Rediscovering 15 Years of Discoveries in Language Resources and Evaluation: The LREC Anthology Analysis.

Joseph Mariani[1,2], Patrick Paroubek[1], Gil Francopoulo[2,3], Olivier Hamon[4]
[1]LIMSI-CNRS, [2]IMMI-CNRS, [3]Tagmatica, [4]formerly at ELDA
E-mail : Joseph.Mariani@limsi.fr, pap@limsi.fr, gil.francopoulo@wanadoo.fr, olihamon@gmail.com

Abstract

This paper aims at analyzing the content of the LREC conferences contained in the ELRA Anthology over the past 15 years (1998-2013). It follows similar exercises that have been conducted, such as the survey on the IEEE ICASSP conference series from 1976 to 1990, which served in the launching of the ESCA Eurospeech conference, a survey of the Association of Computational Linguistics (ACL) over 50 years of existence, which was presented at the ACL conference in 2012, or a survey over the 25 years (1987-2012) of the conferences contained in the ISCA Archive, presented at Interspeech 2013. It contains first an analysis of the evolution of the number of papers and authors over time, including the study of their gender, nationality and affiliation, and of the collaboration among authors. It then studies the funding sources of the research investigations that are reported in the papers. It conducts an analysis of the evolution of the research topics within the community over time. It finally looks at reuse and plagiarism in the papers. The survey shows the present trends in the conference series and in the Language Resources and Evaluation scientific community. Conducting this survey also demonstrated the importance of a clear and unique identification of authors, papers and other sources to facilitate the analysis. This survey is preliminary, as many other aspects also deserve attention. But we hope it will help better understanding and forging our community in the global village.

Keywords: ELRA Anthology, Language Resources, Language Processing Systems Evaluation, Text Analytics, Social Networks, ISLRN, Bibliometrics, Scientometrics.

1. Introduction

1.1. The ELRA community and conference series

Activities in the area of Language Resources and Evaluation greatly increased over the past 30 years, due to the importance of Language Resources to conduct research investigations in language sciences and to develop language processing systems which are based on automatic Machine Learning.

Some milestones may be identified in this area, such as the launching of the evaluation campaigns of speech recognition systems by NIST for DARPA in 1987, the creation of the Linguistic Data Consortium (LDC) and of the Coordinating Committee on Speech Databases and Speech Input/Output Systems Assessment (Cocosda) in 1991. This was followed by the launching of the European Language Resources Association (ELRA) in 1995, which organized the first Language Resources and Evaluation (LREC) conference in 1998. The oriental branch of Cocosda organized the Oriental-Cocosda conference for the first time on the same year. The Language Resources and Evaluation Journal published by Springer was initiated in 2005.

The idea of adding a scientific dimension to the Language Resources distribution activity provided by ELRA through an international conference specifically devoted to Language Resources and Evaluation was first proposed by Joseph Mariani in 1997. The first conference was held in 1998 in Granada (Spain). It was organized and chaired by Antonio Zampolli. Following its great success, the LREC conference has been organized every two years since then and is now chaired by Nicoletta Calzolari.

On the occasion of the 9th LREC conference, it was felt useful to reconsider the last 15 years of research in the area of Language Resources and Evaluation through an analysis of the proceedings of the former 8 LREC conference gathered in the LREC Anthology.

This analysis is similar to comparable exercises which were conducted in the Computational Linguistics community through an analysis of the ACL Anthology which covered 50 years of the ACL conferences and was presented within a specifically dedicated workshop at the 2012 ACL conference in Jeju (Korea), or in the

Spoken Language Processing community through the analysis of the ISCA Archive which covered 25 years of the ECST, Eurospeech, ICSLP and Interspeech conferences and was presented at the 2013 Interspeech conference in Lyon (France). A previous exercise over 15 years of the IEEE ICASSP conference was conducted by the end of the 1980s and served for deciding to launch the ESCA Eurospeech conference. Other analyses may be reported on various NLP conferences, including LREC, in the Saffron project, on the French TALN conference, and on many other conference or social networks (SNAP):

1.2. The ACL Anthology analysis

A similar inspiring exercise has been conducted by the Association for Computational Linguistics (ACL) on the occasion of their 50th anniversary at the ACL 2012 conference (Jeju, Korea), in the form of a one-day workshop entitled "Rediscovering 50 Years of Discoveries in Natural Language Processing" (ACL, 2012). This analysis was conducted by 23 authors within 13 papers addressing various aspects, and using technologies developed in the framework of text analytics, a very active area of research in Natural Language Processing nowadays. They used for this the ACL Anthology[1], which contains data coming from the ACL international conference and workshops, but also from other conferences or journals related to Computational Linguistics (NAACL, EACL, EMNLP, COLING, LREC, ANLP, IJCNLP, Computational Linguistics Journal, etc.). Various analyses of those data, including the Collaboration Graph, Author Citation Graph and Paper Citation Graph, are available at the University of Michigan[2], where data and tools are also available.

1.3. The ISCA community and conference series

Research activities in spoken language processing have been very active for many years. By the end of the 80s, initiatives in Europe and in Asia helped organizing the international community through the creation of the European Speech Communication Association (ESCA) in 1988, followed by the launching of the biennial Eurospeech conference series in 1989 in Europe, and the launching of the biennial International Conference on Spoken Language Processing (ICSLP) in 1990 in Asia, which completed the landscape, previously composed for the most part by the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

From 2000 onwards, Eurospeech and ICSLP merged in a single annual Interspeech conference, under the umbrella of the International Speech Communication Association (ISCA), based on ESCA and on the Permanent Council for the organization of the ICSLPs (PC-ICSLP) (J. Mariani, 2013) (H. Fujisaki, 2013). On the occasion of Interspeech 2013 in Lyon (France) 24 years after the first Eurospeech conference, which took place in Paris in 1989, it was thought interesting to have a look back at the past years and to analyze the steps which resulted in the state-of-the-art in spoken language processing science and technology. This analysis aimed also at providing a good insight of the community and at building up the next steps for the future.

The study covers 25 years of research taking the opportunity of the availability of the ISCA Archive[3] comparable to the ACL Anthology, assembled by Wolfgang Hess, which covered the 1987-2012 period. In a first step, we decided to only consider the conferences, starting with the European Conference on Speech Technology (ECST) organized in 1987 in Edinburgh, followed by the Eurospeech and ICSLP conference series, and by the Interspeech conference series starting in 2000, leaving aside for now other events like workshops (J. Mariani et al., 2013).

1.4. The ICASSP 1976-1990 conference series analysis

A similar, although simpler, analysis was actually conducted by J. Mariani on the IEEE ICASSP conference series (on a 15 years time span from 1976 to 1990), accompanying the launching of the Eurospeech conference in 1989, in his capacity of ESCA president at that time and Technical Chairman of Eurospeech 1989 (J. Mariani, 1990). It appeared that the number of papers at ICASSP, in general but also in speech, increased over those 15 years. The number of speech papers at ICASSP represented overall about 30% of the papers (2,284 on 7,156), but the ratio of speech papers decreased over time from about 50% in 1976 to 30% in 1990. Looking more precisely, it was striking to notice that, even if the US were the largest providers of speech papers overall (more than 50%), whenever the ICASSP conference took place outside the US (Paris (France) in 1982, Tokyo (Japan) in 1986 and Glasgow (UK) in 1989), the total participation increased, the US participation stayed very high, while the European and Asian participation increased a lot and was even on a par with the US one, as it happened typically in Tokyo in 1986. It also resulted in a stronger dynamics of the conference for the following

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

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

[3] http://www.isca-speech.org/iscaweb/index.php/archive/online-archive

years. This advocated for the launching of truly international conferences more specifically devoted to spoken language processing, while covering all the aspects of this research area, as confirmed by Eurospeech and ICSLP, which immediately obtained a large international success.

1.5. The TALN conference series analysis

The proceedings of the TALN conference organized yearly by the French ATALA (*Association pour le Traitement Automatique des Langues*) have been made available online recently[4] and a first study has been presented in 2013 (F. Boudin, 2013).

1.6. SNAP at Stanford University

Stanford University has launched an initiative called the Stanford Large Network Dataset Collection[5] in order to study various kinds of networks, including social networks and the collaboration and citation graphs of scientific conferences (in astrophysics, High Energy Physics, General Relativity and Condensed Matter), where data and tools are also available.

1.7 SAFFRON

At the University of Galway, the Saffron project provides insights in a research community or organization by analyzing its main topics of investigation and the experts associated with these topics. Saffron analysis is fully automatic and is based on text mining and linked data principles. It concerns Natural Language Processing: LREC[6], the ACL Anthology (ACL Annual Conferences, COLING, EACL, HLT, ANL), Information Retrieval (CLEF) and the Semantic Web (Semantic Web Dog Food).

2. Analysis of the series of LREC conferences

As a convention, we will refer to the conference publications as "papers" or "articles". We will refer to individual "authors" and mention their "signatures" or "contributions" to a publication where they act as "contributors". The same author may sign several papers at a given conference, as a single author or together with one or several co-authors.

2.1. The ELRA conferences Anthology

The study covers the series of conferences contained in the LREC Anthology, assembled by Olivier Hamon, which contains the 8 LREC conferences held since 1998 (see Table 1). This covers a time span of 15 years (1998-2013). We did not consider for the time being in this study the workshops, organized as satellite events of LREC, and other ELRA supported events. All the corresponding data, apart from the 1998 Proceedings, is freely available online on the ELRA Web Site[7] and in the ACL Anthology.

Year	Place	# Papers	# Signatures	# Signatures/paper
1998	Granada	212	618	2.92
2000	Athens	280	855	3.05
2002	Las Palmas	354	1,130	3.19
2004	Lisbon	517	1,709	3.31
2006	Genoa	514	1,667	3.24
2008	Marrakesh	620	2,147	3.46
2010	Malta	641	2,199	3.43
2012	Istanbul	670	2,291	3.42
Total		3,808	12,616	3.31

Table 1. *List of conferences with number of papers and authors.*

SAFFRON also analyzed and provides results on the LREC conference analysis, including a dynamic taxonomy of the papers topics[8]. A single LREC conference (LREC 2008) was also analyzed in the ACL Archive Network[9].

[4] www.atala.org/-Conference-TALN-RECITAL

[5] http://snap.stanford.edu/data/

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

[7] http://www.lrec-conf.org/

[8] http://saffron.deri.ie/lrec/

In terms of citation, Google Scholar[10] places LREC at the 4th rank with an h5 index of 35 within the last 5 years and an h5-median mean of 50, on a par with the Coling conference (35 and 50) or the Interspeech conference (33 and 42).

2.2. The resources: data and tools

2.2.1. The data

Regarding the conference series, the LREC proceedings content is freely available online on the ELRA website[11], as well as on the ACL Anthology website[12]. It contains the metadata (List of authors and sessions, Content of the sessions and, for each article, Titles, Authors, Affiliations, Abstract and Bibliographic Reference of the paper), as well as the full content of the articles. All the data is available in its digital content, except the 1998 proceedings, which are only available as images. It was thus necessary to automatically OCRize the corresponding images to get the text in a digital format. In this study, we used the metadata for the chapters 2.3. (Papers) and 2.4. (Authors), and the full content for the chapters 2.5. (Citations), 2.6. (Topics) and 2.7. (Text reuse and plagiarism).

Rapidly, we had the need to set up a benchmark to estimate the quality of the extraction based on a simple hypothesis which is to compute the number of files and the number of known/unknown words when using a broad coverage lexicon. We made the hypothesis that when the process is good, the number of files is high and the number of errors is low. The aim is to be able to take decisions concerning the various parameters of the tools with a quantified evaluation for each variation. The number of errors is computed from the result of the morphological module of TagParser (G. Francopoulo, 2007) which is a deep industrial parser based on a broad English lexicon and Global Atlas (a knowledge base containing more than one million words from 18 Wikipedias) (G. Francopoulo, 2013).

3,808 papers have been published at the 8 LREC conferences and are available in a PDF format. The measure of the quality of the textual data shows that the full series of proceedings contain about 14 million words, and that the overall quality is good, with less than 1% unknown words, except for the year 1998 due to the fact that the textual data was obtained through automatic OCRization for that year (see 2.2.2.).

Year	# papers in metadata	# papers in pdf	# papers in XML (from pdfbox)	# of non-empty papers resulting from the extraction	# of missing abstracts (from the extraction)	# of unknown words	# of words in the content	Quality Evaluation: % (# of unknown words / # of words in the content)	Quality evaluation: % (# of unknown words starting with a lower case letter / # of words in the content)
1998	212	212	212	209	46	32,218	830,231	3.881	2.449
2000	280	280	280	252	29	22,957	990,361	2.318	1.171
2002	354	354	354	339	24	27,713	1,384,150	2.002	0.855
2004	517	517	517	507	41	28,309	1,474,788	1.920	0.805
2006	514	514	514	508	31	40,179	1,826,763	2.199	0.881
2008	620	620	620	617	31	53,477	2,355,240	2.271	0.833
2010	641	641	640	639	37	59,477	2,503,410	2.376	0.982
2012	670	670	670	670	31	60,603	2,693,673	2.250	0.922
Total	3,808	3,808	3,807	3,741	270	324,933	14,058,616	2.311	1.001

Table 2. *Quality of the proceedings*

2.2.2. The tools

The metadata were processed with MS Excel, OpenOffice spreadsheet Calc, the R statistical suite (The R Journal, 2012), iGraph (Csàrdi et al., 2006), the search engine swish-e[13], RankChart and various scripts written in bash shell and C++. The linguistic processing was limited to the use of G. Grefenstette awk implementation of Porter's stemmer (M. F. Porter, 2012) and local grammars compiled either with the Unitex toolkit[14] or flex[15]. The large graph visualization and analysis platform Tulip (D. Auber et al., 2012), was used to browse the co-author

[9] http://clair.eecs.umich.edu/aan/index.php
[10] http://scholar.google.com/citations?view_op=top_venues&hl=en&vq=eng_computationallinguistics
[11] http://www.elra.info/LREC-Conference.html
[12] http://aclweb.org/anthology/
[13] http://swish-e.org/
[14] http://www-igm.univ-mlv.fr/~unitex/
[15] http://flex.sourceforge.net/

and publication graphs. We also sometimes used Google Search for solving authors' gender and harmonizing countries and institutions names.

For analyzing the abstracts and the textual bodies of the articles, we proceeded differently because we needed a deeper and more suitable linguistic processing. As said previously, the 1998 proceedings are only available in hard-copy format. Thus, ELDA manually scanned those proceedings in order to obtain a set of image-format PDF files. It should be noted that for the other years, there is also a small number of files that are in image-format and we wanted to process these contents. On all the files, we then used PdfBox (B. Litchfield, 2005) to extract the content of the non-image files. Associated with PdfBox, we use the "magical" library called "Bouncy Castle" to have access to the small number of encrypted contents[16]. When PdfBox failed to extract the content, we called Tesseract-OCR[17] to produce a textual content. Given a certain number of conditions on the size of the content, the paper was (or was not) retained.

As a first trial, we used ParsCit (I. G. Councill et al. 2008) which was used for the ACL Anthology but we faced different problems due to the fact that the program was not suited for Slavic, German, extended Latin and phonetic alphabets and we did not have the time to retrain the system. We decided to write a small set of rules in Java to extract the abstract and the body and to compare the quality that happened to be 2.5% higher, so we decided not to use ParsCit in this present analysis.

Along with the previous toolkits, we have used the following language resources: the British National Corpus (BNC) (The British National Corpus, 2007), the Open American National Corpus (OANC) (N. Ide et al., 2010), Europarl (Ph. Koehn, 2005), Tagmatica Named Entity database extracted from Wikipedia and various journalistic sources, and a lexicon of 59,850 given names with gender information.

2.3. The papers

The total number of papers published in the conference series amounts to 3,808 (Table 1), with a steadily increase over time from 212 papers in 1998 to 517 at LREC 2004, followed by a stability in 2006 (Genoa), and by a steadily increase since then (Fig. 1). The rejection rate is stable at about 40% of the submitted papers, as mentioned in the conference chair's introductory messages.

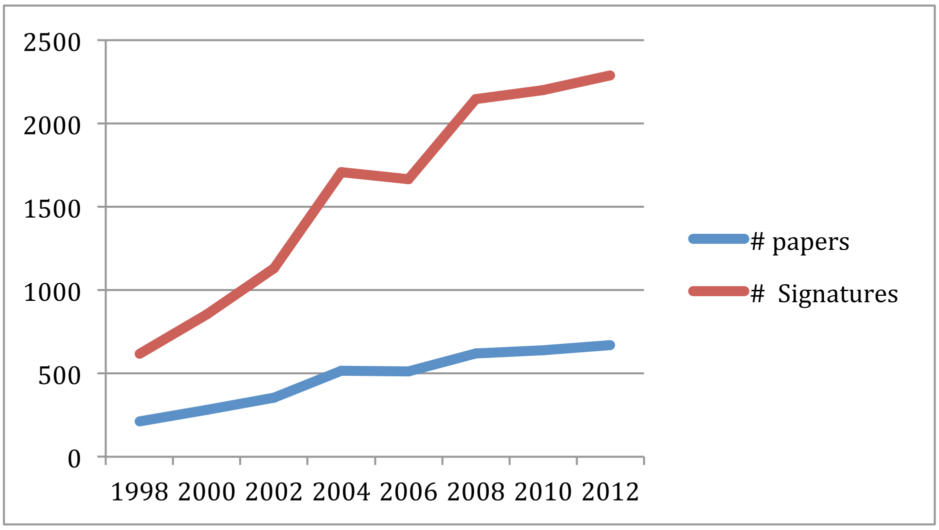

Figure 1. *Number of papers and signatures over time*

2.4. The authors

2.4.1. Number of contributors per conference

Accordingly, the number of signatures also steadily rose up to 1,709 at LREC 2004. It then slightly decreased to 1,667 at LREC 2006 and kept increasing since then to reach 2,291 at LREC 2012 (Fig. 1).

[16] www.bouncycastle.org
[17] https://code.google.com/p/tesseract-ocr/

2.4.2. Number of authors per paper

The number of co-authors per paper is most often 2 to 3 (Fig. 2). The largest number of co-authors for a paper is 21.

