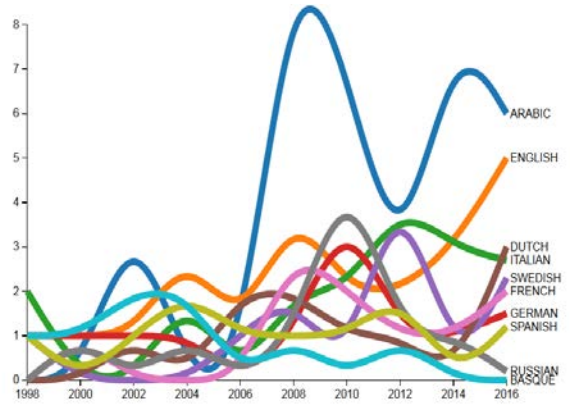
occurrences thanks to the organization of two events (a workshop and a tutorial) on the processing of dialects and local languages. And a workshop on Turkic languages has been organized only in Istanbul in 2012. But that this does not apply to all languages given that, for example, the workshop on Indian languages, firstly organized in Istanbul in 2012 has then been presented again both in Iceland and Slovenia proving then to be unrelated to the geographical location of the conference.
Graph 2 represents the temporal trend of the first ten languages retrieved from the corpus.

3.2.2 Sign Language

The survey on languages includes Sign Language (SL) as well: research on SL has been regularly presented at LREC since 2004 thus becoming a feature of the conference and constituting a "series" of seven consecutive editions. Within our corpus we can identify research on sixteen European and non-European SLs: ARABIC SIGN LANGUAGE - ArSL (2010); AMERICAN SIGN LANGUAGE - ASL (2004, 2010, 2012, 2014); BRITISH SIGN LANGUAGE (2010); CHINESE SIGN LANGUAGE - CSL (2004, 2010); FINNISH SIGN LANGUAGE - FinSS (2010, 2014, 2016); FRENCH SIGN LANGUAGE - LSF (2006, 2014,2016); GERMAN SIGN LANGUAGE - DGS (2010, 2012); GREEK SIGN LANGUAGE- GSL (2004); ITALIAN SIGN LANGUAGE- LIS (2004, 2006, 2010, 2014); HONG KONG SIGN LANGUAGE - HKSL (2008); SLOVENIAN SIGN LANGUAGE - (2016); SWEDISH SIGN LANGUAGE - STS (2012, 2014, 2016); TURKISH SIGN LANGUAGE - TİD (2014); SPANISH SIGN - LSE 2004, (2010); CZECH SIGN LANGUAGE - CzSL (2010); RUSSIAN SIGN LANGUAGE - RSL (2010).

3.2.3 Minority, Less-resourced and Under-resourced Languages

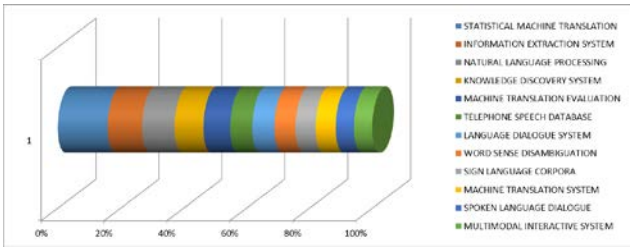
Workshops on this topic have been organized at LREC since the 1998 edition but what is worth mentioning – and investigating - is the terminological shift from the term "minority languages" used in 1998, 2000 and 2004 to "less-resourced languages" in the three following editions and finally to the term "under-resourced language" which establishes itself in 2014 and is used again in 2016.
In 2002 the focus of the workshop titled *Portability Issues in Human Language Technology (HLT*) is slightly different: «The primary objective of the workshop is to bring together participants from academia and industry to discuss and disseminate the current state of the art in multilingual research and development in the context of cross-language HLT transfer» .[2]
On the workshop website, organizers further specified the motivation behind the event: «There are more than 6000 languages in the world, yet only a small number possess the resources required for implementation of Human Language Technologies (HLT). This imbalance in technical resources available to languages of the world is likely to result in a significant linguistic divide that further exacerbates global social and economic inequities unless decisive action is taken relatively soon. One potential means of ameliorating this imbalance in technology resources is through encouraging research in the portability of human language technology for multilingual application».
Research in this field range from the creation of LRs (in the first two editions but again in 2010), to the development of the needed language technologies up to machine translation (in two editions) for these specific languages.



