Linking Language Resources and NLP papers

Gil Francopoulo, LIMSI, CNRS, Université Paris-Saclay + Tagmatica (France)
Joseph Mariani, LIMSI, CNRS, Université Paris-Saclay (France)
Patrick Paroubek, LIMSI, CNRS, Université Paris-Saclay (France)

Abstract

The Language Resources and Evaluation Map (LRE Map) is an accessible database on Language Resources based on records collected during the submission of several major Speech and Natural Language Processing (NLP) conferences, including the Language Resources and Evaluation Conferences (LREC). The NLP4NLP is a very large corpus of scientific papers in the field of Speech and Natural Language Processing covering a large number of conferences and journals in that field. In this article, we establish the link between those two elements in order to study the mention of the LRE Map resource names within the NLP4NLP corpus.

Keywords: Resource Citation, Named Entity Detection, Informetrics, Scientometrics, Text Mining, LRE Map.

1. Introduction

Our work is based on the hypothesis that names, in this case language resource names, correlate with the study, use and improvement of the given referred objects, in this case language resources. We believe that the automatic (and objective) detection is a step towards the improvement of the reliability of language resources as mentioned in [Branco 2013].

We already have an idea on how the resources are used in the recent venues of conferences such as Coling and LREC, as the LRE Map is built according to the resources declared by the authors of these conferences [Calzolari et al 2012]. But what about the other conferences and the other years? This is the subject of the present study.

2. Situation with respect to other studies

The approach is to apply NLP tools on texts about NLP itself, taking advantage of the fact that we have a good knowledge of the domain ourselves. Our work goes after the various studies presented and initiated in the Workshop entitled: “Rediscovering 50 Years of Discoveries in Natural Language Processing” on the occasion of ACL’s 50th anniversary in 2012 [Radev et al 2013] where a group of researchers studied the content of the corpus recorded in the ACL Anthology [Bird et al 2008]. Various studies, based on the same corpus followed, for instance [Bordea et al 2014] on trend analysis and resulted in systems such as Saffron[1] or the Michigan Univ. web site[2]. Other studies were conducted by ourselves specifically on speech-related archives [Mariani et al 2013], and on the LREC archives [Mariani et al 2014a] but the target was to detect the terminology used within the articles, and the focus was not to detect resource names. More focused on the current workshop topic is the study conducted by the Linguistic Data Consortium (LDC) team whose goal was, and still is, to build a language resource (LR) database documenting the use of the LDC resources [Ahtaridis et al 2012]. At the time of the publication (i.e. 2012), the LDC team found 8,000 references and the problems encountered were documented in [Mariani et al 2014b].

3. Our approach

The general principle is to confront the names of the LRE Map with the newly collected NLP4NLP corpus. The process is as follows:

- Consider the archives of (most of) the NLP field,
- Take an entity name detector which is able to work with a given list of proper names,
- Use the LRE Map as the given list of proper names,
- Run the application and study the results.

4. Archives of a large part of the NLP field

The corpus is a large content of our own research field, i.e. NLP, covering both written and speech sub-domains and extended to a limited number of corpora, for which Information Retrieval and NLP activities intersect. This corpus was collected at IMMI-CNRS and LIMSI-CNRS (France) and is named NLP4NLP[3]. It currently contains 65,003 documents coming from various conferences and journals with either public or restricted access. This is a large part of the existing published articles in our field, apart from the workshop proceedings and the published books. Despite the fact that they often reflect innovative trends, we did not include workshops as they may be based on various reviewing processes and as the access to their content may sometimes be difficult. The time period spans from 1965 to 2015. Broadly speaking, and aside from the small corpora, one third comes from the ACL Anthology[4], one third from the ISCA Archive[5] and one third from IEEE[6].

[1] http://saffron.deri.ie

[2] http://clair.eecs.umich.edu/aan/index.php

[3] See www.nlp4nlp.org

[4] http://aclweb.org/anthology

[5] www.isca-speech.org/iscaweb/index.php/archive/online-archive

[6] https://www.ieee.org/index.html

The corpus follows the organization of the ACL Anthology with two parts in parallel. For each document, on one side, the metadata is recorded with the author names and the title. On the other side, the PDF document is recorded on disk in its original form. Each document is labeled with a unique identifier, for instance "lrec2000_1" is reified on the hard disk as two files: "lrec2000_1.bib" and "lrec2000_1.pdf". When recorded as an image, the PDF content is extracted by means of Tesseract OCR[7]. The automatic test leading to the call (or not) of the OCR is implemented by means of some PDFBox[8] API calls. For all the other documents, other PDFBox API calls are applied in order to extract the textual content. See [Francopoulo et al 2015] for more details about the extraction process as well as the solutions for some tricky problems like joint conferences management.

The majority (90%) of the documents come from conferences, the rest coming from journals. The overall number of words is 270M. Initially, the texts are in four languages: English, French, German and Russian. The number of texts in German and Russian is less than 0.5%. They are detected automatically and are ignored. The texts in French are a little bit numerous (3%), so they are kept with the same status as the English ones. This is not a problem because our tool is able to process English and French. The number of different authors is 48,894. The detail is presented in table 1.

5. Named Entity Detection

The aim is to detect a given list of names of resources, provided that the detection should be robust enough to recognize and link as the same entry some typographic variants such as "British National Corpus" vs "British National corpus" and more elaborated aliases like "BNC". Said in other terms, the aim is not to recognize some given raw character strings but also to link names together, a process often labeled as "entity linking" in the literature [Guo et al 2011][Moro et all 2014]. We use the industrial Java-based parser TagParser[9] [Francopoulo 2007] which, after a deep robust parsing for English and French, performs a named entity detection and then an entity linking processing. The system is hybrid, combining a statistical chunker, a large language specific lexicon, a multilingual knowledge base with a hand-written set of rules for the final selection of the named entities and their entity linking.