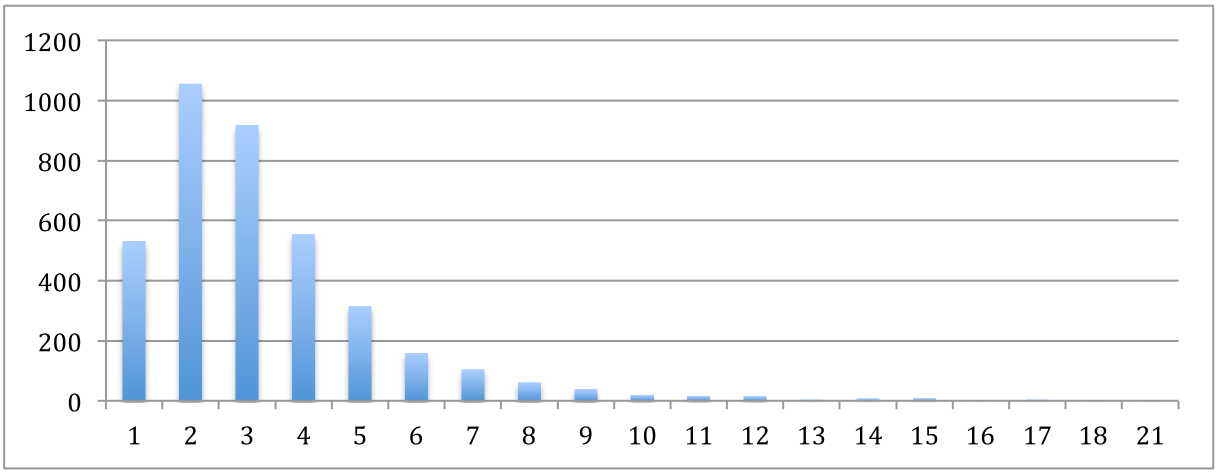

Figure 2. *Number of papers according to the number of co-authors*

2.4.3. Number of signatures per paper over time

However, the average number of co-authors per paper increased over time, from 2.92 in 1998 up to 3.42 in 2012 (i.e. 0.5 more author on average), expressing the fact that more and more scientists collaborate on a research study (Fig. 3).

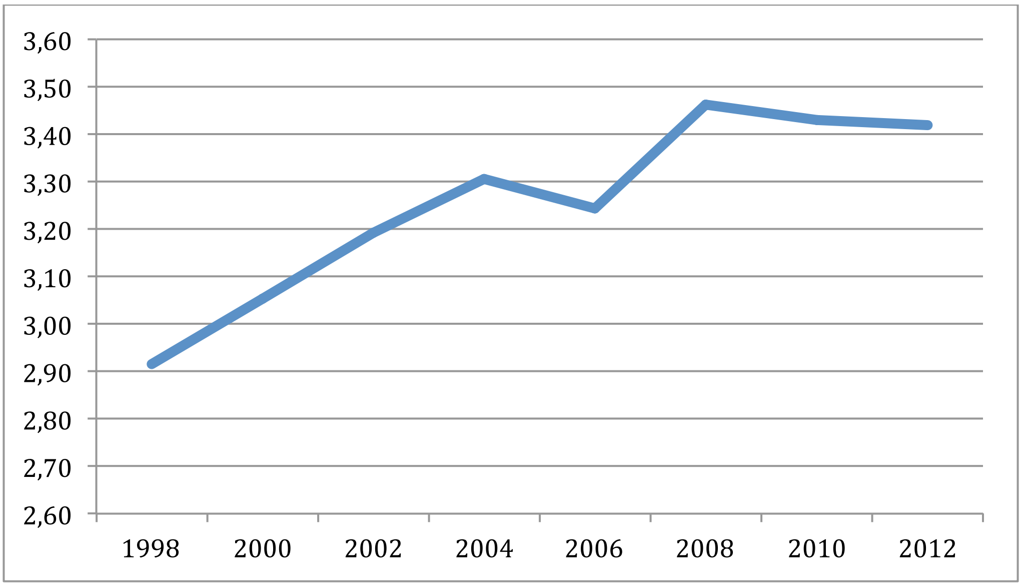

Figure 3. *Average number of authors per paper*

It is interesting to notice that the number of papers with a single author was 11% in 1998 and went down to 3% in 2012, while the number of papers with 3 authors or more was 70% in 1998 and went up to 82% in 2012. This clearly demonstrates the change on the way research is conducted, going progressively from individual research investigations to large projects conducted within teams or in collaboration within consortia, often in international projects and programs.

2.4.4. Number of different authors

The study of the authors is difficult due to the various ways of writing their name (family name and given name, initials, middle initials, ordering, married name, etc.). It therefore necessitated a tedious cleaning process, which was made semi-automatically. On an initial total of 12,474 authors' names, about 6,000 family names or given names had to be corrected, resulting in a list of 6,118 different authors (i.e. a 50% reduction). This clearly demonstrates the need for identifying uniquely each researcher.

2.4.5. Renewal of authors

We first studied the number of authors at each following conference (Table 3). We identified at each conference the authors who didn't publish at the previous conference, that we call "New Authors". We also studied those who never published at any previous conference, that we call "Completely New Authors".

Year	# Authors Signatures	# Different Authors	Authors' redundancy	# New Authors compared with previous LREC	% New Authors	# Completely New Authors	% Completely New Authors
1998	618	506	22%	506	100%	506	100%
2000	855	704	21%	551	78%	551	78%
2002	1,130	894	26%	666	74%	613	69%
2004	1,709	1,288	33%	940	73%	837	65%
2006	1,667	1,281	30%	892	70%	770	60%
2008	2,147	1,668	29%	1,190	71%	954	57%
2010	2,199	1,699	29%	1,142	67%	953	56%
2012	2,291	1,768	30%	1,199	68%	934	53%
Total	12,616					6,118	

Table 3. *Authors' renewal and redundancy*

The difference between the number of signatures and the number of different authors reflects the number of authors whose name appear in several papers, what we may call the "authors' variety", and conversely the "authors' redundancy". It appears that this redundancy slightly increased over time, showing a concentration of the papers authors, and is now stabilized at about 30% (Fig. 4).

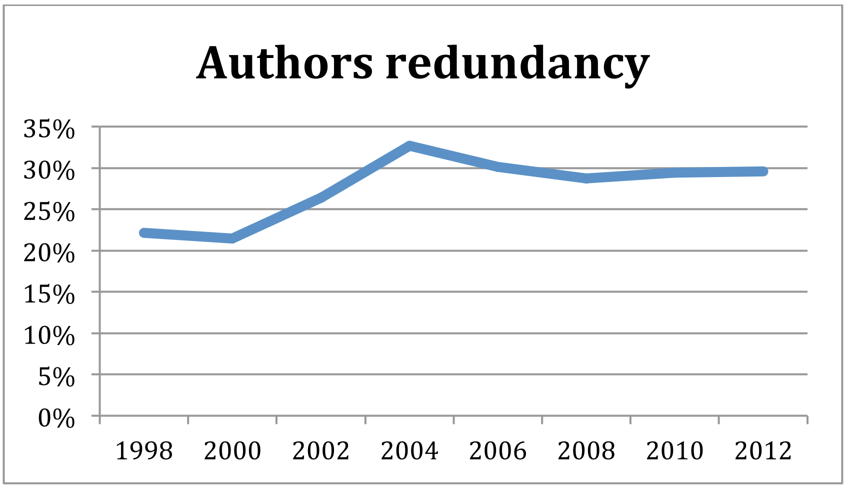

Figure 4. *Authors' redundancy over time*

We then studied the authors' renewal. It clearly showed (Fig. 5) that the number of different authors from one conference to the next conference has been high and increased over time, apart from a slight decrease in 2006 due to the lesser number of papers, until LREC 2008, where there were about 1,200 new authors compared with LREC 2006. It then stayed steadily important with a turn over of about 1,200 different authors each year. The same appears for the number of totally new authors which increased every year up to LREC 2008, apart from the 2006 accident, with 954 new authors that year, but then slightly decreased over time to 934 in 2012. This also appears in terms of percentages (Fig. 6) showing that the percentage of different authors from one year to the next decreased from 78% in 2000 to 68% in 2012, while the number of totally new authors decreased from 78% in 2000 to about 50% in 2012. This shows the stabilization of the research community over time, but may also reflect the commencement of a lack of "new blood", even if this number is still much higher than in other more established communities (only 30% in the case of Interspeech).

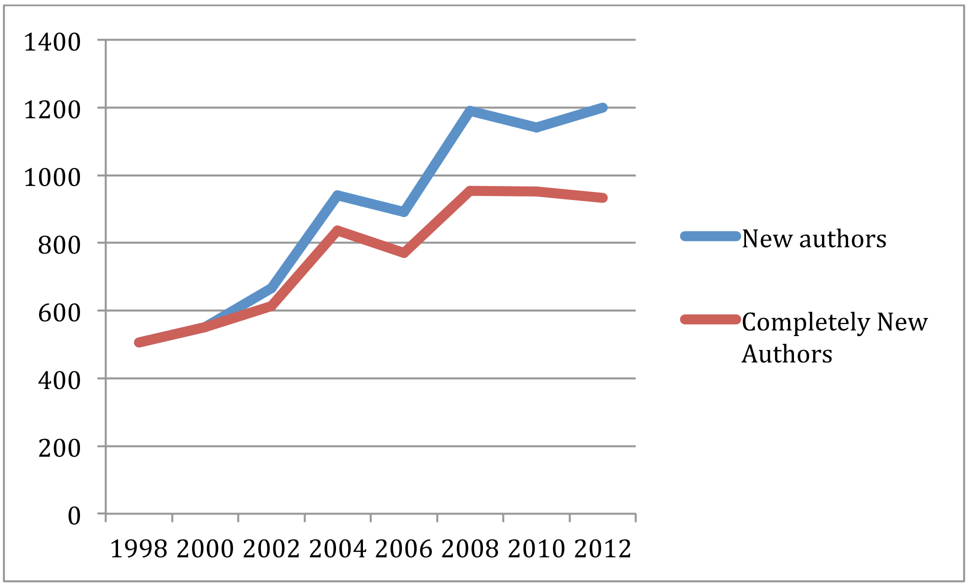

Figure 5. *Number of authors, new authors and completely new authors over time*

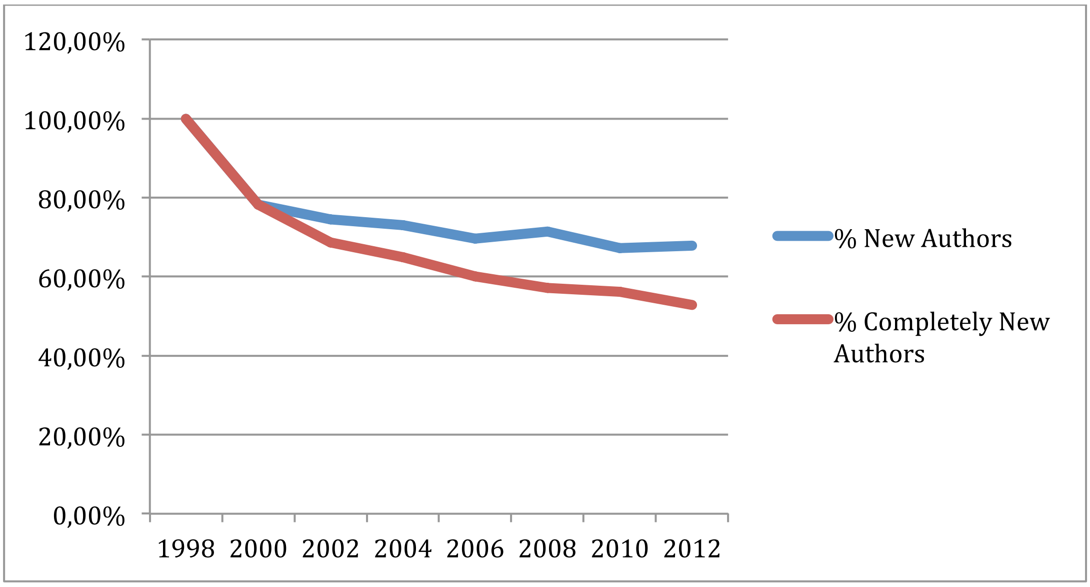

Figure 6. *Percentage of new authors and completely new authors over time.*

2.4.6. Gender of authors

The author gender study was performed with the help of a lexicon of 59,850 given names with gender information (54% male, 44% female, 2% epicene). Variations due to different cultural habits for naming people (single versus multiple given names, family versus clan names, inclusion of honorific particles, ordering of the components etc.) (Yu Fu et al., 2010), changes in editorial practices and sharing of the same name by large groups of individuals, all contribute to make identifying the person referred to by a name a difficult problem, so much that initiatives exist to provide world-wide unique identifiers for researchers (B. Joerg et al., 2012). In this preliminary study we have used a crude normalization of proper names in ASCII, separating them into two components: given name and family name, allowing for compound forms in both parts. Note that for some of them, we only had an initial for the first name, which made gender guessing impossible, unless the same person also appears with his/her first name in full somewhere else. Although the result of the automatic processing was hand-checked by an expert of the domain for the most frequent names, the results presented here need to be considered with caution allowing for an error margin.

The analysis over the 8 conferences shows that 55% of the authors are male, while 28% of the authors are female, with 1% of epicene gender and 16% are of unknown gender (Fig. 7 and 8). If we consider that the authors of unknown gender have the same gender distribution than the ones which are categorized, the ratio of male authors would be 66%, while the female authors would be 34% (Fig. 9).

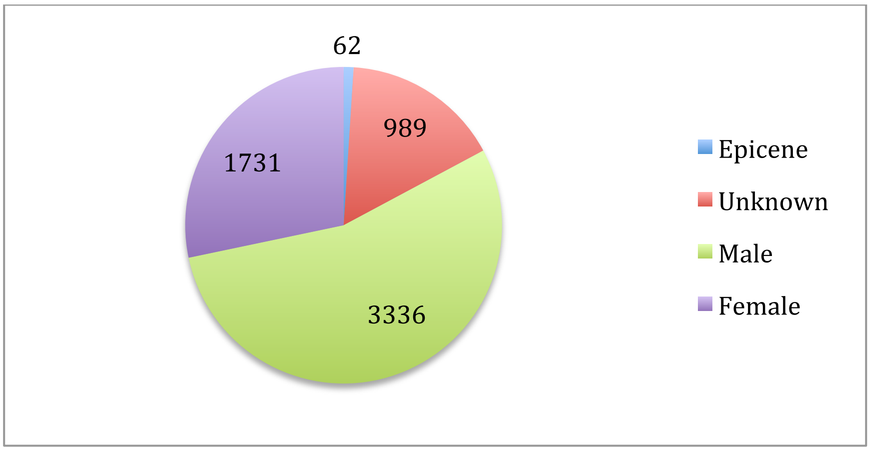

Figure 7. *Gender of the 6,118 authors overall*

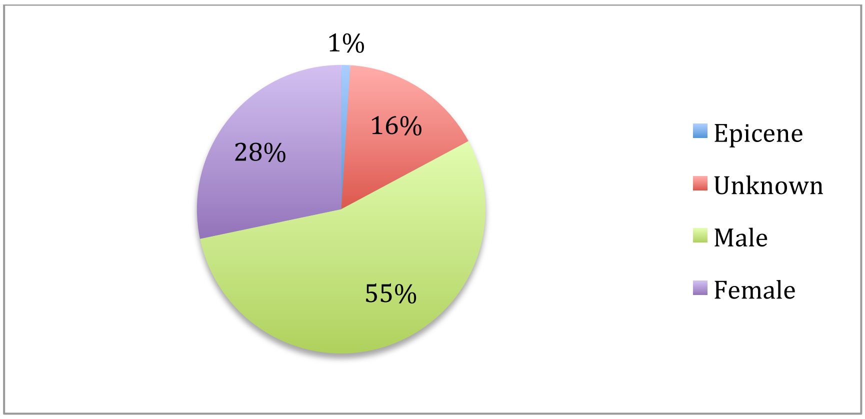

Figure 8. *Percentages of gender of the 6,118 authors overall*

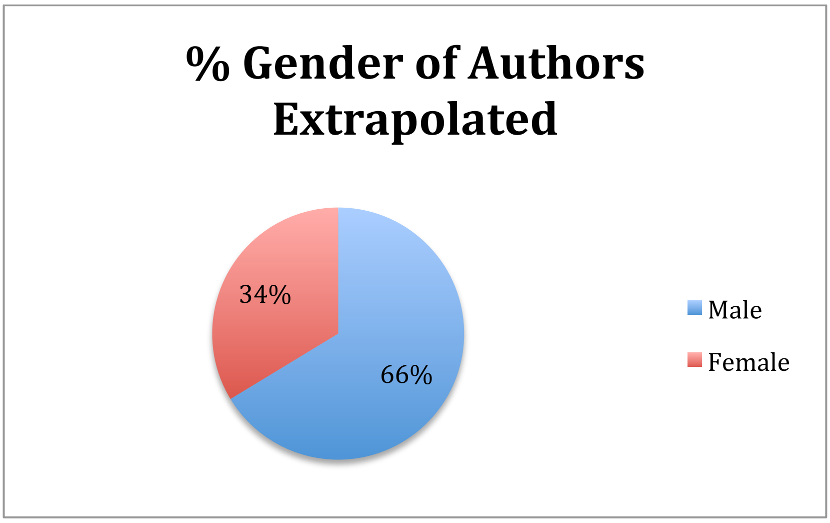

Figure 9. *Percentages of genders overall under the assumption that the distribution on unknown gender is similar*

If we now consider the signatures by gender over the 3,808 papers (Fig. 10), we find even a slight increase in the male share (68% against 32%).

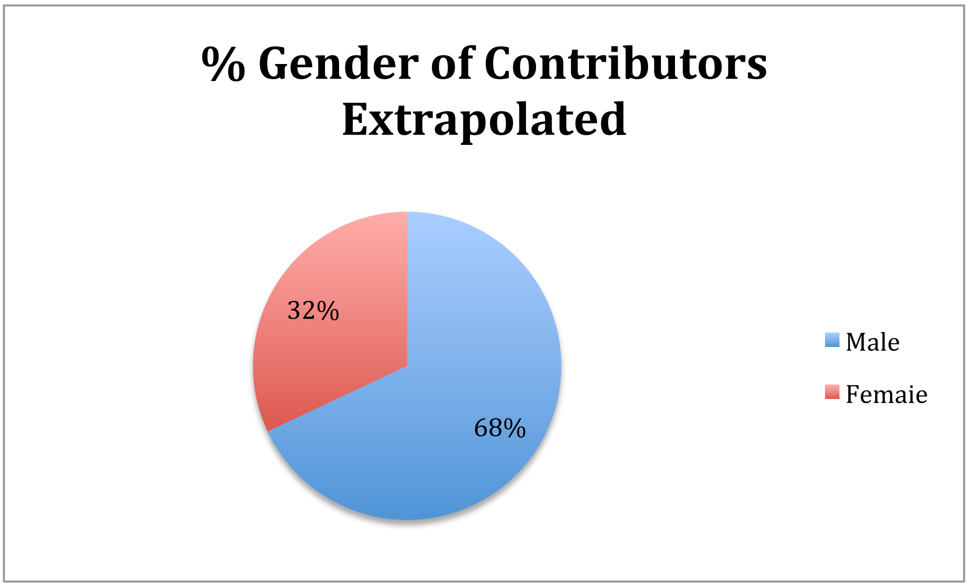

Figure 10. *Gender of the authors' contributions overall*

The analysis of the authors' gender over time (Fig. 11) shows a relative stability of male authors around 65% to 70% and of female authors around 35 % to 30%.

