REDISCOVERING 50 YEARS OF DISCOVERIES IN SPEECH AND LANGUAGE PROCESSING: A SURVEY.

Joseph Mariani[1], *Gil Francopoulo*[2], *Patrick Paroubek*[1], *Frédéric Vernier*[1]
[1]LIMSI-CNRS, [2]Tagmatica
Joseph.Mariani@limsi.fr, gil.francopoulo@wanadoo.fr, pap@limsi.fr, frederic.vernier@limsi.fr

ABSTRACT

We have created the NLP4NLP corpus to study the content of scientific publications in the field of speech and natural language processing. It contains articles published in 34 major conferences and journals in that field over a period of 50 years (1965-2015), comprising 65,000 documents, gathering 50,000 authors, including 325,000 references and representing approximately 270 million words. Most of these publications are in English, some are in French, German or Russian. Some are open access, others have been provided by the publishers. In order to constitute and analyze this corpus several tools have been used or developed. Some of them use Natural Language Processing methods that have been published in the corpus, hence its name. Numerous manual corrections were necessary, which demonstrated the importance of establishing standards for uniquely identifying authors, publications or resources. We have conducted various studies: evolution over time of the number of articles and authors, collaborations between authors, citations between papers and authors, evolution of research themes and identification of the authors who introduced them, measure of innovation and detection of epistemological ruptures, use of language resources, reuse of articles and plagiarism in the context of a global or comparative analysis between sources.

Index terms: Speech Processing, Natural Language Processing, Text Analytics, Bibliometrics, Scientometrics, Informetrics.

1. INTRODUCTION

1.1. Objective

Our goal is to use Natural Language Processing (NLP) tools to study the bibliography in Natural Language Processing. We already conducted such investigations in 1991, with a study of the IEEE ICASSP conference series over a period of 15 years between 1976 and 1990 [20]. This study helped initializing the launching of the Eurospeech conference, now Interspeech [21]. The Association for Computational Linguistics (ACL) has produced an Anthology [1] [31] and organized a workshop entitled "*Rediscovering 50 Years of Discoveries in Natural Language Processing*" on the occasion of ACL's 50th anniversary in 2012 at Jeju (Korea) [1]. We have been invited to give a keynote talk entitled "*Rediscovering 25 Years of Discoveries in Spoken Language Processing"* on the occasion of the 25th anniversary of the International Speech Communication Association (ISCA) during the Interspeech'2013 conference in Lyon (France), based on the ISCA Archive [2] assembled by Wolfgang Hess [22]. Then another analysis of 15 years of research contained in the Language Resources and Evaluation Conference (LREC) proceedings between 1998 and 2012 at the LREC 2014 conference in Reykjavik (Iceland) [23], which was followed by an article "*Rediscovering 15+2 Years of Discoveries in Language Resources and Evaluation*" published in the *Language Resources and Evaluation Journal* in March 2016 [26]. And finally an invited talk "*Rediscovering 10 to 20 Years of Discoveries in Language and Technology"* for the 20th anniversary of the L&TC conference in Poznan (Poland) in 2015 [25]. Our objective was then to integrate and extend those studies to half a century of research investigations in Language Processing. The present article provides a survey of the main results, without getting into too many details on the data and methods used.

1.2. A hot topic

The application of text analytics to bodies of scientific papers has become an active area of research in recent years (see for example [19], [33], [7], [28], [5], [14], [17]), the *Stanford Large Network Dataset Collection* (SNAP) [3] or the Saffron[4] project. On our side, we participated in the *Workshop on Mining Scientific Publications* (WOSP'2015) at Fort Knox (USA), on June 24-25 2015, which resulted in a special issue of the D-Lib Magazine **(**Nov./Dec. 2015, Vol. 21, N° 11/12) [12], and, at about the same time, in the *Workshop on Computational Linguistics and Bibliometrics* (CLBib), organized within the 15th *Int*[al] *Society of Scientometrics and Informetrics Conference* (ISSI)

[1] http://aclweb.org/anthology/
[2] http://www.isca-speech.org/iscaweb/index.php/archive/online-archive
[3] http://snap.stanford.edu/data/
[4] http://saffron.deri.ie

in Istanbul (Turkey), on June 29, 2015 [11]. More recently, we participated in *BIRNDL: Joint Workshop on Bibliometric-enhanced IR (BIR) and NLP for digital libraries (NLPIR4DL)*, organized in the framework of the ACM/IEEE *Joint Conference on Digital Libraries'2016* in Newark (USA) on June 23, 2016, which resulted in a special issue of the *International Journal on Digital Libraries* in March 2017 [27].