6. The LRE Map

The LRE Map is a freely accessible large database on resources dedicated to Natural Language Processing (NLP). The original feature of LRE Map is that the records are collected during the submission of different major NLP conferences[10]. These records were collected directly from the authors. We use the version of the LRE Map collected from 10 conferences from 2010 to 2012 within the EC FlaReNet project as described in [Mariani et al 2015].

The original version was a list of resource descriptions: this does not mean that this is a list of resource names which could be directly used in a recognition system, because what we need for each entry is a proper name, possibly associated with some alternate names. The number of entries was originally 4,396. Each entry has been defined with a headword like "British National Corpus" and some of them are associated with alternate names like "BNC". We further cleaned the data, by regrouping the duplicate entries, by omitting the version number which was associated with the resource name for some entries, and by ignoring the entries which were not labeled with a proper name but through a textual definition and those which had no name. Once cleaned, the number of entries is now 1,301, all of them with a different proper name. All the LRE Map entries are classified according to a very detailed set of resource types. We reduced the number of types to 5 broad categories: NLPCorpus, NLPGrammar, NLPLexicon, NLPSpecification and NLPTool, with the convention that when a resource is both a specification and a tool, the "specification" type is retained. An example is ROUGE which is both a set of metrics and a software package implementing those metrics, for which we chose the "specification" type.

7. Connection of LRE Map with TagParser

TagParser is natively associated with a large multilingual knowledge base made from Wikidata and Wikipedia and whose name is Global Atlas [Francopoulo et al 2013]. Of course, at the beginning, this knowledge base did not contain all the names of the LRE Map. Only 30 resource names were known like "Wikipedia" or "WordNet". During the preparation of the experiment, a data fusion has been applied between the two lists to incorporate the LRE Map into the knowledge base.

8. Running session and post-processing

The entity name detection is applied to the whole corpus on a middle range machine, i.e. one Xeon E3-1270V2 with 32Gb of memory. A post-processing is done in order to filter only the linked entities of the types: NLPCorpus, NLPGrammar, NLPLexicon, NLPSpecification and NLPTool. Then the results are gathered to compute a readable synthesis as an HTML file which is too big to be presented here, but the interested reader may consult the file "lremap.html" on www.nlp4nlp.org. Let's add that the whole computation takes 95 minutes.

[7] https://code.google.com/p/tesseract-ocr

[8] https://pdfbox.apache.org

[9] www.tagmatica.com

[10] As defined in https://en.wikipedia.org/wiki/LRE_Map

short name	# docs	format	long name	language	access to content	period	# venues
acl	4264	conference	Association for Computational Linguistics Conference	English	open access *	1979-2015	37
acmtslp	82	journal	ACM Transaction on Speech and Language Processing	English	private access	2004-2013	10
alta	262	conference	Australasian Language Technology Association	English	open access *	2003-2014	12
anlp	278	conference	Applied Natural Language Processing	English	open access *	1983-2000	6
cath	932	journal	Computers and the Humanities	English	private access	1966-2004	39
cl	776	journal	American Journal of Computational Linguistics	English	open access *	1980-2014	35
coling	3813	conference	Conference on Computational Linguistics	English	open access *	1965-2014	21
conll	842	conference	Computational Natural Language Learning	English	open access *	1997-2015	18
csal	762	journal	Computer Speech and Language	English	private access	1986-2015	29
eacl	900	conference	European Chapter of the ACL	English	open access *	1983-2014	14
emnlp	2020	conference	Empirical methods in natural language processing	English	open access *	1996-2015	20
hlt	2219	conference	Human Language Technology	English	open access *	1986-2015	19
icassps	9819	conference	IEEE International Conference on Acoustics, Speech and Signal Processing - Speech Track	English	private access	1990-2015	26
ijcnlp	1188	conference	International Joint Conference on NLP	English	open access *	2005-2015	6
inlg	227	conference	International Conference on Natural Language Generation	English	open access *	1996-2014	7
isca	18369	conference	International Speech Communication Association	English	open access	1987-2015	28
jep	507	conference	Journées d'Etudes sur la Parole	French	open access *	2002-2014	5
lre	308	journal	Language Resources and Evaluation	English	private access	2005-2015	11
lrec	4552	conference	Language Resources and Evaluation Conference	English	open access *	1998-2014	9
ltc	656	conference	Language and Technology Conference	English	private access	1995-2015	7
modulad	232	journal	Le Monde des Utilisateurs de L'Analyse des Données	French	open access	1988-2010	23
mts	796	conference	Machine Translation Summit	English	open access	1987-2015	15
muc	149	conference	Message Understanding Conference	English	open access *	1991-1998	5
naacl	1186	conference	North American Chapter of the ACL	English	open access *	2000-2015	11
paclic	1040	conference	Pacific Asia Conference on Language, Information and Computation	English	open access *	1995-2014	19
ranlp	363	conference	Recent Advances in Natural Language Processing	English	open access *	2009-2013	3
sem	950	conference	Lexical and Computational Semantics / Semantic Evaluation	English	open access *	2001-2015	8
speechc	593	journal	Speech Communication	English	private access	1982-2015	34
tacl	92	journal	Transactions of the Association for Computational Linguistics	English	open access *	2013-2015	3
tal	177	journal	Revue Traitement Automatique du Langage	French	open access	2006-2015	10
taln	1019	conference	Traitement Automatique du Langage Naturel	French	open access *	1997-2015	19
taslp	6612	journal	IEEE/ACM Transactions on Audio, Speech and Language Processing	English	private access	1975-2015	41
tipster	105	conference	Tipster DARPA text program	English	open access *	1993-1998	3
trec	1847	conference	Text Retrieval Conference	English	open access	1992-2015	24
cell total	67937[11]					1965-2015	577

Table 1: Detail of NLP4NLP, with the convention that an asterisk indicates that the corpus is in the ACL Anthology.