Graph 2: Languages

3.2.4 Trigrams

The analysis of the most frequent trigrams extracted from the corpus provides a list of the recurring topics: from statistical machine translation to multimodal systems, from machine translation evaluation to information extraction systems and so on.
Graph 3 shows the trigrams which identify the most investigated domains.



Graph 3: Trigrams

4. A case-study: WordNet

«WordNet® is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be navigated with the browser (link is external). WordNet is also freely and publicly available for download. WordNet's structure makes it a useful tool for computational linguistics and natural language processing»: this is the definiton of the resource available on the Princeton University website[3].

[2] http://www.lrec-conf.org/lrec2002/lrec/wksh/Portability.html

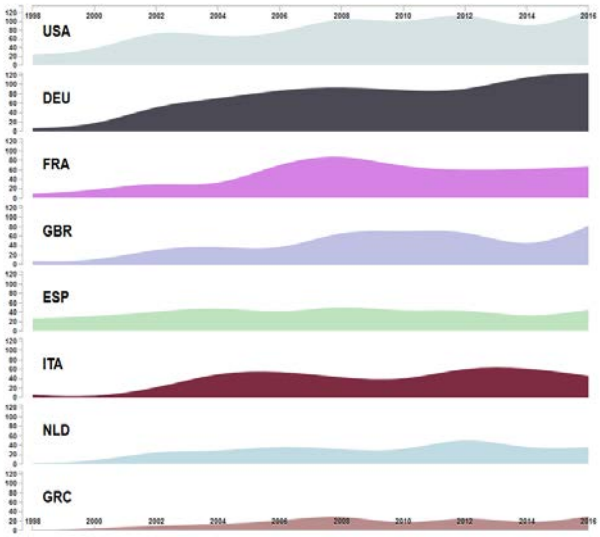
[3] https://wordnet.princeton.edu/

The WordNet resource is one of the most widely used all over the world since the early 90s and many related projects and extensions have been developed over the years: just to mention a few, in addition to the many WordNets in languages other than English, there is «SentiWordNet», a lexical resource for opinion mining; «BabelNet», a very large multilingual semantic network with millions of concepts; «FrameNet», a lexical database similar and referring to WordNet; the Lexical markup framework (LMF) which is an ISO standard for defining a common standardized framework for the construction of lexicons.
Unexpectedly just one workshop on WordNet has been organized at LREC, precisely at the 2002 edition: *Wordnet Structures and Standardization, and how these affect Wordnet Applications and Evaluation*.
In the Call for Papers of the workshop, the organizers state the importance of WordNet as follows: «During the last decade, WordNet has become a powerful resource in (computational) linguistics for various language processing tasks as well as for theoretical research issues. Due to the success of Princeton WordNet, numerous wordnets for further languages have been built or just been started. Therefore, guidelines and principles for comparing existing and acquiring new languages are of utmost importance in the field. Specific lexical properties of new languages should be accounted for. Wordnet developers of less-studied languages can profit from the experience made by the wordnet pioneers, and may also benefit from the feedback provided by wordnet appliers. The workshop will constitute a forum for sharing a common wordnet structure across languages».[4]
Papers on WordNet have obviously been presented at other Workshops during these ten editions and are spread over topics such as lexical semantics, computational lexicography, ontology, information extraction, machine translation, sentiment, annotation, semantic web, linked data, etc. We took note of a total of 38 papers related to WordNet presented at 21 workshops.
From the works presented it is possible to retrieve WordNets for the following languages: *Catalan WordNet, Galician WordNet* (1998); *Hungarian WordNet – Balkanet, Romanian WordNet, Estonian WordNet and GermaNet* (2002); *Czech WordNet - Prague Dependency TreeBank* (2004); *French WordNet* (2008); *Arabic WordNet – YAGO ontology, Estonian WordNet* (2010); *Slovene WordNet, Croatian WordNet* (2012); *IndoWordNet - Indian languages from Indo-Aryan, Dravidian, Quranic Arabic WordNet, Irish language - WordNet Gaeilge* (2014); *Konkani SentiWordNet* (2016).