2. THE NLP4NLP CORPUS

We apply NLP methods to analyze NLP bibliography, hence the name we gave to the corpus: NLP4NLP corpus, ([10], [11]). We consider here Language Processing in the broad sense, which includes written, spoken and sign language processing and Information Retrieval. The NLP4NLP corpus contains papers from thirty-four publications over 50 years (1965-2015), including major conferences (ACL, IEEE-ICASSP (only the speech part), ISCA-Interspeech, ELRA-LREC, etc.) and journals (IEEE-TASLP, *Computational Linguistics*, *Speech Communication*, *Computer Speech and Language*, *Language Resources and Evaluation*, etc.). This represents 558 *events* (by "events" we mean either a conference venue, either annual or with a variable frequency, or a journal number, which often corresponds to a calendar year). This regroups 65,003 articles written by 48,894 different authors representing about 270 MWords and containing 324,422 bibliographical references.

Table 1 provides the list of the elements in the corpus, with the name and acronym of the publication, conference or journal, the number of documents it contains, the language (mostly English, but some are in French and a few in German or Russian), the access modality, either open or proprietary (in this case we received the agreement of the publisher for using the data in the present study), the period and the number of events. In order to get the total number of papers and events contained in the corpus, it is necessary to cancel the duplicate data corresponding to some joint conferences.

short name	# docs	format	long name	language	access to content	period	# venues
acl	4264	conference	Association for Computational Linguistics Conference	English	open access *	1979-2015	37
acmtslp	82	journal	ACM Transaction on Speech and Language Processing	English	private access	2004-2013	10
alta	262	conference	Australasian Language Technology Association	English	open access *	2003-2014	12
anlp	278	conference	Applied Natural Language Processing	English	open access *	1983-2000	6
cath	932	journal	Computers and the Humanities	English	private access	1966-2004	39
cl	776	journal	American Journal of Computational Linguistics	English	open access *	1980-2014	35
coling	3813	conference	Conference on Computational Linguistics	English	open access *	1965-2014	21
conll	842	conference	Computational Natural Language Learning	English	open access *	1997-2015	18
csal	762	journal	Computer Speech and Language	English	private access	1986-2015	29
eacl	900	conference	European Chapter of the ACL	English	open access *	1983-2014	14
emnlp	2020	conference	Empirical methods in natural language processing	English	open access *	1996-2015	20
hlt	2219	conference	Human Language Technology	English	open access *	1986-2015	19
icassps	9819	conference	IEEE International Conference on Acoustics, Speech and Signal Processing - Speech Track	English	private access	1990-2015	26
ijcnlp	1188	conference	International Joint Conference on NLP	English	open access *	2005-2015	6
inlg	227	conference	International Conference on Natural Language Generation	English	open access *	1996-2014	7
isca	18369	conference	International Speech Communication Association	English	open access	1987-2015	28
jep	507	conference	Journées d'Etudes sur la Parole	French	open access *	2002-2014	5
lre	308	journal	Language Resources and Evaluation	English	private access	2005-2015	11
lrec	4552	conference	Language Resources and Evaluation Conference	English	open access *	1998-2014	9
ltc	656	conference	Language and Technology Conference	English	private access	1995-2015	7
modulad	232	journal	Le Monde des Utilisateurs de L'Analyse des Données	French	open access	1988-2010	23
mts	796	conference	Machine Translation Summit	English	open access	1987-2015	15
muc	149	conference	Message Understanding Conference	English	open access *	1991-1998	5
naacl	1186	conference	North American Chapter of ACL	English	open access *	2000-2015	11
paclic	1040	conference	Pacific Asia Conference on Language, Information and Computation	English	open access *	1995-2014	19
ranlp	363	conference	Recent Advances in Natural Language Processing	English	open access *	2009-2013	3
sem	950	conference	Lexical and Computational Semantics / Semantic Evaluation	English	open access *	2001-2015	8
speechc	593	journal	Speech Communication	English	private access	1982-2015	34
tacl	92	journal	Transactions of the Association for Computational Linguistics	English	open access *	2013-2015	3
tal	177	journal	Revue Traitement Automatique du Langage	French	open access	2006-2015	10
taln	1019	conference	Traitement Automatique du Langage Naturel	French	open access *	1997-2015	19
taslp	6612	journal	IEEE/ACM Transactions on Audio, Speech and Language Processing	English	private access	1975-2015	41
tipster	105	conference	Tipster DARPA text program	English	open access *	1993-1998	3
trec	1847	conference	Text Retrieval Conference	English	open access	1992-2015	24
Total incl. duplicates	67937					1965-2015	577
Total excl. duplicates	65,003					1965-2015	558

Table 1. *The NLP4NLP Corpus of Conferences (24) and Journals (10)*[5]
(: included in the ACL Anthology)*

[5] Joint conferences and the corresponding papers are counted once in the total number of venues and documents.