9. Global counting over the whole history

In order to avoid any misleading, we adopt the same conventions as in our other studies, as follows:

- the number of occurrences of a resource name is N when the name is mentioned N times in a document,
- the number of presences of a resource name is 1 when the name is mentioned M times in a document, with M > 0.

We think that the number of presences is a better indicator than the number of occurrences because a resource name may be mentioned several times in a paper for wording reasons, for instance in the body and the conclusion, but

[11] In the general counting, for a joint conference (which is a rather infrequent situation), the paper is counted once (giving 65,003), so the sum of all cells in the table is slightly more important (giving 67,937). Similarly, the number of venues is 558 when the joint conferences are counted once, but 577 when all venues are counted.

what is important is whether the resource is used or not. Year after year, the number of documents per year increases, as presented in figure 1 with the orange line. The number of presences of Language Resources also increases as presented with the blue line.

That means that year after year, more and more LR are mentioned, both as raw counting and as number of presences per document. But we must not forget that there is a bias which boosts the effect: the point is that only recent and permanent resources are recorded in the LRE Map. For instance a resource invented in the 80s' and not used since the creation of the LRE Map in 2010 is not recorded in the LRE Map and will therefore be ignored in our analysis. We see that the number of the presences of Language Resource gets equal to the number of documents in 2006-2007 (it means that on average a Language Resource is mentioned in each paper, as it also appears in figure 2). This period may therefore be considered as the time when the research paradigm in Language Processing turned from mostly model-driven to mostly data-driven. The number of presences then gets even larger than the number of documents.

10. Global top 10 over the history

Over the whole history, when only the top 10 resources are considered, the result is as follows in table 2, ordered by the number of presences in decreasing order. The evolution over the history is presented in figure 3.

There was no mention until 1989, as the earliest LR, TIMIT, appeared at that time. We however see that TIMIT is still much in use after 26 years. The evolution from 1989 until 2015 for these top 10 resources shows for instance that during the period 2004-2011 the resource name "WordNet" was more popular than "Wikipedia", but since 2011, it is the contrary. We can notice also the ridges on even years due to some conferences related to Language Resources that are biennial, such as LREC and Coling on even years.

11. Top 10 for each year

Another way to present the results is to compute a top 10 for each year, as in table 3.

Resource	Type	# pres.	# occur.	First authors mentioning the LR	First corpora mentioning the LR	First year of mention	Last year	Rank
WordNet	NLPLexicon	4203	29079	Daniel A Teibel, George A Miller	hlt	1991	2015	1
Timit	NLPCorpus	3005	11853	Andrej Ljolje, Benjamin Chigier, David Goodine, David S Pallett, Erik Urdang, Francine R Chen, George R Doddington, H-W Hon, Hong C Leung, Hsiao-Wuen Hon, James R Glass, Jan Robin Rohlicek, Jeff Shrager, Jeffrey N Marcus, John Dowding, John F Pitrelli, John S Garofolo, Joseph H Polifroni, Judith R Spitz, Julia B Hirschberg, Kai-Fu Lee, L G Miller, Mari Ostendorf, Mark Liberman, Mei-Yuh Hwang, Michael D Riley, Michael S Phillips, Robert Weide, Stephanie Seneff, Stephen E Levinson, Vassilios V Digalakis, Victor W Zue	hlt, isca, taslp	1989	2015	2
Wikipedia	NLPCorpus	2824	20110	Ana Licuanan, J H Xu, Ralph M Weischedel	trec	2003	2015	3
Penn Treebank	NLPCorpus	1993	6982	Beatrice Santorini, David M Magerman, Eric Brill, Mitchell P Marcus	hlt	1990	2015	4
Praat	NLPTool	1245	2544	Carlos Gussenhoven, Toni C M Rietveld	isca	1997	2015	5
SRI Language Modeling Toolkit	NLPTool	1029	1520	Dilek Z Hakkani-Tür, Gökhan Tür, Kemal Oflazer	coling	2000	2015	6
Weka	NLPTool	957	1609	Douglas A Jones, Gregory M Rusk	coling	2000	2015	7
Europarl	NLPCorpus	855	3119	Daniel Marcu, Franz Josef Och, Grzegorz Kondrak, Kevin Knight, Philipp Koehn	acl, eacl, hlt, naacl	2003	2015	8
FrameNet	NLPLexicon	824	5554	Beryl T Sue Atkins, Charles J Fillmore, Collin F Baker, John B Lowe, Susanne Gahl	acl, coling, lrec	1998	2015	9
GIZA++	NLPTool	758	1582	David Yarowsky, Grace Ngai, Richard Wicentowski	hlt	2001	2015	10