5. Workshops Community

5.1 Community and Country

This paragraph is dedicated to give a first idea about the profile of LREC Workshops authors.
The inter-disciplinary dimension, the specialized themes and the geographical dislocation of its stakeholders are the requisites of attraction of the satellite LREC workshops community: over the years universities, research centers, governmental bodies and industries presented their own research experiences, the technological solutions tested and/or adopted thus facilitating the introduction of new paradigms as well as the giving up of obsolete models.
During these 18 years, 6039 researchers «participated» in the LREC workshops: 4116 from Universities, 402 from industries, 159 from Academies of Science and the rest from National Reserch Bodies (as the French CNRS or the Italian CNR), private research centers, foundations, governmental institutions, national or international organizations from all over the world. Coming to the geographical representativeness, Graph 4 illustrates the participation of the first ten countries of the ranking.
It must be said that «participation» does not necessarily means «attendance»; that is, we have records from the workshop proceedings and not from the list of actual participants to the event. There will surely be a discrepancy between these two figures, meaning that real attendance is much lower than the figures about authors coming from the published papers might suppose.
On the other hand, the following are the countries represented only once: Argentina (2012); Chile (2016); Cuba (2008); Cyprus (2012); Ethiopia (2012); Latvia (2012); Pakistan (2008); Thailand (2000); Vietnam (2012); Zambia (2006). Two countries record two occurrences (that is, two authors): Macao in 2004 and Colombia in 2016.
As stated above, the presence of two authors from these countries does not automatically means that they actually attended the workshops.



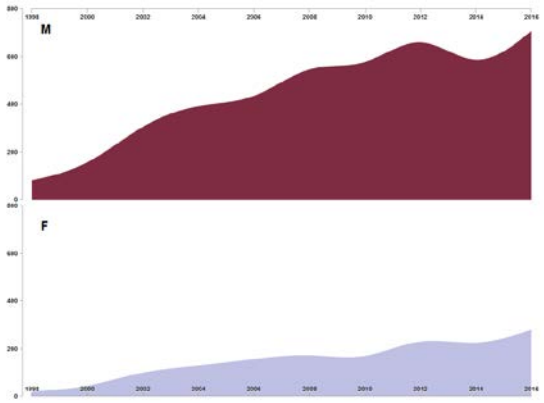
Graph 4: Countries

5.2 Community and Gender

In addition to the nationality of authors, we decided to extrapolate the information on their gender as well: this type of analysis is usually difficult due to the various ways of writing the names (full name, initials, middle initials). It was therefore needed a cleaning process for

[4] http://www.lrec-conf.org/lrec2002/lrec/wksh/Wordnets.html

being able to divide the authors by gender: for disambiguating the initials of the first names we used portals such as ACL Anthology, LREC Proceedings – Section author, in alphabetical order for all editions), Scopus, ISI WOS, Google Scholar Citations and social networks such as LinkedIn.
A few unresolved cases have been annotated with a question mark. Graph 4 shows the participation by gender to LREC Workshops: of course the names of those who participated to more than one edition – or even presented more than one paper at the same edition - have been counted only once. The results talk about a sort of clear preponderance of men: 1528 female participants vs 4435 male (76 are the unidentified authors).



Graph 5: Gender

6. Conclusions

This preliminary analysis of the WS-CORPUS gives a partial terminological overview of the research presented at LREC workshops; further investigations will have to be performed in order to being able to provide a diachronic view on the evolution of the topics treated, for highlighting the new ones which have emerged and those which rather disappeared from the scene.
In these 18 years many topics have been dealt with in these workshops: some major themes have been faced regularly over the conference editions, like for example the various aspects of multimodality (workshops at LREC have been organized since 2002), or issues related to standards and interoperability. Some specific topics such as emotion and sentiment analysis or biomedicine started being investigated at LREC 2006 and since then have been regularly presented; and themes such as application of visualization tools to LRs, controlled natural language applications and NLP applications to the Digital Humanities have recently emerged.
This study cannot be considered an exhaustive review of the field seen through the lenses of the LREC workshops because for the time being just a few features have been taken into account. More thorough analyses can be carried out both on the research side for detecting the thematic trends and on the "societal" side for better defining this heterogeneous community constituted by the scholars who participated to the numerous LREC workshops.