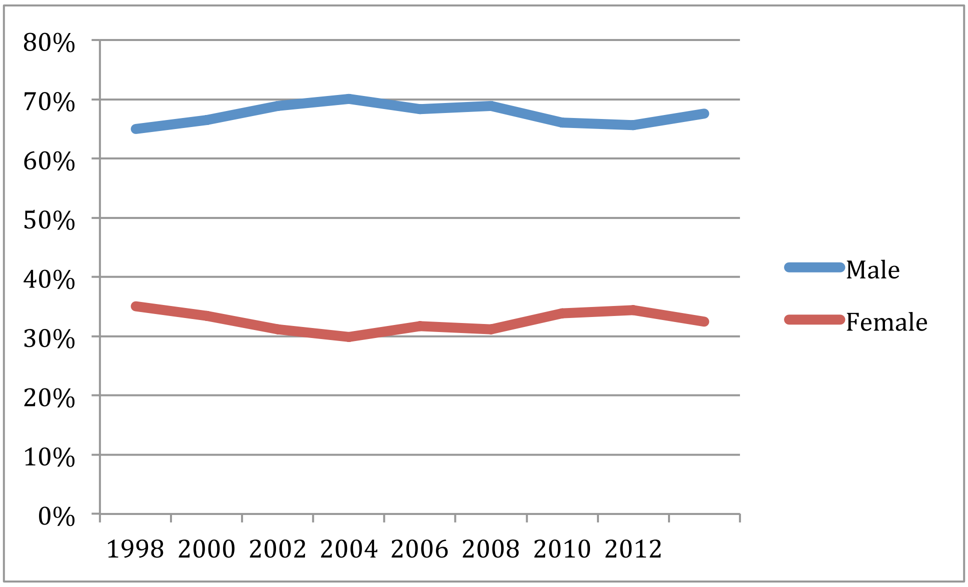

Figure 11. *Gender of the authors' contributions over time.*

2.4.7. Nationality of authors

We studied the nationality of the papers signatures. When an author mentions several affiliations from different countries, it is counted as one contributor for the two first countries he/she mentions. Over the 8 LREC conferences, papers have been published by authors of 75 different countries (Fig. 12a and 12b).

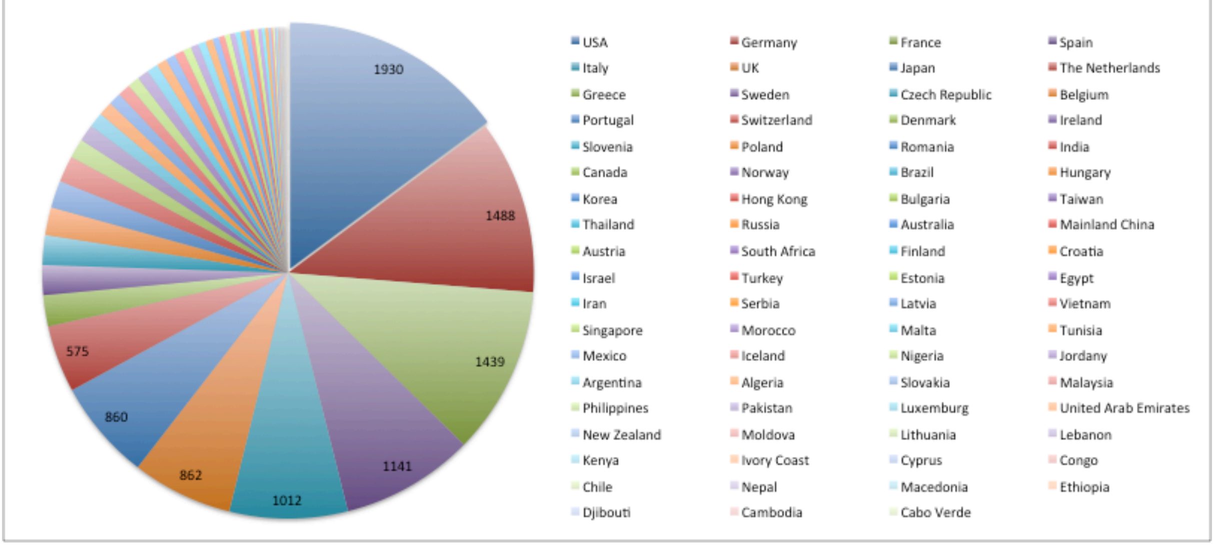

Figure 12a. *Number of signatures per country (over 1998-2012)*

The 12 most publishing countries represent 80% of the contributors: USA (15%), Germany (11%), France (11%). Spain (9%), Italy (8%), UK (7%), Japan (7%), The Netherlands (4%), Greece, Sweden, Czech Republic, Belgium, Portugal and Switzerland (2%) (Fig. 13).

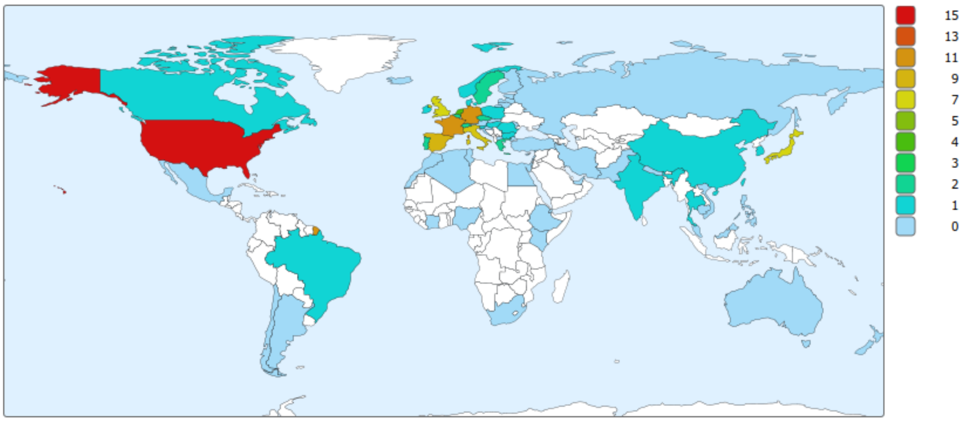

Fig. 12b. *Percentage of signatures per country (over 1998-2012)*

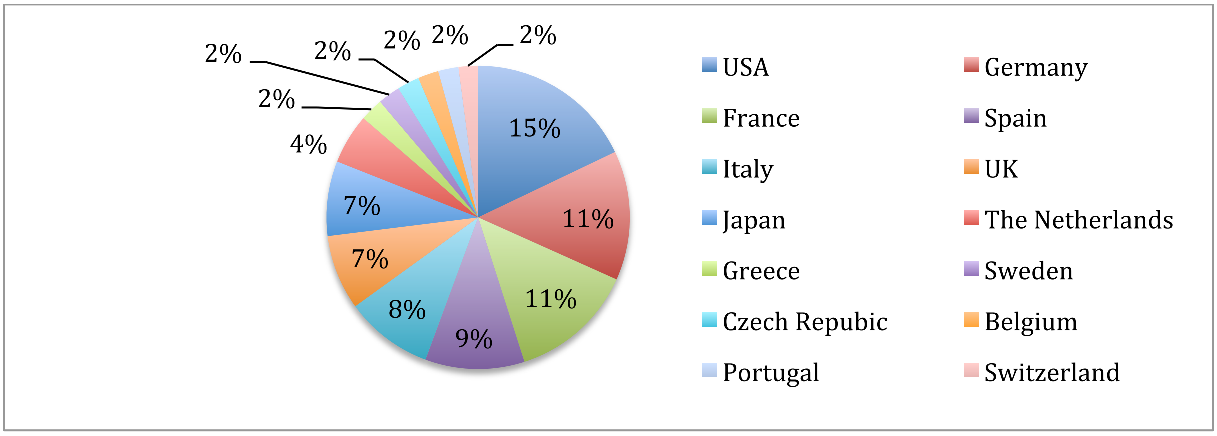

Figure 13. *Percentages of signatures per country for the 14 most cited countries*

If we consider continents, we see that Europe has the largest share (70%), followed by America (16%) and Asia (9%). Africa, Oceania and the Middle East only represent 4% (Fig. 14).

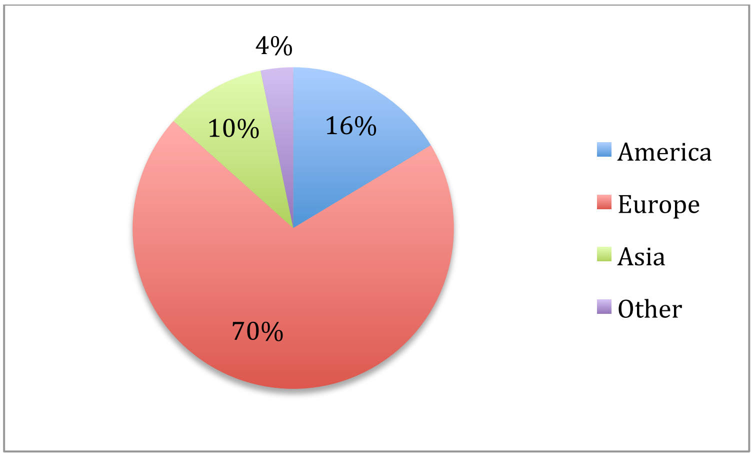

Figure 14. *Percentages of signatures according to continents.*

If we now consider the evolution of the share of signatures per country over time, for the 8 countries totaling 4% or more of the signatures overall (Fig. 15a), we see that the share of the USA remained stable at about 15%. The share of France initially ranked first and then strongly decreased, but is now back in the top 3 countries with about 12%, together with the USA and Germany, which increased its share over time. A second set at around 8% comprises Spain and Italy, both slightly decreasing their share, with a larger participation when the conference takes place in their country (Granada in 1998 and Las Palmas in 2002 for Spain, or Genoa in 2006 for Italy). UK, Japan and The Netherlands are in a third set around 5%, with large variations over time for UK and Japan, and a smoother stability for The Netherlands.

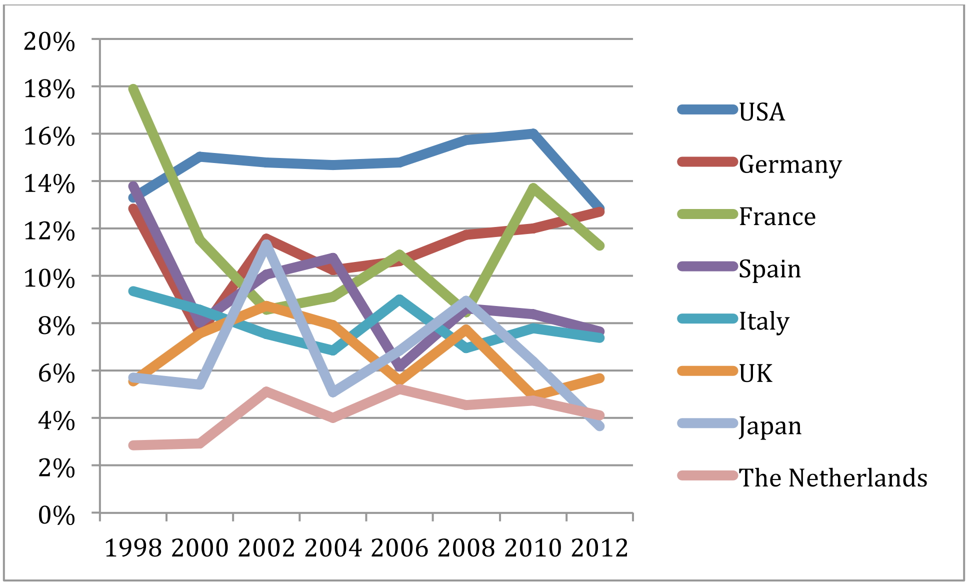

Figure 15a. *Evolution of the share of signatures per country over time for the 8 most cited countries.*

The number of publishing countries (Fig. 15b) and the share of the emerging countries, especially India, PR China and Brazil (Fig. 15c), have largely increased over the years.

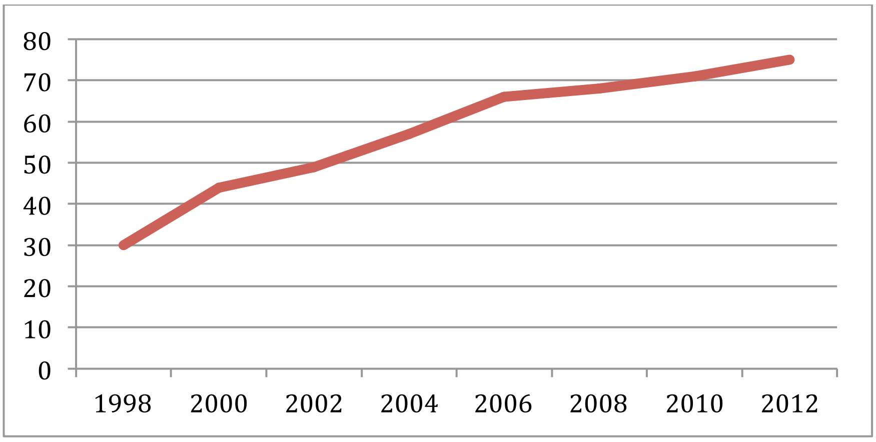

Fig. 15b *Evolution over time of the number of countries having published at LREC*

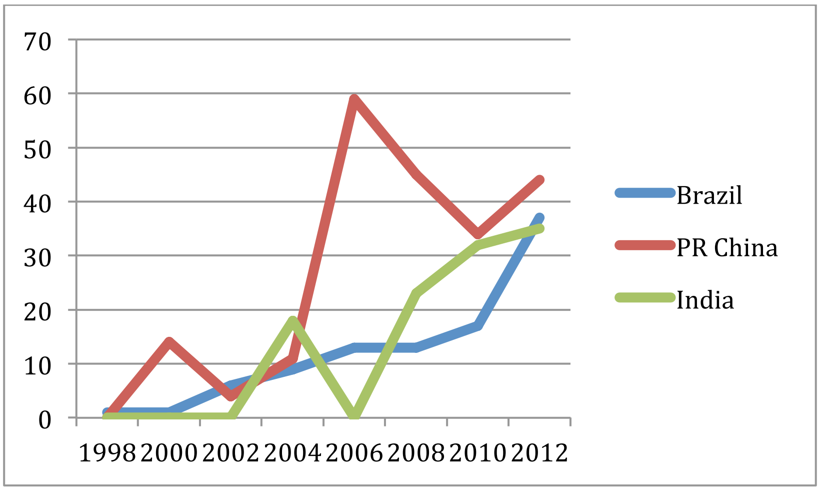

Fig. 15c *Evolution over time of the share of signatures for Brazil, PR China and India*

If we cluster the countries into "Continents" (Fig. 16), we see that the share of Europe around 70% slightly decreased over time until 2008, and increased since then, while America stayed very stable at around 20% and Asia around 10% with more fluctuations. The share of the other countries is slowly and slightly increasing, but is still very low.

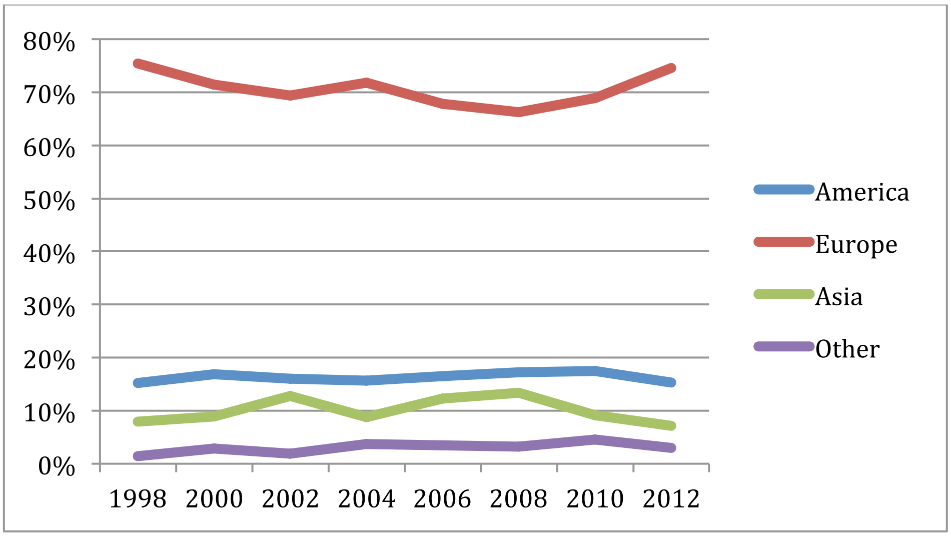

Figure 16. *Evolution of the share of signatures per continent.*

2.4.8. US States

The US contributors come from 28 different states. The 5 most active states are Pennsylvania (567 signatures), New-York State (269), California (198), Massachusetts (194) and Maryland (164) (Fig. 17).

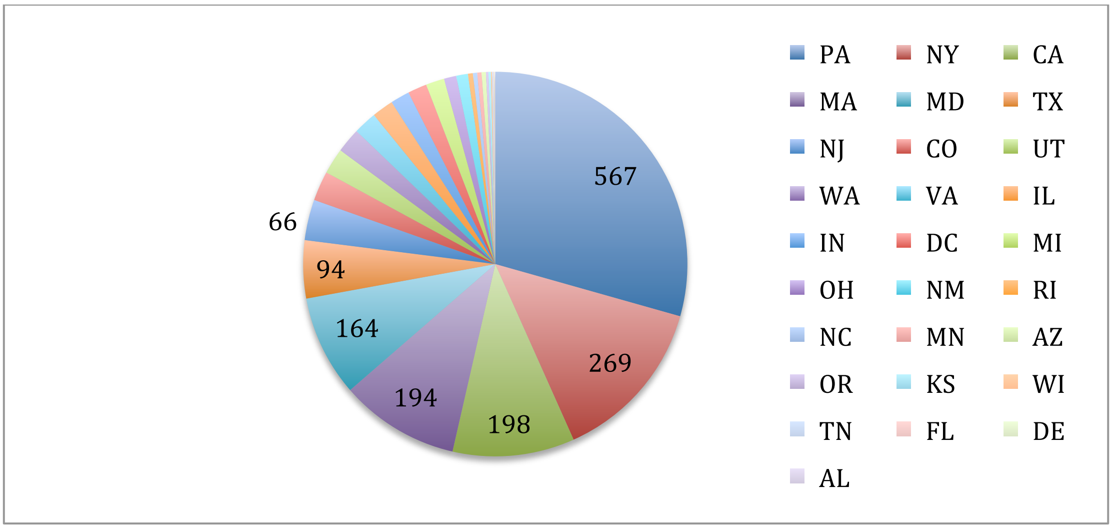

Figure 17. *Share of US signatures per state.*

2.4.9. Affiliations of authors

The authors come from 1,227 different institutions. 20 institutions represent 100 signatures or more, with a total of 3,424 signatures (representing 25% of all the signatures (Table 4)).