Table 2: Top 10 most mentioned resources over the history

Year	# pres.of LR	# doc. in the year	Top10 cited resources (ranked)
1965	7	24	C-3, LLL, LTH, OAL, Turin University Treebank
1966	0	7	
1967	6	54	General Inquirer, LTH, Roget's Thesaurus, TFB, TPE
1968	3	17	General Inquirer, Medical Subject Headings
1969	4	24	General Inquirer, Grammatical Framework GF
1970	2	18	FAU, General Inquirer
1971	0	20	
1972	2	19	Brown Corpus, General Inquirer
1973	7	80	ANC Manually Annotated Sub-corpus, Grammatical Framework GF, ILF, Index Thomisticus, Kontrast, LTH, PUNKT
1974	8	25	General Inquirer, Brown Corpus, COW, GG, LTH
1975	15	131	C-3, LTH, Domain Adaptive Relation Extraction, ILF, Acl Anthology Network, BREF, LLL, Syntax in Elements of Text, Unsupervised incremental parser
1976	13	136	Grammatical Framework GF, LTH, C-3, DAD, Digital Replay System, Domain Adaptive Relation Extraction, General Inquirer, Perugia Corpus, Syntax in Elements of Text, Talbanken
1977	8	141	Grammatical Framework GF, Corpus de Referencia del Español Actual, Domain Adaptive Relation Extraction, GG, LTH, Stockholm-Umeå corpus
1978	16	155	Grammatical Framework GF, C-3, General Inquirer, Digital Replay System, ILF, LLL, Stockholm-Umeå corpus, TDT
1979	23	179	Grammatical Framework GF, LLL, LTH, C-3, C99, COW, CTL, ILF, ItalWordNet, NED
1980	38	307	Grammatical Framework GF, C-3, LLL, LTH, ANC Manually Annotated Sub-corpus, Acl Anthology Network, Automatic Statistical SEmantic Role Tagger, Brown Corpus, COW, CSJ
1981	33	274	C-3, Grammatical Framework GF, LTH, Index Thomisticus, CTL, JWI, Automatic Statistical SEmantic Role Tagger, Brown Corpus, Glossa, ILF
1982	40	364	C-3, LLL, LTH, Brown Corpus, GG, ILF, Index Thomisticus, Arabic Gigaword, Arabic Penn Treebank, Automatic Statistical SEmantic Role Tagger
1983	59	352	Grammatical Framework GF, C-3, LTH, GG, LLL, Unsupervised incremental parser, LOB Corpus, OAL, A2ST, Arabic Penn Treebank
1984	55	353	LTH, Grammatical Framework GF, PET, LLL, C-3, CLEF, TLF, Arabic Penn Treebank, Automatic Statistical SEmantic Role Tagger, COW
1985	53	384	Grammatical Framework GF, LTH, C-3, LOB Corpus, Brown Corpus, Corpus de Referencia del Español Actual, LLL, DCR, MMAX, American National Corpus
1986	92	518	LTH, C-3, LLL, Digital Replay System, Grammatical Framework GF, DCR, JRC Acquis, Nordisk Språkteknologi, Unsupervised incremental parser, OAL
1987	63	669	LTH, C-3, Grammatical Framework GF, DCR, Digital Replay System, LOB Corpus, CQP, EDR, American National Corpus, Arabic Penn Treebank
1988	105	546	C-3, LTH, Grammatical Framework GF, Digital Replay System, DCR, Brown Corpus, FSR, ISOcat Data Category Registry, LOB Corpus, CTL
1989	145	965	Grammatical Framework GF, Timit, LTH, LLL, C-3, Brown Corpus, Digital Replay System, LTP, DCR, EDR
1990	175	1277	Timit, Grammatical Framework GF, LTH, C-3, LLL, Brown Corpus, GG, LTP, ItalWordNet, JRC Acquis
1991	240	1378	Timit, LLL, C-3, LTH, Grammatical Framework GF, Brown Corpus, Digital Replay System, LTP, GG, Penn Treebank
1992	361	1611	Timit, LLL, LTH, Grammatical Framework GF, Brown Corpus, C-3, Penn Treebank, WordNet, GG, ILF