7. Bibliographical References

Benjamin M., Radetzky P. (2014). Multilingual Lexicography with a Focus on Less-Resourced Languages: Data Mining, Expert Input, Crowdsourcing, and Gamification. *Proceedings of CCURL 2014 Workshop: Collaboration and Computing for Under Resourced Languages in the Linked Open Data Era - Ninth International Conference on Language Resources and Evaluation* (LREC'14). Reykjavik, Iceland. European Language Resource Association (ELRA), Paris. Pages 9-16.

Cunliffe D., Herring Susan C.(2005). Introduction to Minority Languages, Multimedia and the Web. *New Review of Hypermedia and Multimedia*, 11(2), 2005. Taylor & Francis. Pages 131-137.
doi = {10.1080/13614560512331392186}.

Francopoulo G., Mariani J., Paroubek. P. Vernier F. (2016). Providing and Analyzing NLP Terms for our Community, *Proceedings of the 5th International Workshop on Computational Terminology (Computerm2016)*, Osaka, Japan. The COLING 2016 Organizing Committee. Pages 94–103, http://www.aclweb.org/anthology/W16-4711

Goggi et al. (2015). Marine Planning and Service Platform (MAPS): An Advanced Research Engine for Grey Literature in Marine Science. *The Grey Journal* Volume 11: 3 (2015), ISSN: 1574-1796.
Also in D. Farace and J. Frantzen (Eds.), Proceeding of the *Sixteenth International Conference on Grey Literature Grey Literature Lobby: Engines and Requesters for Change, GL'16,* (Library of Congress Washington D.C., USA (GL-conference series, ISSN 1386-2316, Volume 16). TextRelease, Amsterdam, 2015. Pages 108-115.

Goggi S., Pardelli G., Bartolini R., Frontini F., Monachini M., Manzella G., De Mattei M. e Bustaffa F. (2016). A semantic engine for grey literature retrieval in the oceanography domain, *The Grey Journa*l, (12), 3, pp.155-161, ISSN: 1574-1796. Also in Dominic Farace, Jerry Frantzen (Eds.) (2016), Proceedings, Seventeenth International Conference on Grey Literature. A New Wave of Textual and Non-Textual Grey Literature, Amsterdam, TextRelease. Pages 104-111.

Johnston T. (2010). Adding value to, and extracting of value from, a signed language corpus through secondary processing: implications for annotation schemas and corpus creation. *Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies. - Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA), Paris. Pages 137-142.

Mariani J., Paroubek, P., Francopoulo G., Hamon O. (2016). Rediscovering 15 + 2 years of Discoveries in Language Resources and Evaluation. *Language Resources and Evaluation*, 50(2):165–220.

Mapelli, V., Arranz, V., Carré, M., Mazo, H., Mostefa, D., and Choukri, K. (2012). ELRA in the heart of a cooperative HLT world. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation* (LREC'12). Istanbul, Turkey. European Language Resources Association (ELRA), Paris. Pages 55-59.

Nirenburg S., McShane M.(2009). Computational Field Semantics: Acquiring an Ontological-Semantic Lexicon for a New Language. In *Language Engineering for Lesser-Studied Languages*, Volume 21. Ioss Press. Pages 183-206.

Simpson H., Cieri C., Maeda K., Baker K., Onyshkevych B. (2008). Human Language Technology Resources for Less Commonly Taught Languages: Lessons Learned Toward Creation of Basic Language Resources. *Proceedings of the SALTMIL Workshop: Collaboration: interoperability between people in the creation of language resources for less-resourced languages - 6th International Conference on Language Resources and Evaluation* (LREC '08), Marrakech, Marocco. European Language Resource Association (ELRA), Paris. Pages 7-11.

Zampolli, A. (2000). Introduction of the Conference Chairman. In M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis, G. Stainhaouer (Eds), *Proceedings of the 2nd International Conference on Language Resources and Evaluation* (LREC'00*)*. Volume I, Athens, Greece. European Language Resource Association (ELRA), Paris. Pages XV-XXI.

http://www.elra.info/en/lrec/proceedings/

http://www.lrec-conf.org/lrec1998/elra/workshops.html

https://wordnet.princeton.edu/

https://en.wikipedia.org/wiki/WordNet