# signatures	Presence in # papers	Institution	Country
312	119	Istituto di Linguistica Computazionale (ILC) – CNR "A. Zampolli"	Italy
287	109	LIMSI-CNRS	France
248	99	University of Sheffield	UK
234	76	Linguistic Data Consortium (LDC)	USA
196	95	German Research Center for Artificial Intelligence (DFKI) GmbH	Germany
190	49	University of the Basque Country (UPV/EHU)	Spain
172	84	Universität des Saarlandes	Germany
172	72	Charles University in Prague	Czech Republic
170	74	Universitat Politecnica de Catalunya (UPC)	Spain
169	66	University of Pennsylvania	USA
167	57	Max Planck Institute for Psycholinguistics	The Netherlands
140	72	Evaluation and Language Resources Distribution Agency (ELDA)	France
136	68	Universität Stuttgart	Germany
134	56	Universitat Pompeu Fabra (UPF)	Spain
130	53	Carnegie Mellon University	USA
124	61	Radboud University Nijmegen	The Netherlands
121	40	Institute for Language and Speech Processing (ILSP)	Greece
116	53	Université de Genève	Switzerland
103	35	MITRE	USA
103	51	Kobenhavns Universitet	Denmark

Table 4. Top 20 Institutions with more than 100 signatures attached to those institutions.

2.4.10. Authors production

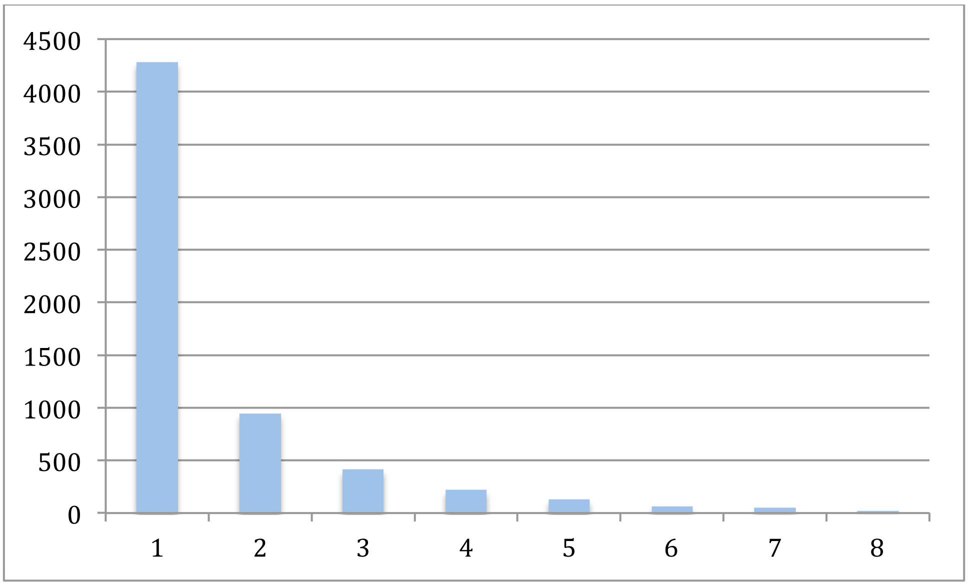

Figure 18. *Number of Authors per Number of Conferences*

20 authors published at all 8 conferences (Nicoletta CALZOLARI, Nick CAMPBELL, Khalid CHOUKRI, Christopher CIERI, Thierry DECLERCK, Robert GAIZAUSKAS, Eva HAJICOVA, Nancy IDE, Sadao KUROHASHI, Mark LIBERMAN, Bernardo MAGNINI, Simonetta MONTEMAGNI, Patrick PAROUBEK, Uwe QUASTHOFF, Bolette SANDFORD PEDERSEN, Diana SANTOS, Takenobu TOKUNAGA, Dan TUFIS, Hans USZKOREIT, Henk VAN DEN HEUVEL) (Fig. 18).

4,280 authors published at a single conference (70% of the 6,118 authors)

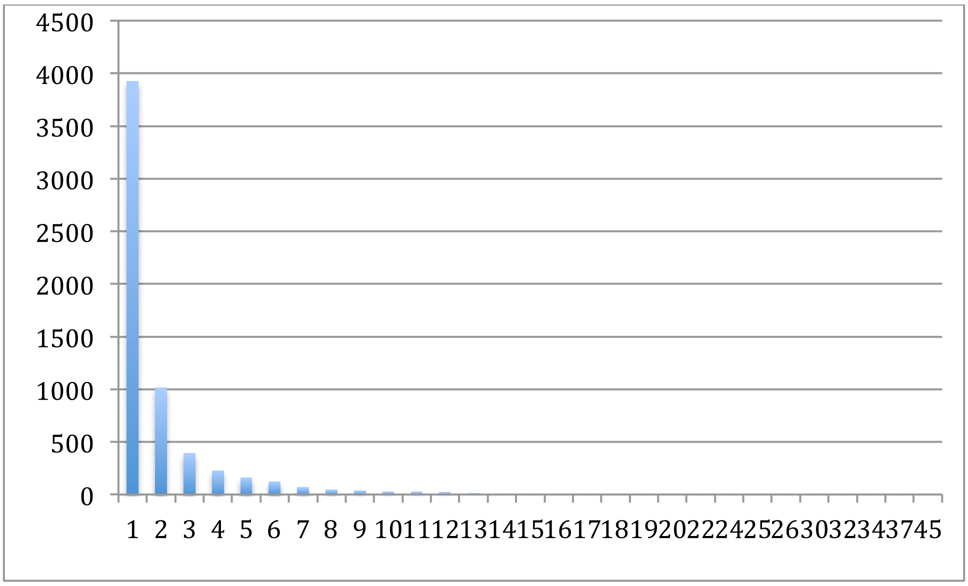

Figure 19. *Number of Papers per Number of Authors*

5 authors published 30 papers or more: Khalid CHOUKRI (45 papers), Nicoletta CALZOLARI (37), Peter WITTENBURG (34), Hitoshi ISAHARA (32), Stephanie M. STRASSEL (30).

20 authors published 20 papers or more, while 130 authors published 10 papers or more, and 3,924 (64% of the 6,118 authors) published only 1 paper (Fig. 19).

2 authors published 5 papers as single authors (Jörg Tiedemann and Nick Campbell) and 3 authors published 4 papers as single authors (Kiril Ribarov, Serge A. Yablonsky and Tomaž Erjavec), while 5,688 authors (93% of the authors) never published alone.

2.4.11. Co-authors

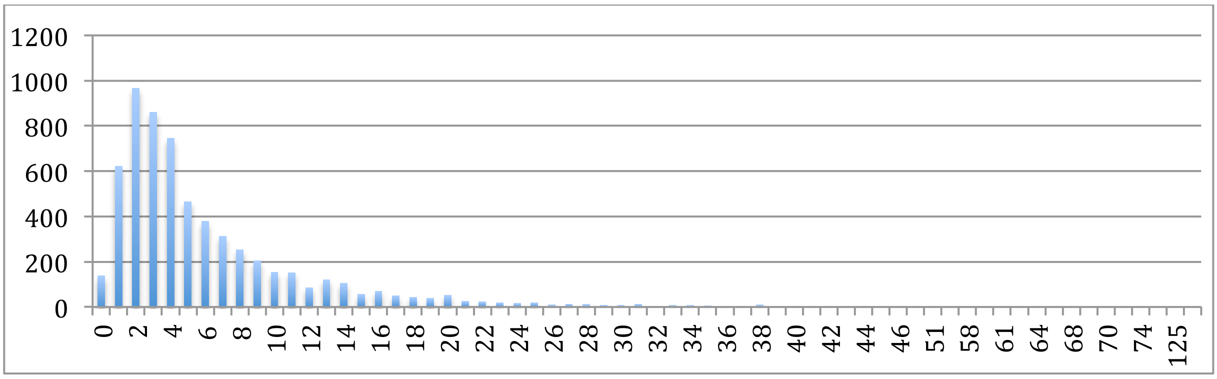

Figure 20. *Number of authors as a function of the number of different co-authors*

8 authors published with 70 or more different co-authors: Khalid CHOUKRI (145), Nicoletta CALZOLARI (125), Monica MONACHINI (82), Djamel MOSTEFA (74), Nuria BEL (74), Alessandro LENCI (73), Stephanie M. STRASSEL (70), Ulrich HEID (70), while 137 authors published alone (Fig. 20).

2.4.12. Cliques

The study of the cliques, i.e. publishing groups of authors (if author A published a paper with author B, and author B published a paper with author C, authors A, B and C belong to the same clique), extracted from the Collaboration graph that links two author nodes when they have published a paper in common, results in 453 cliques. The largest one regroups 4,758 authors, which means that 78 % of the 6,118 LREC authors are somehow connected through a publication path, i.e. have published once together. This may appear as an indicator of the cohesion of the community. The authors of this clique published 3,080 papers, e.g. 81% of the total number of papers. Each of the next two largest cliques contains only 23 authors who never published with any of the 4,758 previous ones. They published respectively 13 and 16 papers. The next clique contains 20 authors, who published 11 papers. The next 11 cliques that contain 10 authors or more published from 1 to 10 papers. Those cliques appear as small communities often related to the study of a specific language. As already mentioned, 2% of the authors (137) have published only alone (Table 5).

Clique size	# cliques	# Authors	% Authors in the cliques	% of Cliques
4,758	1	4,758	78%	0%
23	2	46	1%	0%
20	1	20	0%	0%
15	1	15	0%	0%
14	1	14	0%	0%
12	3	36	1%	1%
11	2	22	0%	0%
10	4	40	1%	1%
9	4	36	1%	1%
8	7	56	1%	2%
7	9	63	1%	2%
6	15	90	1%	3%
5	35	175	3%	8%
4	37	148	2%	8%
3	74	222	4%	16%
2	120	240	4%	26%
1	137	137	2%	30%
Total	453	6,118	100%	100%

Table 5. Cliques in the LREC Collaboration graph

2.4.13. Collaboration Graph

2.4.13.1. Definitions

In mathematics and social science, a **Collaboration graph**[18] is a graph modeling some social network where the **nodes** (or vertices) represent participants of that network (usually individual people) and where two distinct participants are joined by an **edge** whenever there is a collaborative relationship between them of a particular kind. Collaboration graphs are used to measure the closeness of collaborative relationships between the participants of the network.

By construction, the Collaboration graph is a simple graph, since it has no loop-edges and no multiple edges, and is undirected, contrary to a citation graph. The Collaboration graph need not be connected. Thus people who never co-authored a joint paper represent isolated nodes in the Collaboration graph. Those who are connected constitute a clique.

The distance, or path-length, between two people/nodes in a Collaboration graph is called the **Collaboration distance** or the **geodesic distance**. Thus the collaboration distance between two distinct nodes is equal to the smallest number of edges in an edge-path connecting them. The **Diameter** of the Collaboration graph is the largest collaboration path in that graph. If no path connecting two nodes in a collaboration graph exists, the collaboration distance between them is said to be infinite.

The **Degree** of a node is the number of edges attached to this node. It illustrates the amount of co-authors of each author. The **Density** of a graph is the fraction of possible edges that exist in a graph. It provides a measure of the intensity of collaboration. The **Clustering coefficient** of a node is a measure of the degree to which nodes in a graph tend to cluster together, as the fraction of possible edges linked to that node.

2.4.13.2. Measures for LREC and other conferences

We computed those measures for the LREC Anthology and compared with the same measures that we computed for other conferences: the francophone TALN conference series (1997-2013) organized by ATALA and contained in the TALN Archive, the ISCA Conference Archive (1987-2012), including the ECST, Eurospeech, ICSLP and Interspeech conference series, the International ACL (1979-2013), EMNLP (1996-2013) and Coling (1965-2012) conferences, contained in the ACL Anthology[19] and we also included similar measures which are available online for the ACL Anthology as a whole[20], and at the SNAP[21], for Astrophysics, where the number of authors is similar to ISCA, and for General Relativity, where the number of authors is similar to LREC (Table 6) (for more details see 1. Introduction).

Looking at Table 6 and considering language and speech processing, we see that the number of papers as well as the number of different authors, is similar in the ISCA Archive and in the ACL Anthology, and in the LREC, ACL and Coling conference series. The level of collaborations (Average Degree) is much larger at LREC than at the ACL, EMNLP or Coling conferences, and comparable to the ISCA conferences. The largest clique at ISCA conferences gathers 84% of all authors, while it gathers 78% at LREC and only 61% at the ACL International conferences, 64% at EMNLP and 43% at Coling. However LREC has more cliques of intermediary sizes (10 to 25 authors) than ISCA, expressing the existence of small communities working on specific languages. The Average Path length is somehow correlated to the importance of the largest clique. The diameters are very close for LREC, TALN, ISCA and larger for ACL, EMNLP and Coling. The density is larger for LREC and EMNLP, and even more for TALN, reflecting the stronger relationship of the respective communities. It is lower for the ISCA conferences and for the ACL Anthology which is 3 times less dense, given that this anthology regroups 13 different conferences and journals. The Average Clustering Coefficient is similar for all conferences, LREC having the highest score. It was also striking to see that LREC has 4 times less papers than ISCA, while it has only a little bit more than half different authors. We therefore introduced a "Productivity" measure giving the average number of papers by different authors, but which may also be considered as an author "Redundancy"

[18] http://en.wikipedia.org/wiki/Collaboration_graph
[19] http://aclweb.org/anthology/
[20] http://clair.eecs.umich.edu/aan/index.php
[21] http://snap.stanford.edu/data/

measure. Given that the ACL Anthology includes a large set of conferences of interest for the same authors, it has a large productivity/redundancy score.

	LREC Anthology	TALN (francophone)	ISCA Archive	ACL International Conferences	Coling	EMNLP	ACL Anthology	SNAP Astrophysics	SNAP General Relativity
Time span (years)	15	17	26	35	48	18	35	11	11
# venues	8	17	25	35	20	18	342 (20 conferences and journals)	n.a.	n.a.
# papers	3,808	937	16,206	4,029	3,700	1,503	21,212	n.a.	n.a.
# Nodes (# authors)	6,118	1,103	14,583	5,041	5,163	2,339	17,792	18,772	5,242
# Edges (collaborate)	19,629	2,192	43,397	9,546	8,273	4,429	49,561	19,811	14,496
Authors productivity/redundancy	0.62	0.85	1.11	0.80	0.72	0.64	1.32	n.a.	n.a.
Max Degree (# co-authors)	145	31	169	63	64	49	n.a.	n.a.	n.a.
Average Degree	6,42	3.97	5.95	3.79	3.21	3 .79	6,16	n.a.	n.a.
Number of cliques	453	162	953	916	1,18	272	n.a.	n.a.	n.a.
Largest clique (%authors)	4,758 (78%)	702 (64%)	12,295 (84%)	3,088 (61%)	2,233 (43%)	1,508 (64%)	13,259 (82%)	17,903 (95%)	4,158 (79%)
Diameter	16	14	15	19	21	19	15	14	17
Average Path Length	3.51	2.29	4.01	2.48	1.47	3.03	n.a.	n.a.	n.a.
Density	0.0011	0.0036	0.0004	0.0008	0.0006	0.0016	0.0004	0.0011	0.0010
Average Clustering Coefficient	0.73	0.59	0.65	0.58	0.56	0.65	n.a.	0.63	0.53

Table 6. Comparison across various conferences in Language science and technology and elsewhere
(n.a.: not available)

Looking at the SNAP conferences, we see that the degree of collaboration is much higher in Astrophysics, where the largest cliques includes 95% of all authors and where it seems that almost everyone collaborated with each other one day or another! The indicators in the area of General relativity are very similar to those of LREC, with a comparable time span and number of different authors, apart from the Average Clustering Coefficient, which is higher for LREC, showing a slightly higher level of collaboration.

2.4.13.3. Measures of Centrality

Our aim is now to explore the role of each author in the Collaboration Graph, trying to assess its influence. In Graph Theory, there exist several types of Centrality Measures (L. Freeman, 1978). **Closeness Centrality** was introduced in Human Sciences by A. Bavelas to express the efficiency of a communication Network (A. Bavelas, 1948 and A. Bavelas, 1950). It is based on the shortest distance between an author and another author whatever the number of collaborations between the two authors. The Closeness Distance is then computed as the average geodesic distance of that author with all the other authors belonging to the same clique. It is very close to the **Average Distance** of a node measure. The exact formula that we use is the harmonic centrality. The **Degree Centrality** is simply the number of co-authors of each author, or the number of edges attached to the corresponding node. The **Betweenness Centrality** is based on the number of paths crossing a node. It reflects the importance of an author as a bridge across different sets of authors, or communities.

Looking at Table 7, we see that the ranking may drastically change for some authors depending on the kind of Centrality that is considered. If we only consider the Closeness Centrality, some authors who have high ranking in the other types of centrality may not appear (see those marked in grey in Table 7). In order to analyze the reasons for those differences, we considered the relationship between the Closeness Centrality and various measures attached to the authors' productivity in a collaborative environment: number of collaborations, number of articles, number of different collaborators and number of articles published as single author (Table 8). It appears that the Degree Centrality better illustrates the collaborative activity of the authors, while the Betweeness Centrality is somehow correlated with the productivity of the authors. The Closeness Centrality tends to favor those who publish together with very active authors. On the contrary, authors who are very active but publish with fewer co-authors or with less active co-authors, are under-represented even if they are very central in a sub-community.