1993	243	1239	Timit, WordNet, Penn Treebank, Brown Corpus, EDR, LTP, User-Extensible Morphological Analyzer for Japanese, BREF, Digital Replay System, James Pustejovsky
1994	292	1454	Timit, LLL, WordNet, Brown Corpus, Penn Treebank, C-3, Digital Replay System, JRC Acquis, LTH, Wall Street Journal Corpus
1995	290	1209	Timit, LTP, WordNet, Brown Corpus, Digital Replay System, LLL, Penn Treebank, Grammatical Framework GF, TEI, Ntimit
1996	394	1536	Timit, LLL, WordNet, Brown Corpus, Digital Replay System, Penn Treebank, Centre for Spoken Language Understanding Names, LTH, EDR, Ntimit
1997	428	1530	Timit, WordNet, Penn Treebank, Brown Corpus, LTP, HCRC, Ntimit, BREF, LTH, British National Corpus
1998	883	1953	Timit, WordNet, Penn Treebank, Brown Corpus, EuroWordNet, British National Corpus, Multext, EDR, LLL, PAROLE
1999	481	1603	Timit, WordNet, Penn Treebank, TDT, Maximum Likelihood Linear Regression, EDR, Brown Corpus, TEI, LTH, LLL
2000	842	2271	Timit, WordNet, Penn Treebank, British National Corpus, PAROLE, Multext, EuroWordNet, Maximum Likelihood Linear Regression, TDT, Brown Corpus
2001	648	1644	WordNet, Timit, Penn Treebank, Maximum Likelihood Linear Regression, TDT, Brown Corpus, CMU Sphinx, Praat, LTH, British National Corpus
2002	1105	2174	WordNet, Timit, Penn Treebank, Praat, EuroWordNet, British National Corpus, PAROLE, NEGRA, TDT, Grammatical Framework GF
2003	1067	1984	Timit, WordNet, Penn Treebank, AQUAINT, British National Corpus, AURORA, FrameNet, Praat, SRI Language Modeling Toolkit, OAL
2004	2066	2712	WordNet, Timit, Penn Treebank, FrameNet, AQUAINT, British National Corpus, EuroWordNet, Praat, PropBank, SemCor
2005	2006	2355	WordNet, Timit, Penn Treebank, Praat, AQUAINT, PropBank, British National Corpus, SRI Language Modeling Toolkit, MeSH, TDT
2006	3532	2794	WordNet, Timit, Penn Treebank, Praat, PropBank, AQUAINT, FrameNet, GALE, EuroWordNet, British National Corpus
2007	2937	2489	WordNet, Timit, Penn Treebank, Praat, SRI Language Modeling Toolkit, Wikipedia, GALE, GIZA++, SemEval, AQUAINT
2008	4007	3078	WordNet, Wikipedia, Timit, Penn Treebank, GALE, PropBank, Praat, FrameNet, SRI Language Modeling Toolkit, Weka
2009	3729	2637	WordNet, Wikipedia, Timit, Penn Treebank, Praat, SRI Language Modeling Toolkit, GALE, Europarl, Weka, GIZA++
2010	5930	3470	WordNet, Wikipedia, Penn Treebank, Timit, Europarl, Praat, FrameNet, SRI Language Modeling Toolkit, GALE, GIZA++
2011	3859	2957	Wikipedia, WordNet, Timit, Penn Treebank, Praat, SRI Language Modeling Toolkit, Weka, GIZA++, Europarl, GALE
2012	6564	3419	Wikipedia, WordNet, Timit, Penn Treebank, Europarl, Weka, Praat, SRI Language Modeling Toolkit, GIZA++, FrameNet
2013	5669	3336	Wikipedia, WordNet, Timit, Penn Treebank, Weka, SRI Language Modeling Toolkit, Praat, GIZA++, Europarl, SemEval
2014	6700	3817	Wikipedia, WordNet, Timit, Penn Treebank, Praat, Weka, SRI Language Modeling Toolkit, SemEval, Europarl, FrameNet
2015	5597	3314	Wikipedia, WordNet, Timit, SemEval, Penn Treebank, Praat, Europarl, Weka, SRI Language Modeling Toolkit, FrameNet