	Closeness Centrality			Degree Centrality			Betweenness Centrality			Average distance	
Author's name	Rank	Harmonic Centrality	Norm on first	Rank	Index	Norm on first	Rank	Index	Norm on first	Rank	Average distance
Nicoletta Calzolari	1	1604	**1,00**	2	0,95	**0,95**	2	165085	**0,74**	1	**3,41**
Khalid Choukri	2	1576	**0,98**	1	1,00	**1,00**	1	224316	**1,00**	2	**3,53**
Monica Monachini	3	1472	**0,92**	3	0,58	**0,58**	21	61157	**0,27**	3	**3,67**
Laurent Romary	4	1468	**0,92**	14	0,40	**0,40**	9	83727	**0,37**	4	**3,68**
Núria Bel	5	1465	**0,91**	11	0,44	**0,44**	7	86089	**0,38**	6	**3,69**
Alessandro Lenci	6	1464	**0,91**	6	0,48	**0,48**	10	83715	**0,37**	5	**3,69**
Stelios Piperidis	7	1456	**0,91**	9	0,45	**0,45**	27	57795	**0,26**	7	**3,70**
Claudia Soria	8	1443	**0,90**	10	0,45	**0,45**	30	56212	**0,25**	8	**3,72**
Bente Maegaard	9	1415	**0,88**	22	0,32	**0,32**	14	67044	**0,30**	9	**3,80**
Martha Palmer	10	1407	**0,88**	24	0,29	**0,29**	4	97442	**0,43**	11	**3,82**
Djamel Mostefa	11	1402	**0,87**	11	0,44	**0,44**	19	64229	**0,29**	17	**3,89**
Nancy Ide	12	1401	**0,87**	38	0,23	**0,23**	62	35879	**0,16**	10	**3,80**
Thierry Declerck	13	1392	**0,87**	23	0,30	**0,30**	12	71110	**0,32**	13	**3,85**
Peter Wittenburg	14	1386	**0,86**	4	0,56	**0,56**	8	84155	**0,38**	16	**3,89**
Dan Tufiş	15	1381	**0,86**	24	0,29	**0,29**	13	69177	**0,31**	14	**3,87**
Valérie Mapelli	16	1378	**0,86**	41	0,23	**0,23**	163	20053	**0,09**	15	**3,87**
Maria Gavriilidou	17	1378	**0,86**	86	0,18	**0,18**	252	14489	**0,07**	12	**3,85**
Olivier Hamon	18	1373	**0,86**	44	0,22	**0,22**	117	25065	**0,11**	18	**3,89**
Christopher Cieri	19	1358	**0,85**	21	0,33	**0,33**	24	60203	**0,27**	19	**3,91**
Patrick Paroubek	20	1356	**0,85**	19	0,34	**0,34**	45	43655	**0,20**	31	**3,99**
Bernardo Magnini	21	1355	**0,85**	18	0,34	**0,34**	11	83555	**0,37**	20	**3,93**
Hitoshi Isahara	23	1336	**0,83**	8	0,46	**0,46**	6	90464	**0,40**	34	**4,01**
Sophie Rosset	30	1328	**0,83**	15	0,39	**0,39**	37	49360	**0,22**	47	**4,06**
Daan Broeder	33	1326	**0,83**	16	0,38	**0,38**	42	44723	**0,20**	33	**4,01**
Lori Lamel	45	1302	**0,81**	27	0,26	**0,26**	20	61245	**0,27**	77	**4,13**
Ulrich Heid	56	1294	**0,81**	17	0,37	**0,37**	5	95955	**0,43**	81	**4,14**
Asunción Moreno	57	1289	**0,80**	7	0,48	**0,48**	22	61045	**0,27**	111	**4,22**
Henk van den Heuvel	97	1256	**0,78**	13	0,43	**0,43**	18	64526	**0,29**	169	**4,32**
Stephanie M Strassel	99	1255	**0,78**	5	0,50	**0,50**	3	104638	**0,47**	161	**4,30**
Steven Bird	128	1238	**0,77**	49	0,22	**0,22**	15	66916	**0,30**	153	**4,28**
Nikos Fakotakis	161	1215	**0,76**	19	0,34	**0,34**	16	66472	**0,30**	231	**4,41**
Hans Uszkoreit	168	1208	**0,75**	52	0,21	**0,21**	17	64881	**0,29**	207	**4,38**

Table 7. Computation and comparison of the Closeness Centrality, Degree Centrality, Betweenness Centrality and Average Distance for the most central authors. The authors are ranked according to the Closeness Centrality measure, with a selection of the 20 authors top ranked with that measure, and of the ones ranked among the top 20 for the other measures while being ranked in the top 200 according to the Closeness Centrality.

Authors' Name	Closeness Centrality			Collaborations		Different collaborators		number of Articles (= number of signatures)		number of articles as single author	
	Rank	Harmonic Centrality	Norm on first	Rank	#	Rank	#	Rank	#	Rank	#
Nicoletta Calzolari	1	1604	**1,00**	2	**227**	2	**125**	2	**37**	77	**1**
Khalid Choukri	2	1576	**0,98**	1	**240**	1	**145**	1	**45**	77	**1**
Monica Monachini	3	1472	**0,92**	3	**139**	3	**82**	14	**22**	431	**0**
Laurent Romary	4	1468	**0,92**	14	**95**	12	**67**	6	**26**	431	**0**
Núria Bel	5	1465	**0,91**	11	**105**	4	**74**	14	**22**	77	**1**
Alessandro Lenci	6	1464	**0,91**	6	**116**	6	**73**	14	**22**	431	**0**
Stelios Piperidis	7	1456	**0,91**	9	**109**	16	**63**	22	**18**	77	**1**
Claudia Soria	8	1443	**0,90**	10	**108**	9	**69**	25	**17**	431	**0**
Bente Maegaard	9	1415	**0,88**	22	**77**	17	**61**	61	**12**	77	**1**
Martha Palmer	10	1407	**0,88**	24	**70**	18	**59**	25	**17**	431	**0**
Djamel Mostefa	11	1402	**0,87**	11	**105**	4	**74**	19	**20**	431	**0**
Nancy Ide	12	1401	**0,87**	38	**56**	70	**33**	10	**24**	77	**1**
Thierry Declerck	13	1392	**0,87**	23	**71**	20	**58**	21	**19**	19	**2**
Peter Wittenburg	14	1386	**0,86**	4	**135**	12	**67**	3	**34**	431	**0**
Dan Tufiş	15	1381	**0,86**	24	**70**	23	**51**	10	**24**	19	**2**
Valérie Mapelli	16	1378	**0,86**	41	**54**	49	**37**	107	**10**	431	**0**
Maria Gavriilidou	17	1378	**0,86**	86	**43**	99	**29**	278	**6**	431	**0**
Olivier Hamon	18	1373	**0,86**	44	**53**	40	**38**	43	**14**	77	**1**
Christopher Cieri	19	1358	**0,85**	21	**78**	49	**37**	6	**26**	431	**0**
Patrick Paroubek	20	1356	**0,85**	19	**81**	21	**52**	29	**16**	77	**1**
Bernardo Magnini	21	1355	**0,85**	18	**82**	14	**64**	38	**15**	431	**0**
Hitoshi Isahara	23	1336	**0,83**	8	**111**	9	**69**	4	**32**	77	**1**
Sophie Rosset	30	1328	**0,83**	15	**94**	18	**59**	22	**18**	431	**0**
Daan Broeder	33	1326	**0,83**	16	**92**	26	**46**	9	**25**	431	**0**
Lori Lamel	45	1302	**0,81**	27	**63**	21	**52**	107	**10**	431	**0**
Ulrich Heid	56	1294	**0,81**	17	**88**	7	**70**	6	**26**	77	**1**
Asunción Moreno	57	1289	**0,80**	7	**114**	11	**68**	10	**24**	431	**0**
Henk van den Heuvel	97	1256	**0,78**	13	**102**	14	**64**	10	**24**	431	**0**
Stephanie M Strassel	99	1255	**0,78**	5	**120**	7	**70**	5	**30**	77	**1**
Steven Bird	128	1238	**0,77**	49	**52**	34	**41**	38	**15**	431	**0**
Nikos Fakotakis	161	1215	**0,76**	19	**81**	25	**49**	14	**22**	19	**2**
Hans Uszkoreit	168	1208	**0,75**	52	**51**	30	**43**	51	**13**	431	**0**

Table 8. Comparison of the Closeness Centrality with measures of the collaborativeness and productivity of authors.

We believe the influence of an author may be related to the number of publications, the number of collaborations, and especially with co-authors who themselves publish and collaborate a lot, their ability to innovate and the impact factor of their papers (number of citations, that we didn't consider for the time being in this paper). It seemed that weighting the nodes with the number of published papers, or by the number of collaborations (considering however only the papers published with a least one co-author), would better reflect the essence of the network, as it appears in Table 9. We think that the Harmonic Centrality weighted by the number of articles gives the best picture of the collaboration graph essence, as it allows for placing on the forefront representatives of various trends in Language Resource and Evaluation.

Authors' name	Harmonic Centrality weighted by Articles			Harmonic Centrality weighted by Collaborations			Harmonic Centrality		
	Rank	Index	Norm on first	Rank	Index	Norm on first	Rank	Index	Norm on first
Khalid Choukri	1	70909	**1,00**	1	378182	**1,00**	2	1576	**0,98**
Nicoletta Calzolari	2	59354	**0,84**	2	364146	**0,96**	1	1604	**1,00**
Peter Wittenburg	3	47110	**0,66**	4	187055	**0,50**	14	1386	**0,86**
Hitoshi Isahara	4	42752	**0,60**	10	148296	**0,39**	23	1336	**0,83**
Laurent Romary	5	38160	**0,54**	13	139431	**0,37**	4	1468	**0,92**
Stephanie M Strassel	6	37660	**0,53**	9	150641	**0,40**	99	1255	**0,78**
Christopher Cieri	7	35301	**0,50**	21	105902	**0,28**	19	1358	**0,85**
Ulrich Heid	8	33643	**0,47**	17	113868	**0,30**	56	1294	**0,81**
Nancy Ide	9	33631	**0,47**	28	78472	**0,21**	12	1401	**0,87**
Dan Tufiş	10	33150	**0,47**	25	96688	**0,26**	15	1381	**0,86**
Daan Broeder	11	33150	**0,47**	16	121991	**0,32**	33	1326	**0,83**
Monica Monachini	12	32395	**0,46**	3	204676	**0,54**	3	1472	**0,92**
Núria Bel	13	32239	**0,46**	8	153869	**0,41**	5	1465	**0,91**
Alessandro Lenci	14	32199	**0,45**	5	169779	**0,45**	6	1464	**0,91**
Asunción Moreno	15	30946	**0,44**	12	146995	**0,39**	57	1289	**0,80**
Henk van den Heuvel	16	30148	**0,43**	14	128130	**0,34**	97	1256	**0,78**
Djamel Mostefa	17	28040	**0,40**	11	147212	**0,39**	11	1402	**0,87**
Nikos Fakotakis	18	26725	**0,38**	24	98397	**0,26**	161	1215	**0,76**
Thierry Declerck	19	26452	**0,37**	22	98847	**0,26**	13	1392	**0,87**
Stelios Piperidis	20	26206	**0,37**	6	158694	**0,42**	7	1456	**0,91**
Claudia Soria	23	24530	**0,35**	7	155835	**0,41**	8	1443	**0,90**
Sophie Rosset	25	23911	**0,34**	15	124871	**0,33**	30	1328	**0,83**
Patrick Paroubek	26	21696	**0,31**	19	109838	**0,29**	20	1356	**0,85**
Bernardo Magnini	29	20329	**0,29**	18	111130	**0,29**	21	1355	**0,85**
Olivier Hamon	33	19215	**0,27**	35	72744	**0,19**	18	1373	**0,86**
Bente Maegaard	42	16976	**0,24**	20	108932	**0,29**	9	1415	**0,88**
Valérie Mapelli	74	13782	**0,19**	32	74425	**0,20**	16	1378	**0,86**
Maria Gavriilidou	206	8267	**0,12**	61	59250	**0,16**	17	1378	**0,86**

Table 9. Computation and comparison of Closeness Centrality unweighted or weighted by the number of articles or by the number of collaborations. The authors are ranked according to the Closeness Centrality weighted by the number of articles, in the same way as in Table 7.

2.4.13.4. Visualization of the Collaboration Graph

Here are some views of the LREC Collaboration Graph, obtained with the iGraph Software (Csàrdi et al., 2006). Figure 21 provides a complete view of the LREC Anthology, using the *Fruchterman-Reingolg layout* (T.M. Fruchterman and E.M. Reingold, 1991). The center is constituted by the large main clique, while the periphery is constituted by the smallest cliques. The central node corresponds to the most central author, N. Calzolari, and her coauthor nodes appear in orange.

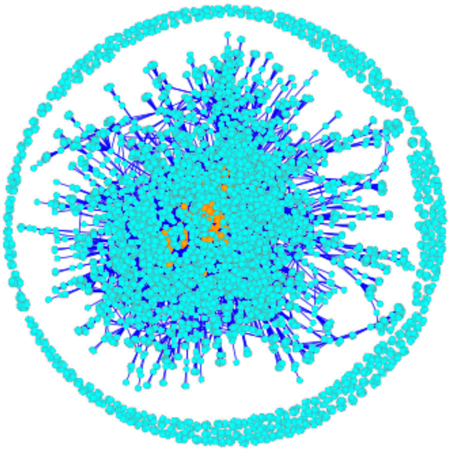

Figure 21. LREC Anthology Collaboration Graph
(order one NC coauthor nodes in orange, other nodes in light blue, all edges in blue)

Figure 22 provides a view of the second order sub-network attached to N. Calzolari (her co-authors, and the co-authors of her co-authors).

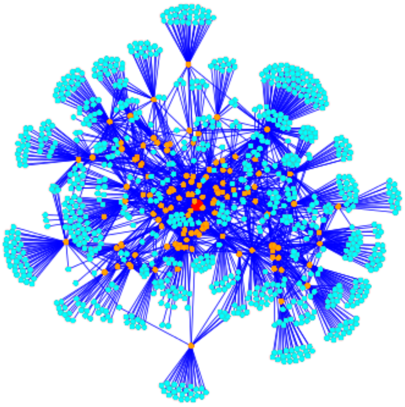

Figure 22. 2nd order sub-network attached to N. Calzolari
(NC node in red, order one NC coauthor nodes in orange, order 2 coauthor nodes in light blue, all edges in blue)

Figure 23 presents the first order sub-network attached to N. Calzolari (the 136 co-authors who once published a joint paper with her) and Figure 24 presents those 136 co-authors.

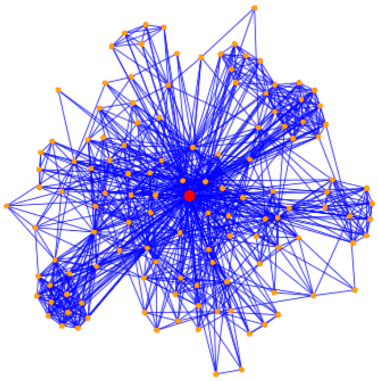

Fig. 23 1st order sub-network attached to N. Calzolari
(NC node in red, coauthor nodes in orange, edges in deep blue)

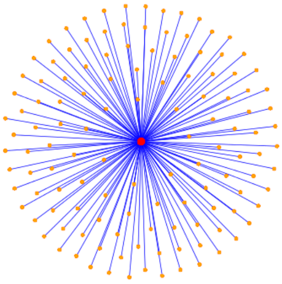

Figure 24. 125 N. Calzolari's co-authors
(NC node in red, coauthor nodes in orange, edges in deep blue)

2.5. Citations

We studied the citations in the content of the papers that are accessible in their digital form, either after OCRization for the 1998 proceedings or directly from 2000 to 2012. Given the lack of time, we only studied the citation of funding agencies, and postponed the analysis of the cited authors and cited papers sources (various conferences, journals or books) that we conducted in the case of the ISCA Archive. We also still miss for both archives the study of the cited papers, including the opinion analysis of the citing context as it has been performed for the ACL archive.

2.5.1. Most cited Funding Agencies

We studied the mention of the Funding Agencies appearing in acknowledgment constructions within the papers (e.g. "supported/funded/financed by…", "support/funding/grant/fellowship from/of…"), in order to estimate the support of public research funding in the different countries, to study later on the way it is organized within those different countries, and analyze whether this funding has an influence on the research topics. We should stress that it may also reflect the requirements of the various agencies to acknowledge their support, or the habits in various countries.

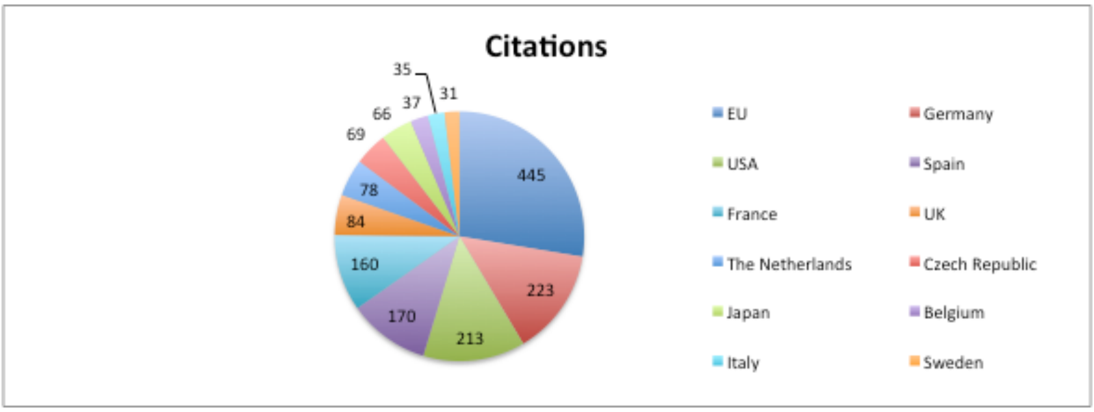

Figure 25. Share of the funding acknowledgement for the 12 most cited countries

If we consider the 12 most cited countries (Fig. 25), we see that the EU at the EC level ranks first (with more than 400 citations), followed by a set comprising Germany and the USA (about 200 citations). A third set comprises Spain and France (160-170). It is followed by a set of 4 (UK, The Netherlands, the Czech Republic and Japan) (60-80) and finally by a set of 3 (Belgium, Italy and Sweden) (30-40).

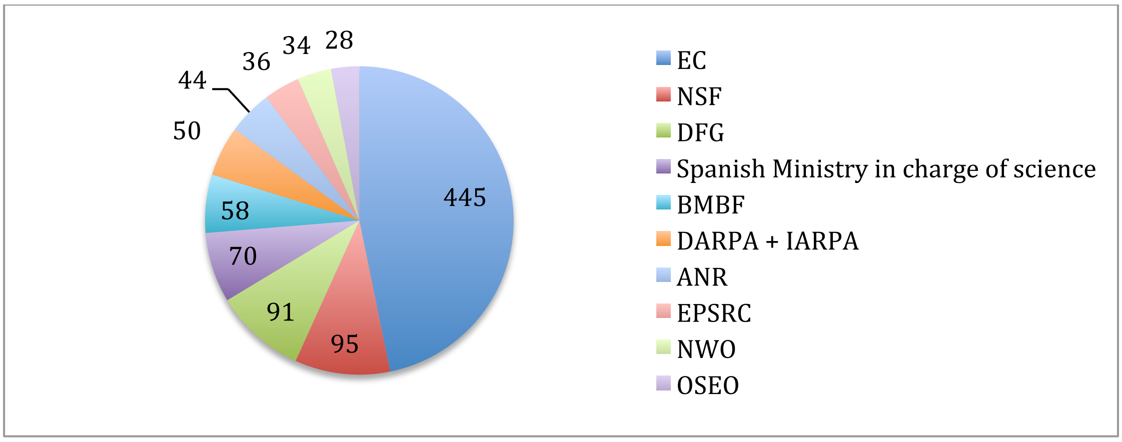

Figure 26. Share of the funding acknowledgement for the 10 most cited agencies

If we now consider the 10 most cited agencies (Fig. 26), we see that the European Commission (EC) comes by far at the first rank, with its various programs. Next come the US National Science Foundation (NSF) and the German Research Foundation (*Deutsche Forschungsgemeinschaft* (DFG)). They are followed by the Spanish Ministry in charge of Science, the German Ministry of Education and Research (*Bundesministerium für Bildung und Forschung* (BMBF)), the US Department of Defense DARPA and IARPA agencies, the French National Research Agency (*Agence Nationale de la Recherche* (ANR)) created in 2005, the UK Engineering and Physical Science Research Council (EPSRC), the Netherlands Organization for Scientific Research (NWO) and the French OSEO, in charge of more industrially oriented research funding.

2.6. Topics

2.6.1. Term based topic analysis

Our objectives were twofold: i) to compute the most frequent terms of the domain, ii) to study their variation over time.

Just as for the study of citations, our initial input is the textual content of the papers, available in a digital format apart from the proceedings of 1998 and a small set of papers coming from the other years, which had to be OCRized. Over these 15 years, the archives contain a grand total of 14,004,022 words, mostly in English, as it appears in Table 2.

As our aim is to study the terms of the Natural Language Processing domain, we did not want to get noise from some frequent formula "ordinarily" used in the English language. We adopted a contrastive approach with the same strategy implemented in TermoStat (P. Drouin, 2004). For this purpose, as a first step, we processed a vast amount of "ordinary" English texts in order to compute a statistical language profile. More precisely, we applied a deep syntactic parser called TagParser[22] and got the noun phrases. For each sentence, we kept only the noun phrases with a plain noun as a head, thus excluding the situations where a pronoun, a date or a number is the head. We also made a special dispatching for co-ordinations. We retained the various combinations of sequence of adjectives, prepositions and nouns excluding initial determiners according to unigrams, bigrams and trigrams sequences, and we stored the resulting statistical language model. This process was applied on a corpus gathering the British National Corpus (aka BNC)[23], the Open American National Corpus (aka OANC[24]), the Suzanne corpus release-5[25], the English EuroParl archives (years 1999 until 2009)[26], plus a small collection of newspapers in the domain of sports, politics and economy. The total of words was 200M words. It should be noted that, in selecting this corpus, we took care to avoid any text dealing with Natural Language Processing.

In a second step, we parsed the LREC Anthology with the same filters and used our language model to distinguish LREC specific terms from common ones. In other words, we made the hypothesis that when a sequence of words is INSIDE the Anthology and NOT INSIDE the "ordinary" profile, we consider that this term

[22] www.tagmatica.com
[23] www.natcorp.ox.ac.uk
[24] www.americannationalcorpus.org
[25] www.grsampson.net/Resources.html
[26] www.statmt.org/europarl

is specific to the field of Language Resources and Evaluation. The 14,004,022-word content reduced to 1,344,129 terms occurrences, provided that this number counts all the occurrences of all the sizes and does not restrict to the longest terms, thus counting a great number of overlapping situations between fragments of texts.

The twenty most frequent terms in Language Resources and Evaluation were computed over the period of 15 years, with the following strategy. First, the most frequent terms were computed in a raw manner, and secondly the synonyms sets (aka synsets) for all most 50 frequent terms of each year (which are frequently the same from one year to another) were manually declared in the lexicon of TagParser. Around the term synset, we gathered the variation in upper/lower case, singular/plural number, US/UK difference, abbreviation/expanded form and absence/presence of a semantically neutral adjective, like "artificial" in "artificial neural network". Thirdly, the most frequent terms were recomputed with the amended lexicon. This processing took 3 hours on a mid-range workstation (a Dell Precision workstation based on a single Xeon E3-1270V2 with 32 Gb of RAM) and gave the results that follow.

The 20 most frequent terms (lemmas) over time (1998-2012) are the following (Table 10):

Terms and variants	# Occurrences	Frequency (%)
annotation : annotation(s)	19315	1.44
POS : POS(s) | Part(s) Of Speech | Part(s) of Speech | Part-Of-Speech | Part-of-Speech | Pos | part(s) of speech | part-of-speech	6642	0.49
annotator : annotator(s)	4705	0.35
ontology : ontology | ontologies	4634	0.34
parser : parser(s)	3727	0.28
NP : NP(s)	3538	0.26
WordNet : WordNet(s) | Wordnet(s) | wordnet(s)	3166	0.24
tagger : tagger(s)	3091	0.23
XML : XML(s) | Extensible Markup Language	2908	0.22
synset : synset(s)	2811	0.21
lemma : lemma(s)	2758	0.21
segmentation : segmentation(s)	2641	0.20
metric : metric(s)	2593	0.19
MT : MT(s) | Machine Translation(s) | machine translation(s)	2340	0.17
semantic : semantic	2270	0.17
treebank : treebank(s)	2150	0.16
classifier : classifier(s)	2073	0.15
predicate : predicate(s)	1886	0.14
syntactic : syntactic	1868	0.14
metadata : metadata(s) | meta-data(s) | meta-datum | metadatum	1865	0.14

Table 10. 20 most frequent terms overall

2.6.2. Change in Topics

We studied the ranking among the 50 most popular terms (mixing unigrams, bigrams and trigrams) representing several topics of interest. The terms are followed by their ranking in 1998 and 2012 (Rank 1998/Rank 2012).

2.6.2.1. Keywords remaining popular (Fig. 27)

We studied in this category the following keywords, which stayed in the 20 top over 15 years: *Annotation* (1/1), *Parser* (4/4), POS (2/2), *Wordnet* (10/15), *NP* (3/17), *Tagger* (5/12) and *Lemma* (12/5).

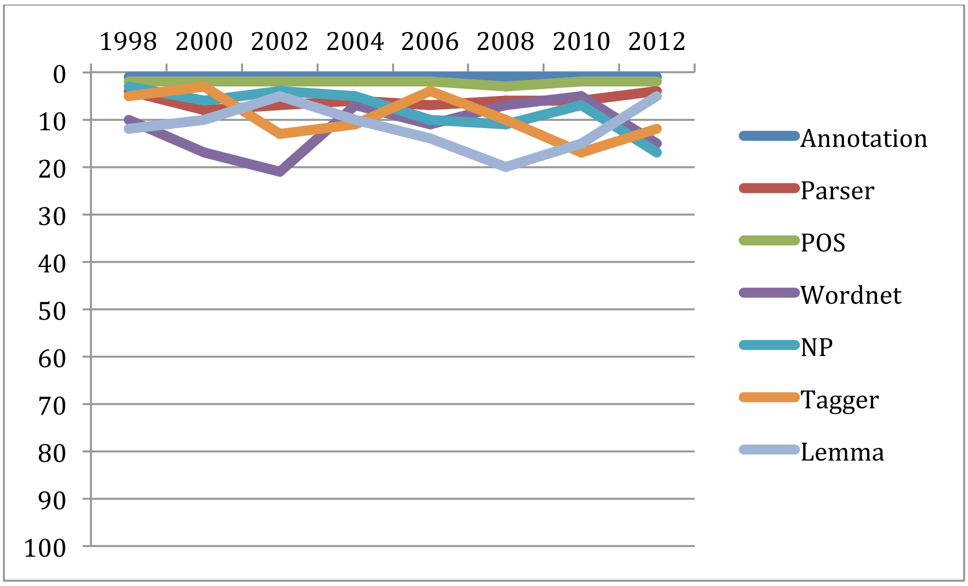

Figure 27. *Terms remaining popular*

2.6.2.2. Keywords becoming popular (Fig. 28)

We studied in this category the following keywords, which became more and more popular over time: *Annotator* (36/3), *Synset* (28/16), *XML* (Less than 100/7), *Wikipedia* (Less than 100/14), *Metadata* (Less than 100/9), *Treebank* (86/13).

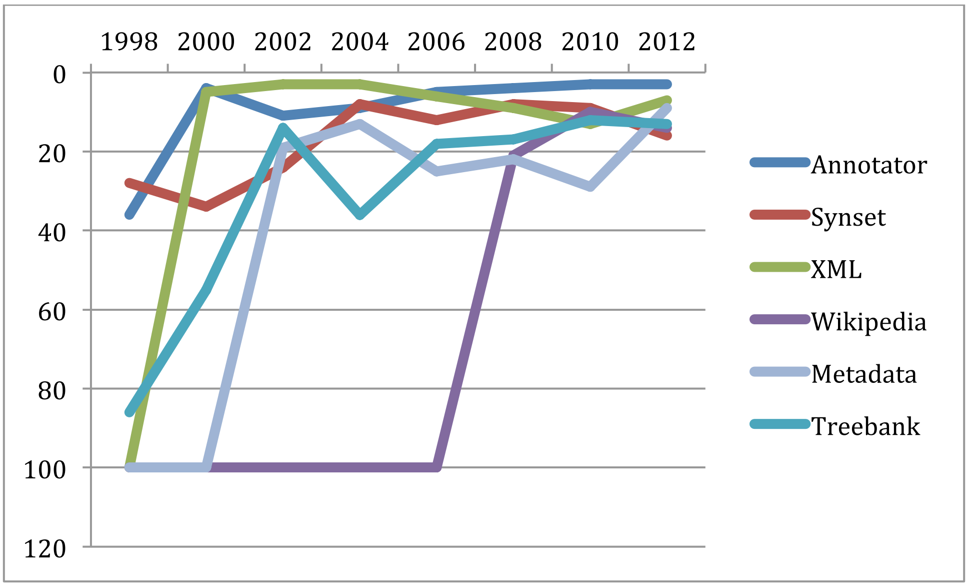

Figure 28. *Terms becoming popular*

2.6.2.3. Keywords losing popularity (Fig. 29)

We studied in this category *Encoding* (6/41) and *Markup* (20/Less than 100).

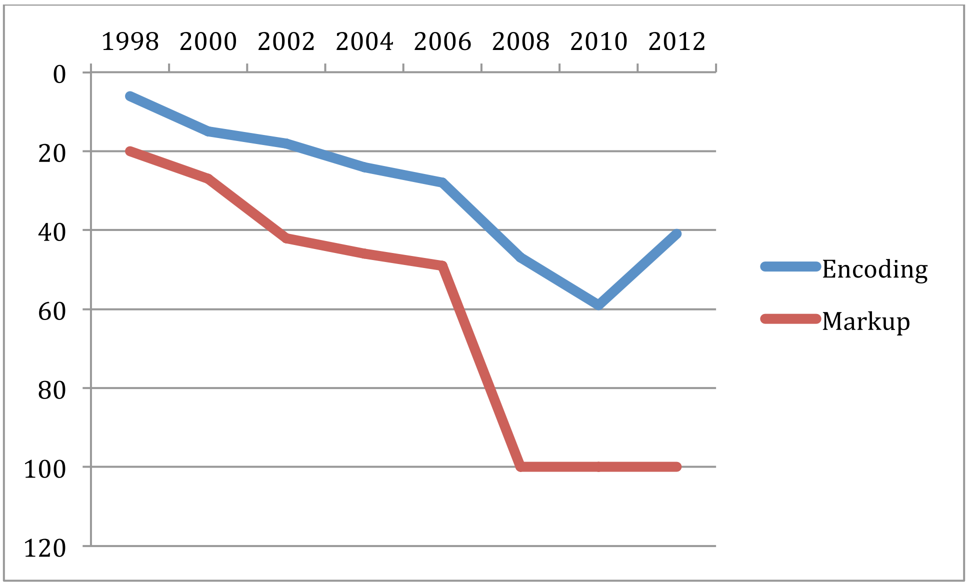

Figure 29. *Terms losing popularity*

We studied especially the disappearing of the term “*SGML*” (7/less than 100) replaced by “*XML*” (Less than 100/7) (Fig. 30).

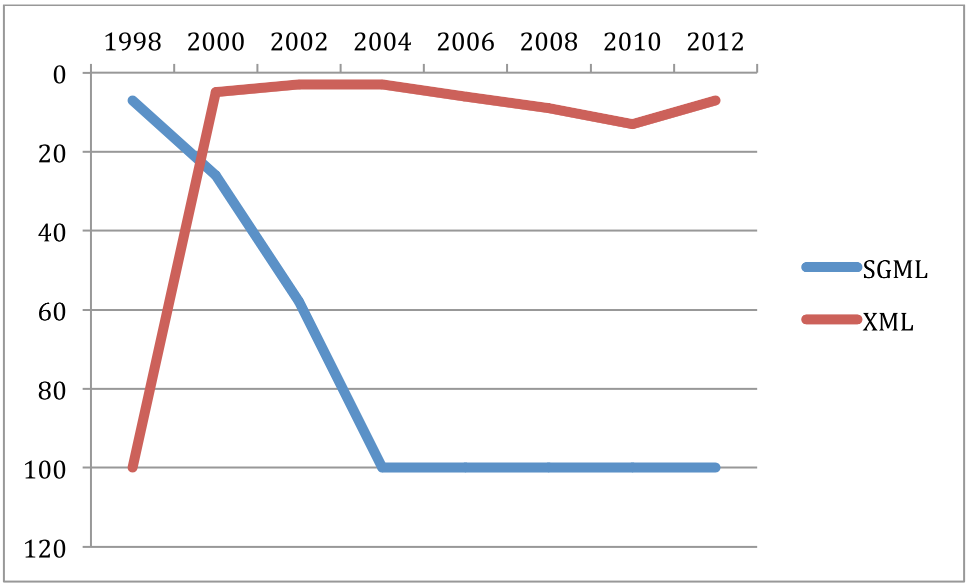

Figure 30. *Comparison of SGML and XML over time*

We also studied the arising of “*bigram*”, “*trigram*” and “*NGram*”, only “*Ngram*” remaining (Fig 31).

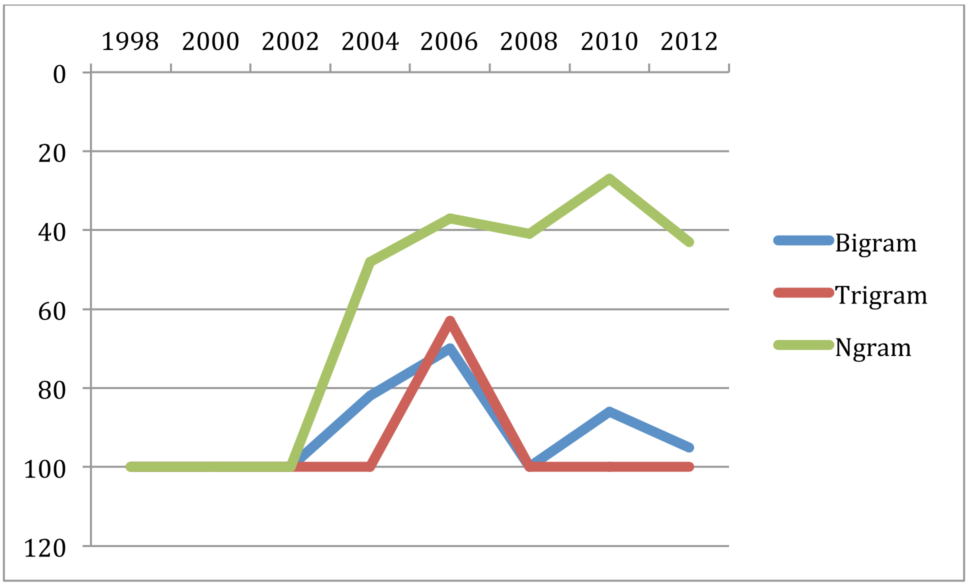

Figure 31. *Comparison of bigram, trigram and Ngram over time*

2.6.2.4. Keywords strongly fluctuating (Fig. 32):

We studied in this category terms that strongly fluctuated: those which were very popular in 1998, then lost popularity in 2002 and recently regained popularity: *LM* (9/75/34) and *Tagset* (13/82/54), one which became popular and lost popularity recently: *Framenet* (Less than 100/21/86), and one which became popular and seem to fluctuate at each conference: *BLEU* (Less than 100/30/80). *Neural Networks* was popular by the end of the 90s, lost its popularity in the 2000s and recently regained popularity (30/89/42).

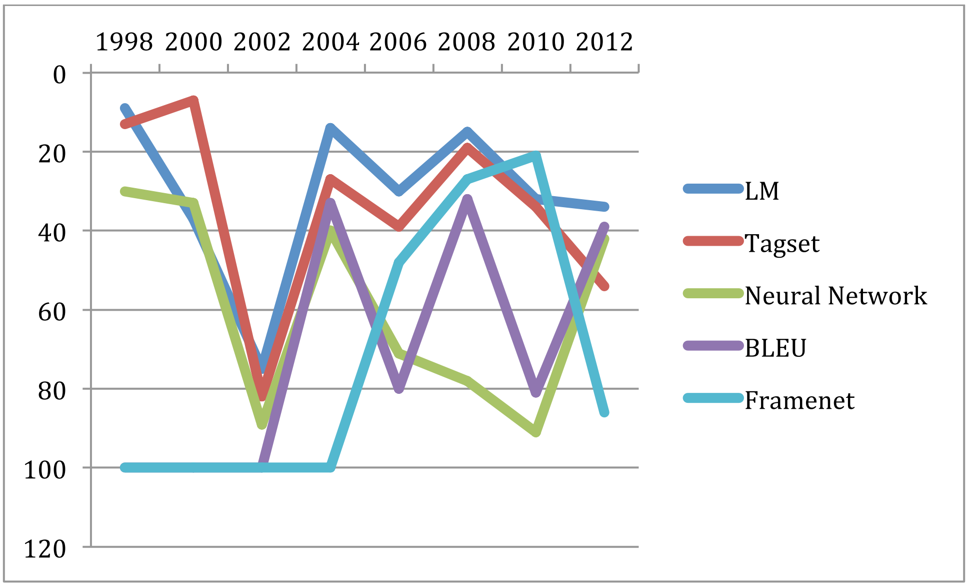

Figure 32. *Terms strongly fluctuating*

2.6.2.5. Keywords slightly fluctuating (Fig. 33):

We also have terms that slightly fluctuate over time, such as *Ontology* (19/4/10), *MT* (11/42/18), *Metric* (18/26/20) and *segmentation* (8/18/6).