Table 3: Top 10 mentioned resources per year

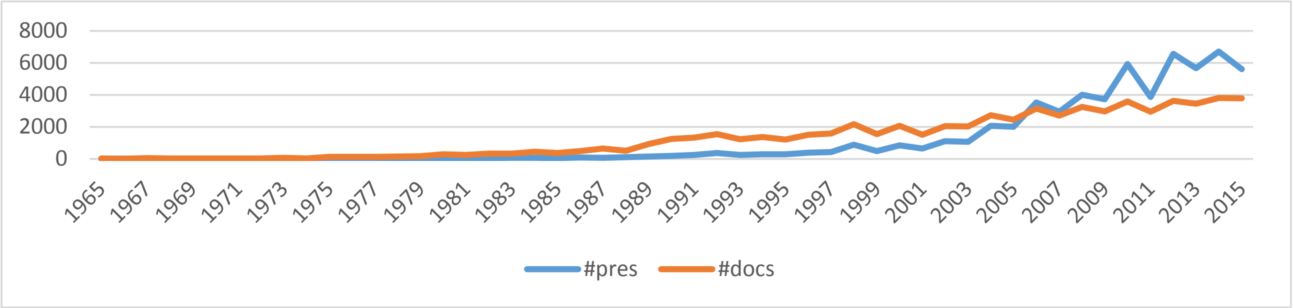

Figure 1: Presence of LR and total number of documents

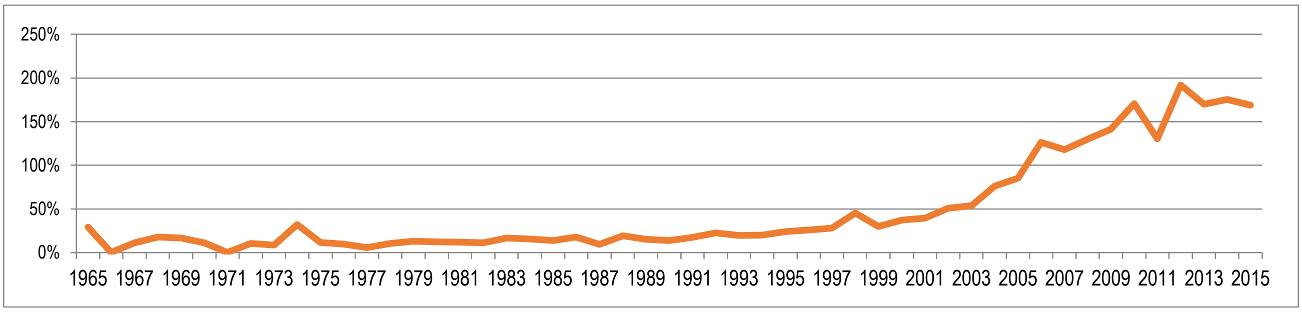

Figure 2: Percentage of LR presence in papers

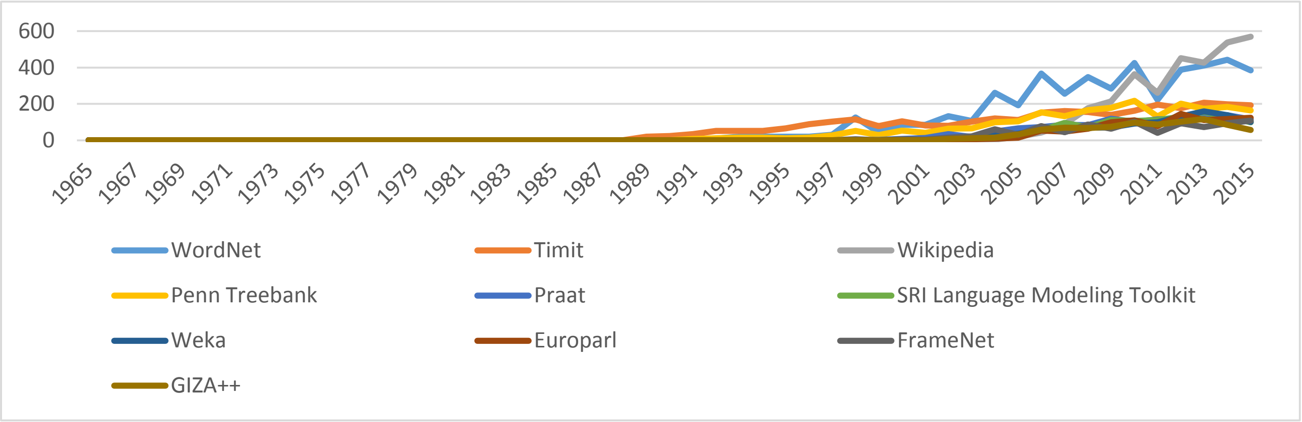

Figure 3: Evolution of the 10 Top LR presences over time

A different way to present the evolution of the terms is to compute a tag cloud at different points in time, for instance every 10 years in 1994, 2004 and 2014 by means of the site Tag Crowd[12]. Let's note that we chose the option to consider 2014 instead of 2015, as LREC and COLING did not occur in 2015.

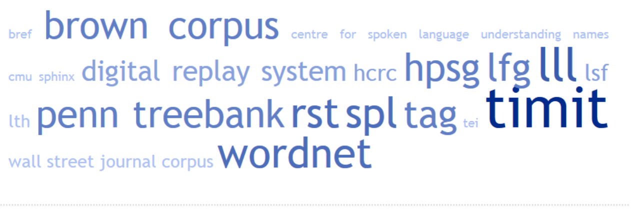
bref brown corpus centre for spoken language understanding names
cmu sphinx digital replay system hcrc hpsg lfg lll lsf
lth penn treebank rst spl tag tei timit
wall street journal corpus wordnet

Figure 4: Tagcloud for 1994

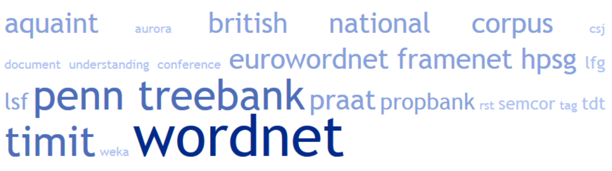
aquaint aurora british national corpus csj
document understanding conference eurowordnet framenet hpsg lfg
lsf penn treebank praat propbank rst semcor tag tdt
timit weka wordnet

Figure 5: Tag cloud for 2004

We see in those figures the sustainable interest over the years for resources such as TIMIT, Wordnet or Penn Treebank. The relative popularity of others such as the Brown Corpus or the British National Corpus decreased over time, while it increased for others such as Wikipedia or Praat, which came to the forefront

[12] http://tagcrowd.com/

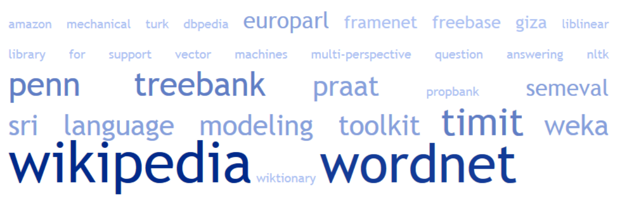

Figure 6: Tag cloud for 2014

12. Targeted study on “wordnet”

Instead of considering the whole set of names, another way to proceed is to select a name, starting from its first mention and to present its evolution, year after year. Let’s consider “WordNet”, starting in 1991 in the figure 7.

Another interesting view is the display the propagation of a specific term from a conference to another by means of a propagation matrix to be read from the top to the bottom. For instance, the first mention of “WordNet” (in our field) was issued in the Human Language Technology (HLT) conference in 1991 (first line). The term propagated in the NLP community through MUC, ACL, TREC and COLING in 1992, then in TIPSTER in 1993 and in the Speech community in 1994 (through the ISCA conference and the Computer Speech and Language journal), as presented in the following matrix of table 4, with the convention that the striped lines indicate that the corresponding corpus doesn’t exist in NLP4NLP, in case of biennal conferences, for example.