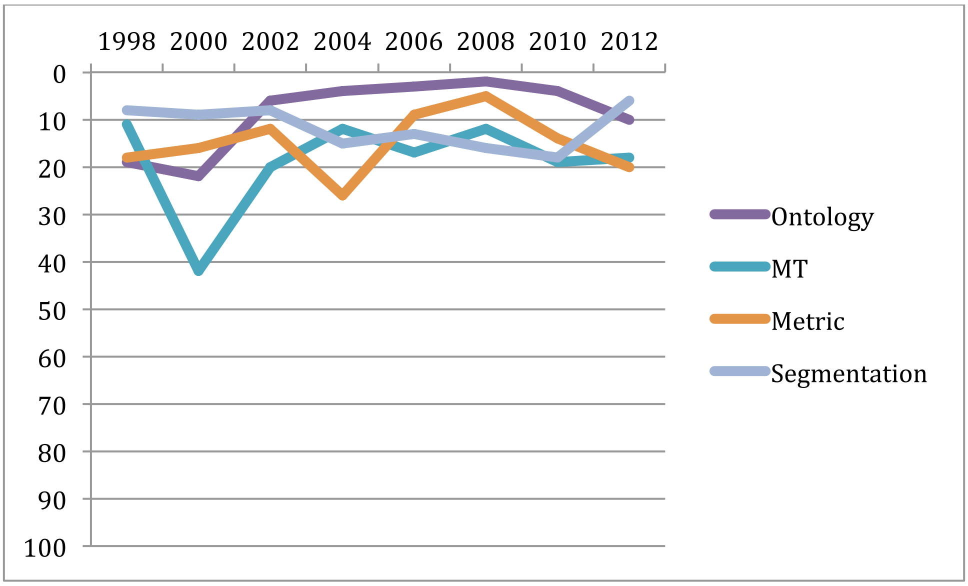

Figure 33. *Terms slightly fluctuating*

2.6.3. Specific study on the "15-year best friends" of "popular" terms

A selection of terms has been studied with respect to both their time-behavior and semantically closeness. The aim is to detect trends and related properties between the terms of the domain.

Let's recall that the previous diagrams have been computed on the whole text. This is efficient for getting a global estimation of the evolution of the various terms of the domain, but for a given paper, the topics mentioned in the text are rather heterogeneous: the paper deals for instance with the state of the art, with tracks which have been abandoned, with future directions and so on. Thus, in order to focus on semantically close terms, we cannot rely on the whole text. Instead, we decided to study the terms that appear in the abstracts and we made the hypothesis that the abstract is more targeted. Of course, this statement is certainly wrong for a small number of abstracts, but we took as hypothesis that this is right in the general case.

We implemented an algorithm that iterates on the "becoming popular" terms. Each of these terms is considered as a "focus" and the objective is to compute the "best friends" of this focus. We define the notion of "best friends" of a focus as simply the terms that appear the most frequently in the same abstract of the focus. So, a selection of terms is computed and then we return to the general ranking algorithm used in the previous sections. Said in other words, we consider the "best friends" as a filter.

2.6.3.1. The case of "Annotation" (Fig. 34)

Annotation stayed very popular over time. The importance of the *Annotator* only came apparent in a second step. Annotation of *POS* and *Treebanks* for *Parsers* and *Taggers* have been followed by the *Annotation* of *Framenets*, while *Semantics* and *Ontology* were always mentioned, and while the *XML* format became usual.

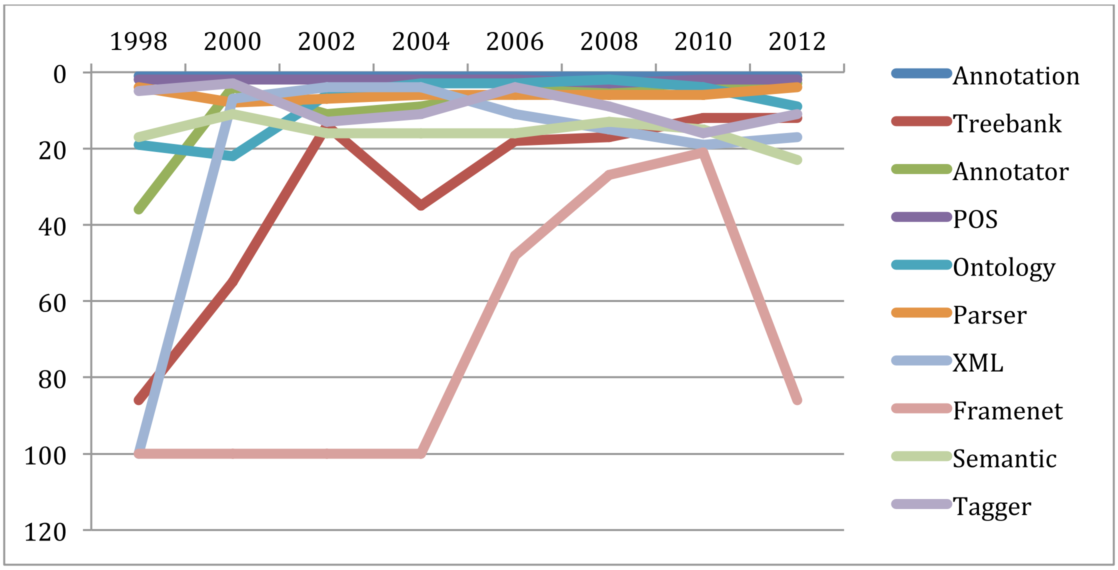

Figure 34. Rank of "Best friends" of "Annotation"

2.6.3.2. The case of "Annotator" (Fig. 35)

Annotator came after *Annotation*. After annotating *Wordnets*, *POS* and *Treebanks*, they annotated *Timebanks* and *Propbanks*. The need for *Annotation Tools* and the problem of *Annotator Agreement* came to surface, while *Ontology* was always in the foreground.

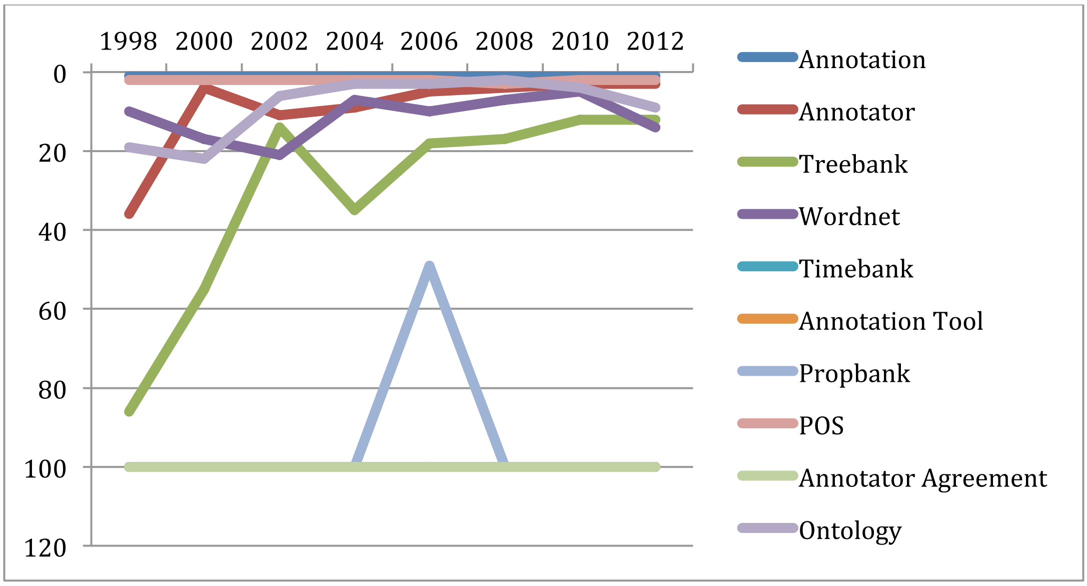

Figure 35. Rank of "Best friends" of "Annotator"

2.6.3.3. The case of "Metric" (Fig. 36)

The need for *Metrics* was clearly identified very early, for the evaluation of Machine Translation (*MT*), or of *Parsers*. It became especially popular starting in 2006 with the success of the Statistical Machine Translation (*SMT*) approach. *BLEU* is a very popular metrics for *MT*, while *NIST* appears in that area.

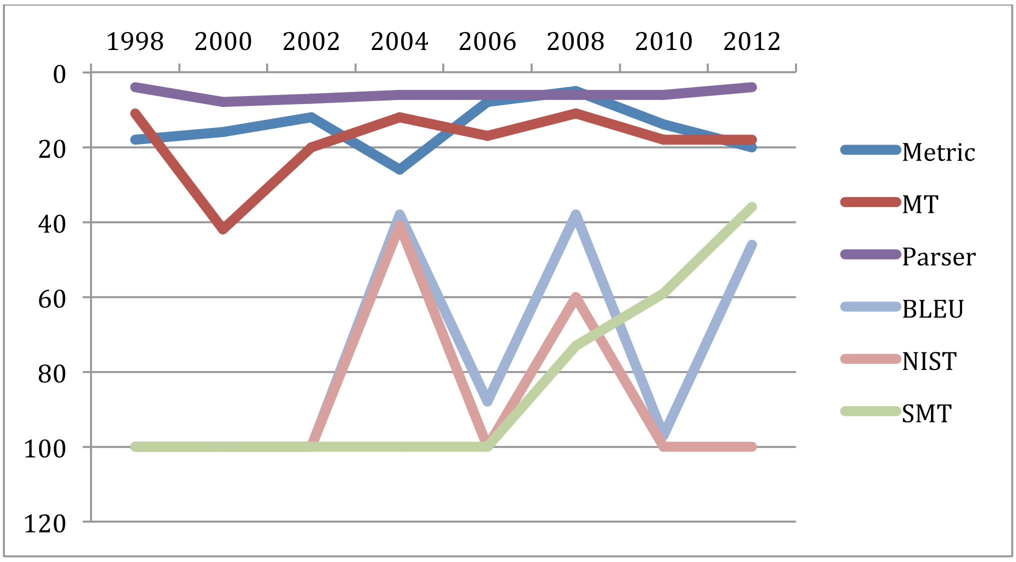

Figure 36. *Rank of "Best friends" of "Metric"*

2.6.3.4. The case of "Synset" (Fig. 37)

Synset went of course along with *Wordnet* over the years, but also with *Ontology*. *Framenet* and *Sentiwordnet* came later, while the reference to *Princeton* and *Eurowordnet* can be noticed. The use for *Disambiguation* in general, and to Word Sense Disambiguation (*WSD*) in particular can also be mentioned.

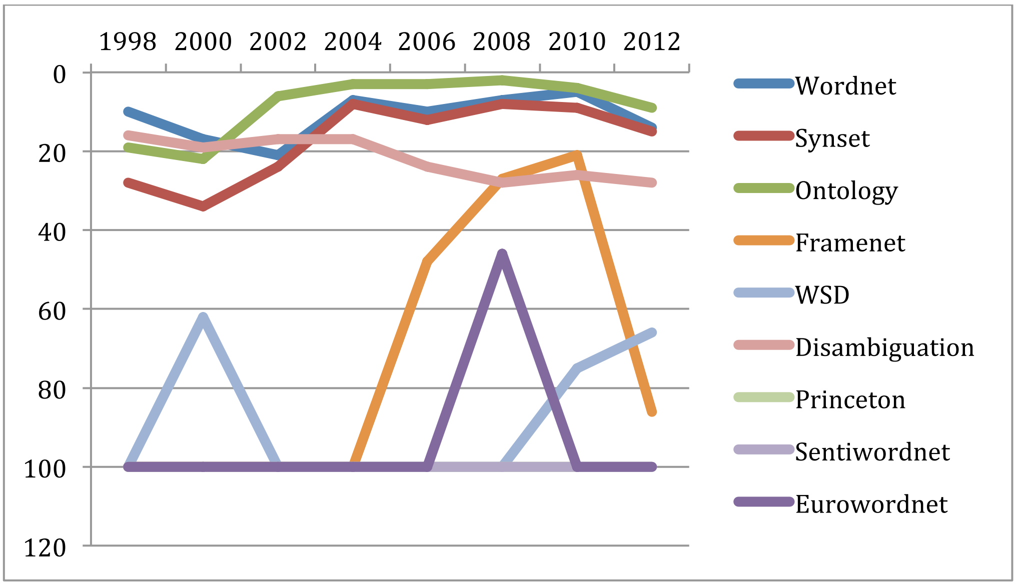

Figure 37. *Rank of "Best friends" of "Synset"*

2.6.3.5. The case of "WER" (Fig. 38)

Word Error Rate (*WER*) and more generally *Error Rate* accompanied the development of Speech Recognition (*SR*), but also of Machine Translation (*MT*) and the use of Language Models (*LM*). *NIST* appears as a major actor in organizing evaluation campaigns, as well as the US Evaluation Pilot Advisory Committee (*EPAC*). The Broadcast News (*BN*) task was especially very popular, while *LIMSI-CNRS* and *Philips* were the first non-US laboratories to participate in the Speech Recognition evaluation campaigns organized by NIST and appear as such, even if they stay in the background as their rank is lower than 100.

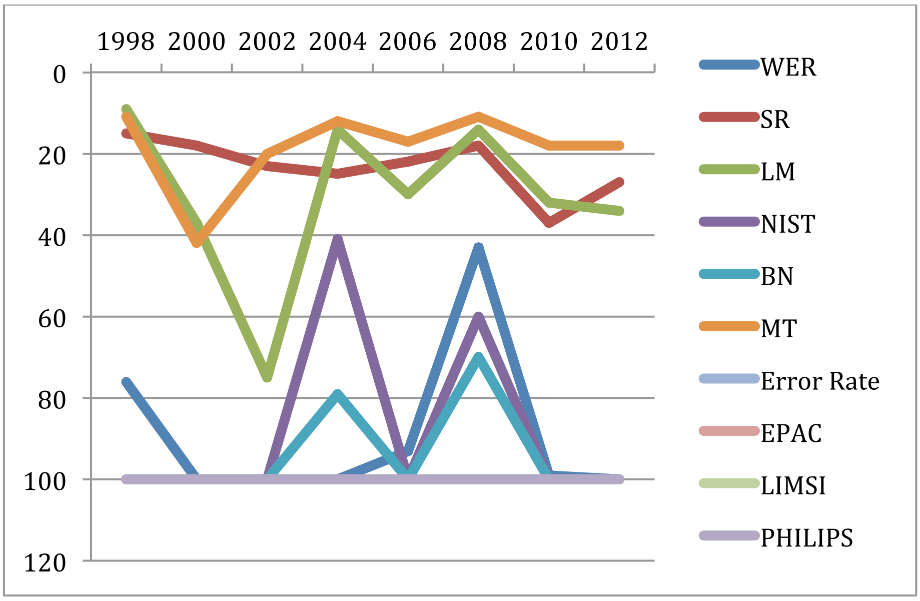

Figure 38. *Rank of "Best friends" of "WER"*

2.6.4. *Tag Clouds for frequent terms*

The aims of this current section is to have a global estimation of the main terms of a specific year and to have an idea of the stability of the terms over the years. The line-based diagrams presented in the previous section allow for a fine grain presentation but they do not permit a global view. For this purpose, we decided to experiment Tag Clouds.

From the extracted terms considered as the terms of the domain as stated in the previous sections, we run a web service called TagCrowd[27], and we thank Daniel Steinbock for providing it. This service has size limitations and it was not possible to compute the Tag Clouds from the terms coming from the body of the papers. We therefore only selected the terms taken from the abstracts.

We present here the Tag Clouds for the first and second conferences (1998 and 2000), and then every 4 years (2004, 2008 and 2012) in order to better stress the differences.

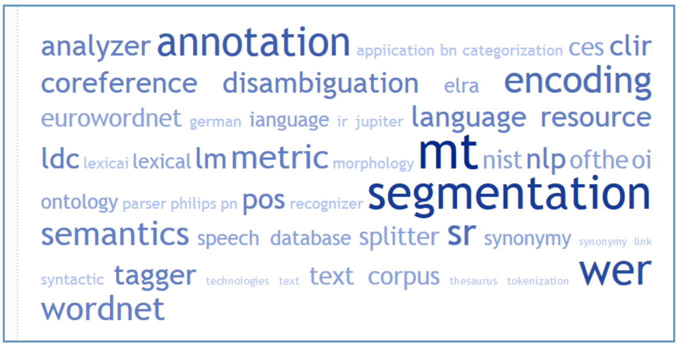
analyzer annotation application bn categorization ces clir
coreference disambiguation elra encoding
eurowordnet german ianguage ir jupiter language resource
ldc lexicai lexical lm metric morphology mt nist nlp ofthe oi
ontology parser philips pn pos recognizer segmentation
semantics speech database splitter sr synonymy synonymy link
syntactic tagger technologies text text corpus thesaurus tokenization wer
wordnet

Figure 39. *Tag Cloud based on the 1998 abstracts*

[27] www.tagcrowd.com

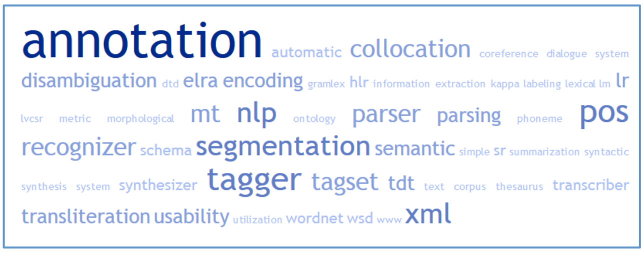

Figure 40. *Tag Cloud based on the 2000 abstracts*

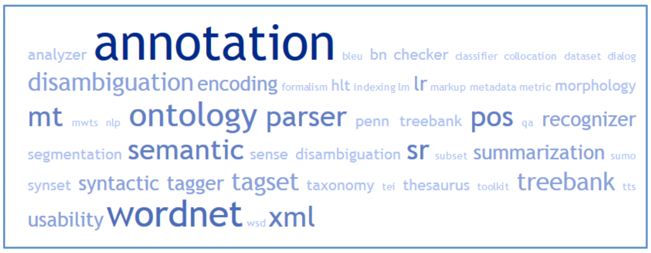

Figure 41. *Tag Cloud based on the 2004 abstract*

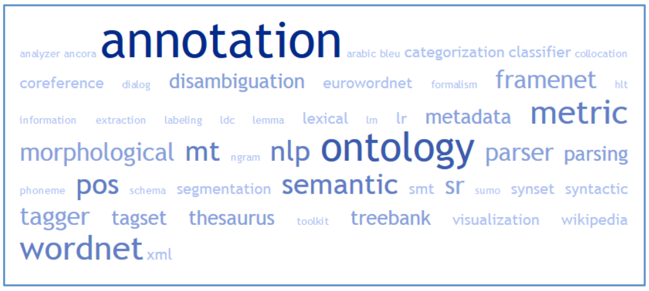

Figure 42. *Tag Cloud based on the 2008 abstracts*

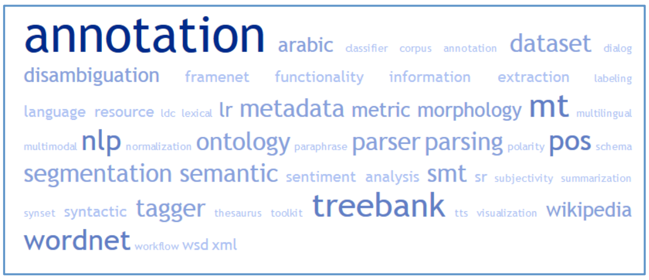

Figure 43. *Tag Cloud based on the 2012 abstracts*

Globally, it appears that most frequent terms remained across the years, and the "pictures" look similar. "*Annotation*" was already popular in 1998 and stayed popular since then. In 2000, "*Segmentation*" and "*Encoding*" got less apparent, while *POS* and *Tagger* increased their presence. In 2004, "*ontology*", "*semantic*" and "*wordnet*" came to te forefront. The presence of "*ontology*" was reinforced in 2008, while "*metric*" strongly appeared. "*Treebank*" and "*MT*" were very present in 2012.