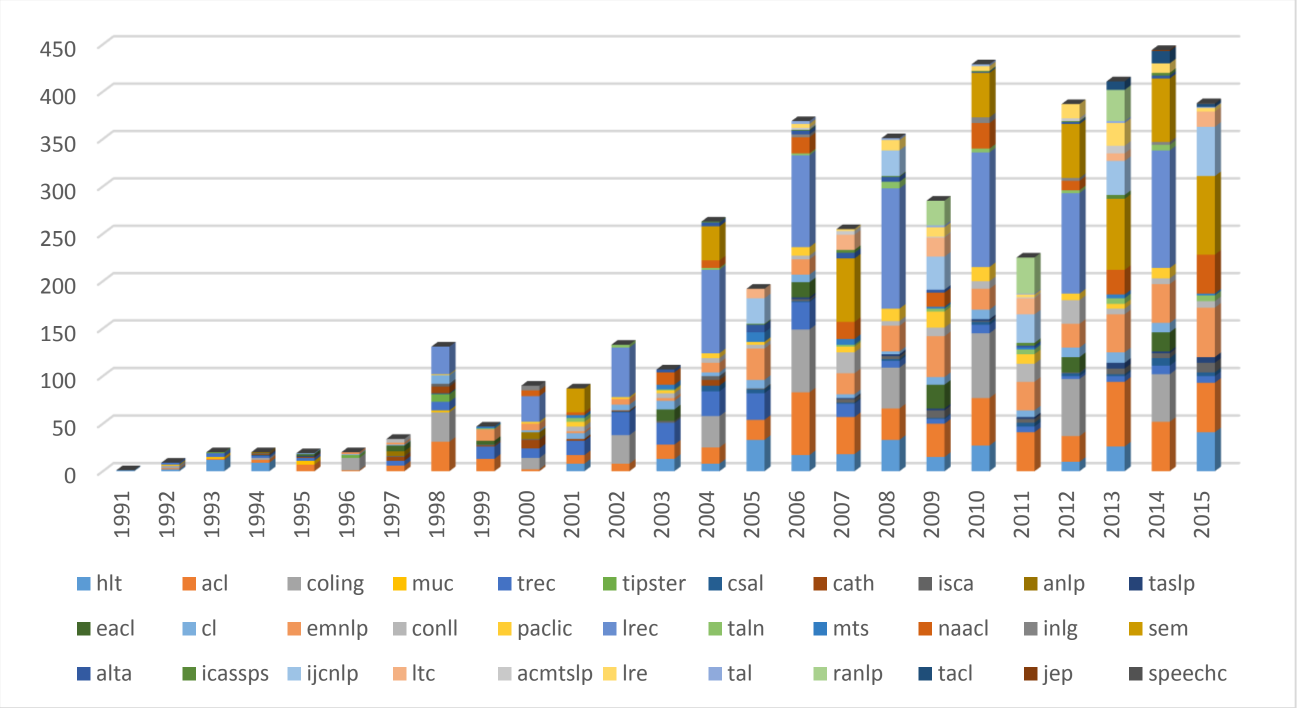

Figure 7: Evolution of "WordNet" presence over time

	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
hlt	■	■	■	■	▤	▤	▤	▤	▤	▤	■	▤	■	■	■	■	■	■	■	■	▤	■	■	▤	■
muc		■	■	▤	■	▤	▤	■	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤
acl		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
trec		■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
coling		■	▤	■	▤	■	▤	■	▤	■	▤	■	▤	■	▤	■	▤	■	▤	■	▤	■	▤	■	▤
tipster			■	▤	▤	■	▤	■	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤
anlp				■	▤	▤	■	▤	▤	■	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤
isca				■		■	■	■	■	■			■	■	■	■	■	■	■	■	■	■	■	■	■
csal				■				■				■		■	■	■	■	■	■	■	■	■	■	■	■
cath				■	■		■	■	■	■	■	■		■	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤
cl					■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	▤
eacl					■	▤	■	▤	■	▤	▤	▤	■	▤	▤	■	▤	▤	■	▤	▤	■	▤	■	▤
taslp					■											■	■	■	■	■	■		■	■	■
emnlp						■	■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
conll							■	▤			■		■	■	■	■	■	■	■	■	■	■	■	■	■
paclic								■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	▤
lrec								■	▤	■	▤	■	▤	■	▤	■	▤	■	▤	■	▤	■	▤	■	▤
taln									■		■	■	■	■		■	■	■	■	■	■	■	■	■	■
mts									■	▤	■	▤	■	▤	■	▤	■	▤	■	▤	■	▤	■	▤	■
inlg										■	▤	▤	▤	▤	▤	■	▤		▤	■	▤	■	▤	■	▤
naacl										■	■	▤	■	■	▤	■	■	▤	■	■	▤	■	■	▤	■
sem											■	▤	▤	■	▤	▤	■	▤	▤	■	▤	■	■	■	■
icassps												■		■	■	■	■	■		■	■	■	■	■	
alta													■	■	■	■	■	■	■	■	■	■		■	▤
ijcnlp															■	▤	▤	■	■	▤	■	▤	■	▤	■
ltc															■	▤	■	▤	■	▤	■	▤	■	▤	■
tal																■		■	■	■	■		■		■
lre																■	■	■	■	■	■	■	■	■	■
acmtslp																■	■		■		■	■	■	▤	▤
ranlp																			■	▤	■	▤	■	▤	▤
tacl																							■	■	■
jep																								■	▤
speechc																									■

Table 4: Propagation matrix for "WordNet"

13. Targeted study on "Wikipedia"

Let's see the evolution of another term like "Wikipedia", starting in 2003, as follows:

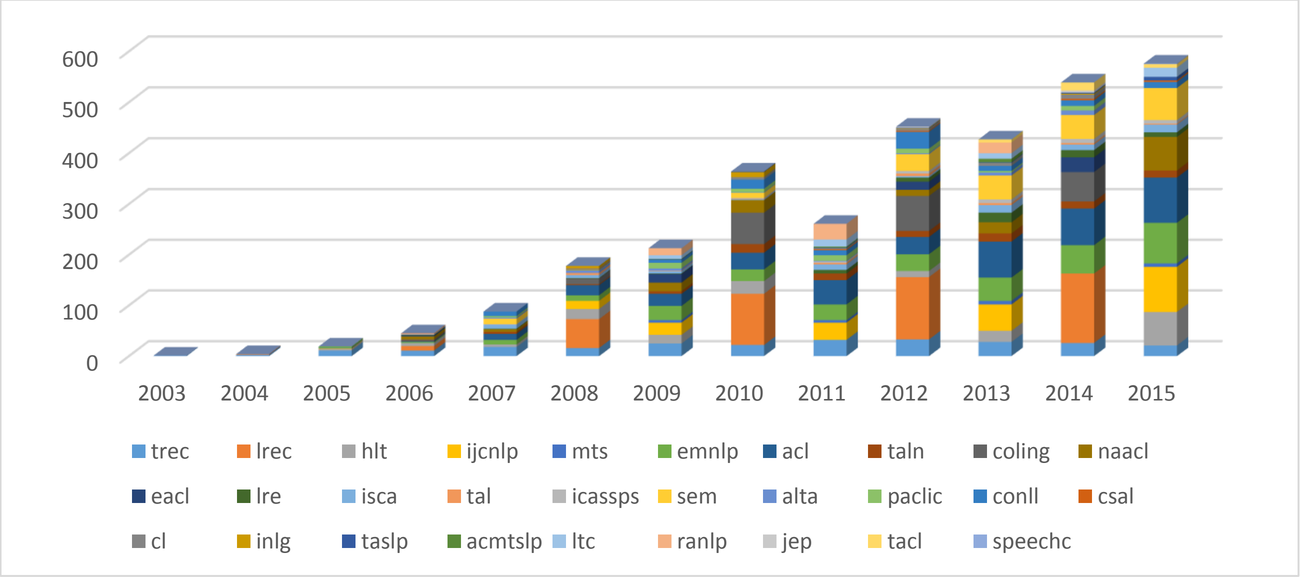

Figure 8: Evolution of "Wikipedia" presence over time

14. Conclusion and Perspective

To our knowledge, this study is the first which matches the content of the LRE Map with the scientific papers published in our field. Beforehand the LRE Map resources were related to the papers of conferences such as Coling and LREC, as the authors were invited to declare these resources during the different paper submission phases, but we had no idea on how these resources were used in other conferences and in other years. Of course, our approach does not cover all the names over the history. For instance a resource invented in the 80s' and not used anymore since 2010 is not recorded in the LRE Map and will therefore be ignored in our analysis. However, we see that Language Resources are more and more used nowadays, and that on average more than one Language Resources is cited in a conference or journal paper. We now plan to consider measuring a resource innovation impact factor for our various sources, conferences and journals: which are the sources where new resources are first mentioned that will later spread in other publications?

14. Acknowledgements

We'd like to thank Wolfgang Hess for the ISCA archive, Douglas O'Shaughnessy, Denise Hurley, Rebecca Wollman and Casey Schwartz for the IEEE data, Nicoletta Calzolari, Helen van der Stelt and Jolanda Voogd for the LRE Journal articles, Olivier Hamon and Khalid Choukri for the LREC proceedings, Nicoletta Calzolari, Irene Russo, Riccardo Del Gratta, Khalid Choukri for the LRE Map, Min-Yen Kan for the ACL Anthology, Florian Boudin for the TALN proceedings and Ellen Voorhees for the TREC proceedings.

15. Bibliographic References

Ahtaridis Eleftheria, Cieri Christopher, DiPersio Denise (2012), LDC Language Resource Database: Building a Bibliographic Database, Proceedings of LREC 2012, Istanbul, Turkey.

Bird Steven, Dale Robert, Dorr Bonnie J, Gibson Bryan, Joseph Mark T, Kan Min-Yen, Lee Dongwon, Powley Brett, Radev Dragomir R, Tan Yee Fan (2008), The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics, Proceedings of LREC, Marrakech, Morocco.

Bordea Georgeta, Buitelaar Paul, Coughlan Barry (2014), Hot Topics and schisms in NLP: Community and Trend Analysis with Saffron on ACL and LREC Proceedings, Proceedings of LREC 2014, 26-31 May 2014, Reykjavik, Iceland.

Branco Antonio (2013), Reliability and Meta-reliability of language resources : ready to initiate the integrity debate ? TLT12 COS, Centre for Open Science.

Calzolari Nicoletta, Del Gratta Riccardo, Francopoulo Gil, Mariani Joseph, Rubino Francesco, Russo Irene, Soria Claudia (2012), The LRE Map. Harmonising Community Descriptions of Resources, Proceedings of LREC, Istanbul, Turkey.

Francopoulo Gil (2007), TagParser : well on the way to ISO-TC37 conformance. ICGL (International Conference on Global Interoperability for Language Resources), Hong Kong, PRC.

Francopoulo Gil, Marcoul Frédéric, Causse David, Piparo Grégory (2013), Global Atlas: Proper Nouns, from Wikipedia to LMF, in LMF Lexical Markup Framework (Francopoulo, ed), ISTE Wiley.

Francopoulo Gil, Mariani Joseph, Paroubek Patrick (2015), NLP4NLP: the cobbler's children won't go unshod, in D-Lib Magazine : The magazine of Digital Library Research[13].

Guo Yuhang, Che Wanxiang, Liu Ting, Li Sheng (2011), A Graph-based Method for Entity Linking, International Joint Conference on NLP, Chiang Mai, Thailand.

Mariani Joseph, Paroubek Patrick, Francopoulo Gil, Delaborde Marine (2013), Rediscovering 25 Years of Discoveries in Spoken Language Processing: a Preliminary ISCA Archive Analysis, Proceedings of Interspeech 2013, 26-29 August 2013, Lyon, France.

Mariani Joseph, Paroubek Patrick, Francopoulo Gil, Hamon Olivier (2014a), Rediscovering 15 Years of Discoveries in Language Resources and Evaluation: The LREC Anthology Analysis, Proceedings of LREC 2014, 26-31 May 2014, Reykjavik, Iceland.

Mariani Joseph, Cieri Christopher, Francopoulo Gil, Paroubek Patrick, Delaborde Marine (2014b), Facing the Identification Problem in Language-Related Scientific Data Analysis, Proceedings of LREC 2014, 26-31 May 2014, Reykjavik, Iceland.

Mariani Joseph, Francopoulo Gil (2015), Language Matrices and a Language Resource Impact Factor, in Language Production, Cognition, and the lexicon (Nuria Gala, Reihard Rapp, Gemma Bel-Enguix editors), Springer.

Moro Andrea, Raganato Alessandro, Navigli Roberto (2014), Entity Linking meets Word Sense Disambiguation : a Unified Approach, Transactions of the Association for Computational Linguistics.

Radev Dragomir R, Muthukrishnan Pradeep, Qazvinian Vahed, Abu-Jbara, Amjad (2013), The ACL Anthology Network Corpus, Language Resources and Evaluation 47: 919–944.

[13] dlib.org/dlib/november15/francopoulo/11francopoulo.html