2.6.5. New terms introduced by authors

We studied when and whom introduced new terms, as a mark of the innovative ability of various authors, which may also be taken into account in the estimate of their influence in the community. We make the hypothesis that an innovation is induced by the introduction of a term which was unused in the community and which is now very popular. We proceeded as follows: first we applied a terminological extraction in order to determine which are the 20 most used terms in the last proceedings (LREC 2012) and which were not used in the first proceedings. Each term thus appeared at a certain point in time. For each of these terms, starting from the second proceedings, we detected the author who introduced the term. This computation may give one or several names, as the papers could be authored by several researchers, or could be mentioned in several papers on the same year. The results when ranked in chronological order are the following (Table 11).

Term	Conference	Authors
SVM	lrec2000	Alex Waibel, Alon Lavie, Klaus Ries, Liza Valle, Lori Levin, Tatsuo Yamashita, Yuji Matsumoto
connective	lrec2000	Jordi Porta Zamorano, Montserrat Marimón Felipe
BLEU	lrec2002	Christopher Cieri, Keith J Miller, Kishore Papineni, Mark Liberman, Michelle Vanni
Google	lrec2002	Atsushi Fujii, Jimmy Lin, Katunobu Itou, Kristina Nilsson, Lars Borin, Tetsuya Ishikawa
MSA	lrec2004	Albino Nogueiras Rodríguez, Anastasios Tsopanoglou, Asunción Moreno, Dorota Iskra, Herbert S Tropf, Imed Zitouni, Irene Castellón, Jamal Borno, Jordi Escribano, Khalid Choukri, Laura Alonso i Alemany, Lluís Padró, Nikos Fakotakis, Oren Gedge, Ossama Emam, Rainer Siemund, Sanda M Harabagiu, Steven J Maiorano, V Finley Lacatusu, Xavier Messeguer
SMS	lrec2004	A Chalamandaris, András Kornai, András Rung, G Giannopoulos, George Carayannis, István Szakadát, László Németh, P Tsiakoulis, Péter Halácsy, Spyros Raptis, Viktor Trón
Wikipedia	lrec2004	Christian Biemann, Christian Wolff, Stefan Bordag, Uwe Quasthoff
UIMA	lrec2008	Alexander Troussov, Anthony Levas, Antonio Pareja-Lora, Branimir Boguraev, Brian Davis, Christian Biemann, Christian Chiarcos, Claire Waast-Richard, Claudia Soria, David Ferrucci, Ekaterina Buyko, Florian Holz, Frederik Cailliau, Gerhard Heyer, Gilles Adda, Guergana Savova, James Masanz, Jean-Luc Gauvain, John Carroll, John Judge, Julien Nioche, Karin Schuler, Lori Lamel, Martine Garnier-Rizet, Mary Neff, Mikhail Sogrin, Monica Monachini, Nicoletta Calzolari, Paul Keyser, Philip Ogren, Riccardo Del Gratta, Roberto Bartolini, Siegfried Handschuh, Stephan Vanni, Sylvie Guillemin-Lanne, Ted Briscoe, Tommaso Caselli, Uwe Quasthoff, Valeria Quochi, Vinod Kaggal, Youssef Drissi, Øistein E Andersen
Wiktionary	lrec2008	Christof Müller, Corina Forăscu, Dan Cristea, Iryna Gurevych, Jonas Sjöbergh, Kenji Araki, Marius Răschip, Michael Zock, Torsten Zesch
tweet	lrec2010	Franciska de Jong, Henk van den Heuvel, Martin Reynaert, Nelleke Oostdijk, Orphée De Clercq

Table 11. Introduction of new terms: date and authors.

2.6.6. Specific study about clustering on 2012 papers using the "Term Frequency Inverse Document Frequency" (TF-IDF)

The objectives were twofold. First, we wanted to study whether it is possible to facilitate the automatic clustering of papers into a limited number of sessions based solely on the parsing of the content. Secondly, we wanted to exhibit possible semantic hidden links between apparently unrelated papers.

Our process relies on the same terminological extraction as the previous sections. Let's recall that this extraction computes the terms of the domain from the difference between a statistical profile of "ordinary" English templates (recorded on a disk) and the syntactic patterns of the papers of the conference. Once the terms are collected, the TF-IDF of each term is computed. Without entering into mathematical details, let's say that the TF-IDF value reflects how important a term is to represent a document within a corpus[28] (C. Manning et al., 2008). A consequence of this computation is that the popular terms over the whole conference (like "Annotation", for instance) do not have a high TF-IDF value: only specific terms have a high value.

We define the notion of "salient terms" of a paper as being the terms with the highest TF-IDF and we consider only the five highest values (see "docMostSalientTerms" in Table 12). Said in other words, the salient terms of a given paper are the terms that distinguish this paper from the rest of the conference. It should be noted that this statement is valid within the paradigm of the "Bags of Words", that means that we do not make any distinction between, for instance the two terms strategy#1 and strategy#2 in the sentence "We apply strategy#1 and not strategy#2 which was used 10 years ago". In our process, strategy#1 and strategy#2 count equally for one because we count only the number of occurrences and do not consider the negation.

Then, we considered these salient terms as the representation of the paper and from these terms, we automatically clustered the papers using a hierarchical clustering algorithm (UPGMA) using the cosine similarity between papers. Once each cluster is built, the terms of the clusters are ranked according to their TF-

[28] See http://en.wikipedia.org/wiki/Tf-idf for details

IDF in order to get a list of terms that are representative of the cluster (see "clusterMostSalientTerms" in Table 12). We finally cleaned the selected papers by eliminating the cases where an acronym was misinterpreted. The clustering process gives the following result.

Cluster#1
number of documents=3
clusterMostSalientTerms=TBAQ,Timex normalization,evaluation component,post-correction,timexes

Paper	Title	docMostSalientTerms
lrec2012_657	A corpus of general and specific sentences from news	AQ,AQUAINT,Berner,NYT science,docid
lrec2012_451	Massively Increasing TIMEX3 Resources: A Transduction Approach	AQUAINT,TBAQ,TimeML corpus,Timex,signal phrase
lrec2012_128	TIMEN: An Open Temporal Expression Normalisation Resource	Timex,Timex normalization,approach to Timex,evaluation component,timexes

Cluster#2
number of documents=3
clusterMostSalientTerms=Visuel,WordVis,actant,record of DiCoInfo,structure of term

Paper	Title	docMostSalientTerms
lrec2012_366	Capturing syntactico-semantic regularities among terms: An application of the FrameNet methodology to terminology	DiCoInfo,DiCoInfo lexicographer,actant,actantial,structure of term
lrec2012_1096	Logic Based Methods for Terminological Assessment	DiCoInfo,Visuel,WordVis,record of DiCoInfo,visualization device
lrec2012_245	Identifying equivalents of specialized verbs in a bilingual comparable corpus of judgments: A frame-based methodology	FS,Judge,actantial,arguido,pair of equivalent

Cluster#3
number of documents=3
clusterMostSalientTerms=MLE,SCF,SSI Dijkstra,monotransitive,set of SCFs

Paper	Title	docMostSalientTerms
lrec2012_390	Customizable SCF Acquisition in Italian	ACCUSARE,MLE,PVF,SCF,SCFs
lrec2012_1063	Reclassifying subcategorization frames for experimental analysis and stimulus generation	SCFs,Valent,monotransitive,plus clause,set of SCFs
lrec2012_269	Using Verb Subcategorization for Word Sense Disambiguation	Dijkstra,Dijkstra algorithm,SCF,SCF model,SSI Dijkstra

Cluster#4
number of documents=3
clusterMostSalientTerms=Kathir,Quran dictionary,hyperlinked, [illegible]

Paper	Title	docMostSalientTerms
lrec2012_646	LAMP: A Multimodal Web Platform for Collaborative Linguistic Analysis	Dukes,Quran,Quran dictionary,hyperlinked,mafūl
lrec2012_190	QurSim: A corpus for evaluation of relatedness in short texts	Kathir,Quran, [illegible]
lrec2012_123	QurAna: Corpus of the Quran annotated with Pronominal Anaphora	Quran, [illegible] , [illegible]

Table 12. *Clustering of LREC 2012 papers using "Term Frequency Inverse Document Frequency" (TF-IDF)*

Due to the fact that the computed TF-IDF weights highlight terms that are specific to a paper, the clusters are built in such a way that some papers that are apparently unrelated are gathered together.

In the first cluster, the three papers appear in different sessions at LREC 2012. The two first ones get linked through the use of the AQUAINT corpus, while the two last ones refer to the study of Temporal Expressions (TIMEX).

In the second cluster, the three papers also belong to different sessions. The two first ones share the use of the DicoInfo resource, while the two last ones refer to the study of the "actantial" structure of the verbs, under the writing of the same author working at different sites.

The three papers in the third cluster also belong to different sessions, and gather under the umbrella of SubCategorization Frame (SCF) approach.

The fourth cluster illustrates the strong presence of the study of the Quran at LREC 2012. Although the three papers come from the same laboratory, they were placed in three different sessions.

2.7. Text reuse and plagiarism

We studied the use across the conference series of parts of former papers written by the same authors (that we will call "reuse") or by different authors (that we will call "plagiarism"). For this, we considered a bag of windows of a certain number of words, after a linguistic processing in each paper published at a conference, and compared them with the windows obtained from papers published at all former conferences. More precisely, we apply the deep parser Tagparser with robust morphological analysis, word tagging and named entity recognition and then stored the result in a huge index which is dynamically used for comparison. Each windows is made of the sequence (the order is important) of lemmas and the parts of speech. After some trials, we found that a size of 7 words gave meaningful results, so we set 7 as an empirical parameter.

We then gave a closer look at the couple of papers which have more than 3% similarity in the case of possible plagiarism, and more than 10% similarity in the case of reuse.

It appears that only one case of possible plagiarism was detected, but it appeared after a manual checking that the two papers came from the same laboratory, even if the authors were different. 29 possible reuses of already published papers were detected, ranging from a similarity ratio of 10% up to 47%. Most of the time, the similar parts are related to the presentation of a program, a project, a problem or a resource shared by the two papers. In case of reuse, the time span is usually related to the former conference, as it appears in Table 13.

Reused Reusing	1998	2000	2002	2004	2006	2008	2010	Total
2000	1							1
2002		2						2
2004		1	1					2
2006			1	7				8
2008			1		5			6
2010						3		3
2012						1	6	7
Total	1	3	3	7	5	4	6	29

Table 13. Number of papers reusing and Number of papers reused at each conference

3. Perspectives

Conducting this analysis has been a heavy work shared by the 4 authors. It is still preliminary, as other aspects would deserve attention.

We plan to investigate more deeply the structure of the research community through the graph of collaboration and the graph of citations among authors, as a social network. This process will help identifying factions of people who publish together or cite each other.

We still need to analyze the cited papers, when we will be able to identify those citations with enough reliability, and to establish the link between the citing authors, the cited authors, the citing papers and the cited papers. We will then conduct an opinion survey, such as the change over time of citation purposes, or of citation polarity (positive, neutral, negative).

We will extend the bottom up term analysis that we already started, and deepen the potential detection of weak signals and emerging trends. In parallel, we will also consider in a top down manner the evolution of the index terms provided by the authors themselves in their papers. We will analyze the evolution of the conference sessions' title and content over time.

Establishing a link between the authors, the citations and the papers' topics will allow us to study the changes in the topics of interest for authors or factions.

Finally, we will keep on carrying on the comparative study of the various conferences, and consider the community they form all together.

4. Conclusions

In this analysis exercise, we benefited of the fact that all the LREC conferences data is freely available online. However, we faced some difficulty in the use of the available data. The eldest information for LREC 1998 could not be used directly, because it was not available in a text format, and we had to use OCR, which inserted some errors.

We spent an important time cleaning the data related to authors' name, laboratory affiliations, countries, funding agencies, with all their variants, that can only be sorted by a human eye. There is a clear need for a better identification of all those entities, which will necessitate an international effort, as the identifiers must be unique. It is a challenge for the scientific community, through their associations, in order to avoid that the charges and privileges attached to this organizational activity be seized by for-profit companies.

The research in Natural Language Processing, for both spoken, written and signed languages, has achieved major advances over this period through constant and steadily scientific efforts, that gained efficiency thanks to the availability of a necessary infrastructure made up of publicly funded programs, largely available language resources, and regularly organized evaluation campaigns initiated in the USA by the 80s. It also very importantly benefited of a scientific social network bridging the community through the LREC conference organized by the European Language Resource Association (ELRA) to share ideas and make progress.

This preliminary analysis allowed us to extract salient or hidden information and trends which, we hope, provide a better understanding of the past 15 years of research in Language Resources and Evaluation worldwide. We hope it will also serve as a precious experience for building up the next 15 years.

5. Acknowledgements

The authors wish to thank the ACL colleagues, Ken Church, Sanjeev Khudanpur, Amjbad Abu Jbara, Dragomir Radev and Simone Teufel, who helped them in the starting phase, Isabel Trancoso, who gave her ISCA Archive analysis on the use of assessment and corpora, Wolfgang Hess, who produced and provided the 14 GBytes ISCA Archive and Florian Boudin, who made available the TALN Anthology. They also thank Khalid Choukri and Nicoletta Calzolari, who provided information about the past LREC conferences, Victoria Arranz, Ioanna Giannopoulou, Johann Gorlier, Jérémy Leixa, Valérie Mapelli and Hélène Mazo, who helped in recovering the metadata for LREC 1998, and all the organizers, reviewers and authors over the 15 years conferences without whom this analysis could not be conducted!

6. Apologies

This survey has been conducted on textual data, which covers a 15 years long period, including an OCRized content for 1998. The analysis uses tools, which automatically process the content of the scientific papers and may make errors. Therefore, the results should be considered as containing an error margin, and the authors wish to apologize for any errors that the reader may detect and that they will be glad to take into account in future releases of the present survey.

7. References

ACL (2012), Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries, ACL 2012, Jeju, July 10 2012, ISBN 978-1-937284-29-9

Auber, David; Archambault, Daniel; Bourqui, Romain; Lambert, Antoine; Mathiaut, Morgan; Mary, Patrick; Delest, Maylis; Dubois, Jonathan and Mélançon, Guy (2012), Research Report, p31, January 2012, http://hal.archives-ouvertes.fr/hal-00659880

Bavelas, Alex (1948) "A mathematical model for small group structures." *Human Organization 7: 16-30.*

Bavelas, Alex (1950) "Communication patterns in task oriented groups." *Journal of the Acoustical Society of America 22: 271-282.*

Boudin, Florian (2013) TALN archives: une archive numérique francophone des articles de recherche en traitement automatique de la langue. TALN-RÉCITAL 2013, 17-21 Juin 2013, Les Sables d'Olonne

Councill, Isaac G.; Giles, C. Lee and Kan, Min-Yen (2008), ParsCit: An open-source CRF reference string parsing package. In Proceedings of the Language Resources and Evaluation Conference (LREC 2008), Marrakesh, Morocco, May 2008.

Csárdi, Gábor and Nepusz, Tamás (2006). The igraph software package for complex network research. InterJournal 2006. *Complex Systems* 1695: 1-9.

Drouin, Patrick (2004) Detection of Domain Specific Terminology Using Corpora Comparison. In Proceedings of the Language Resources and Evaluation Conference (LREC 2004), Lisbon, Portugal, May 2004.

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

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

Freeman, Linton C. (1978) Centrality in Social Networks, Conceptual Clarifications. Social Networks. 1 (1978/79) 215-239

Fruchterman, T. M. J. and Reingold, E. M. (1991). Graph Drawing by Force-Directed Placement. Software: Practice and Experience, 21(11).

Fujisaki, Hiroya (2013), History of ICSP and PC-ICSLP, ISCA Web site – About ISCA – History http://www.isca-speech.org/iscaweb/index.php/about-isca/history

Ide, Nancy; Suderman, Keith and Simms, Brian (2010) ANC2Go: A Web Application for Customized Corpus Creation, Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), May 2010, Valletta, Malta, European Language Resources Association (ELRA), 2-9517408-6-7.

Joerg, Brigitte; Höllrigl, Thorsten and Sicilia, Miguel-Angel (2012) Entities and Identities in Research Information Systems, 2012. In 11th International Conference on Current Research Information Systems (CRIS2012): "e-Infrastructures for Research and Innovation: Linking Information Systems to Improve Scientific Knowledge Production", Prague, Czech Republic, June 6-9, 2012.

Koehn, Philipp (2005), Europarl: A Parallel Corpus for Statistical Machine Translation,, MT Summit 2005.

Litchfield, Ben (2005), Making PDFs Portable: Integrating PDF and Java Technology, March 24, 2005, Java Developers Journal, http://java.sys-con.com/node/48543 (pdfbox is available at http://pdfbox.apache.org/)

Manning, Christopher D.; Raghavan, Prabhakar and Schütze, Hinrich (2008), *Introduction to Information Retrieval*, Cambridge University Press. 2008., ISBN: 0521865719

Mariani, Joseph (1990), La Conférence IEEE-ICASSP de 1976 à 1990 : 15 ans de recherches en Traitement Automatique de la Parole, Notes et Documents LIMSI 90-8, Septembre 1990

Mariani, Joseph (2013) The ESCA Enterprise, ISCA Web site – About ISCA – History http://www.isca-speech.org/iscaweb/index.php/about-isca/history

Mariani, Joseph; Paroubek, Patrick; Francopoulo, Gil and 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

Porter, M. F. (2012), An algorithm for suffix stripping, 1980, Program 14 (3), 130-137 (awk implementation by Gregory Grefenstette, July 5 2012, is available at http://tartarus.org/martin/PorterStemmer/awk.txt)

The British National Corpus (2007), version 3 (BNC XML Edition). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/

The R Journal (2012), 4(2):5-12, December 2012, ISSN 2073-4859, http://journal.r-project.org/

Fu, Yu; Xu, Feiyu and Uszkoreit, Hans (2010), Determining the Origin and Structure of Person Names, Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), May 2010, pp 3417-3422, Valletta, Malta, European Language Resources Association (ELRA), isbn: 2-9517408-6-7.