3. DATA PROCESSING

The papers present in the corpus have been obtained either after scanning or directly as textual documents. In the first case, it has been necessary to transform them by using an OCR software. In some cases, the papers also include metadata. In other cases, metadata had to be extracted from the text. The information obtained through automatic extraction is related to various aspects: name of the different authors, with their affiliation, their nationality or their gender, scientific terms, language resources, citations (authors, title, sources), funding agencies, etc. Several processing used the TagParser deep syntactic parser [8], based on a large multilingual lexicon and on the *Global Atlas* knowledge base built on the content of 18 Wikipedias [9].

4. PRODUCTION ANALYSIS

4.1. Evolution of the number of publications

As we can see on fig. 1, the number of publications increased over the years but tends to stabilize.

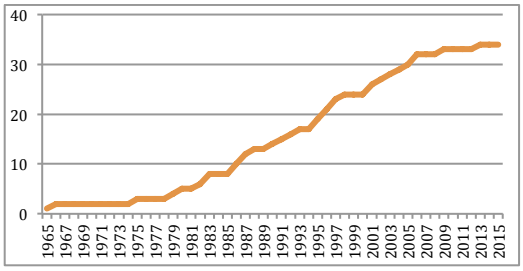

Figure 1. *Cumulated number of sources (conferences and journals) over the years.*

4.2. Evolution of the number of articles

The number of papers constantly increases in a quasi-exponential way, and reaches more than 65,000 documents in 2015 (Fig. 2)

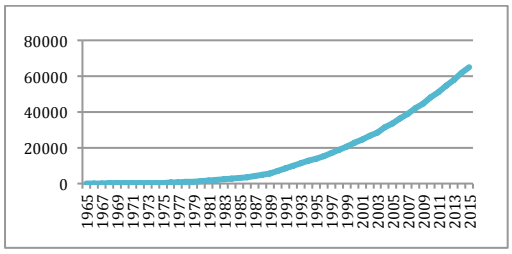

Fig. 2. *Cumulated number of papers over the years.*

The number of documents provided by each publication is also very variable, from 18,369 documents from the ISCA conference series down to 82 in the case of the ACM Transactions on Speech and Language Processing (ACM-TLSP) (Fig. 3).

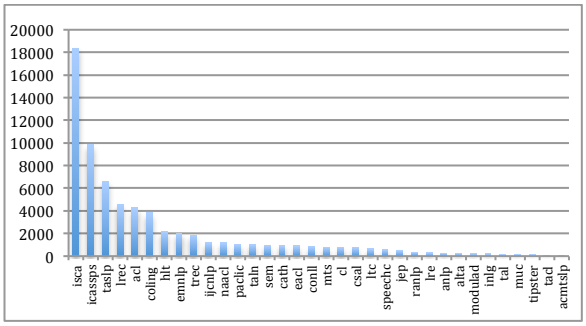

Figure 3. *Number of documents for each source*

This is linked to the publication age, to its frequency and to the number of papers that are published for each event that is very variable (Fig. 4). The ISCA-Interspeech conferences are those which publish the largest number of articles at each event (656 on average), followed by LREC (506), ICASSP-Speech (378) IJCNLP (198) and Coling (182). *ACM-TLSP* only has 8 articles on average for each number.

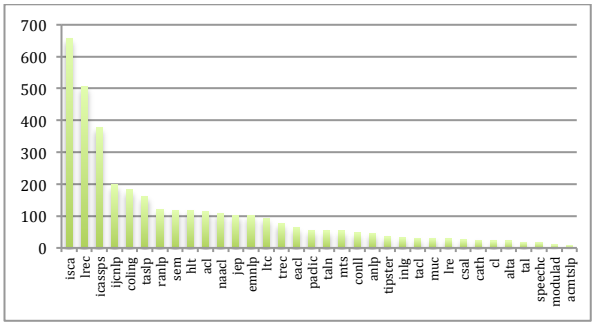

Figure 4. *Average number of documents at each venue (conferences) or issue (journals)*

4.3. Authors analysis

The study of authors is problematic due to variations in rendering of names (family name and given name, initials, middle initials, ordering, married name, etc.). It therefore required a tedious semi-automatic cleaning process [24], which resulted in a list of 48,894 different authors. This suggests a need to determine ways to uniquely identify researchers, which has been proposed [18], and may also be solved through organisms such as ORCID[6].

4.3.1. Evolution of the average number of co-authors per paper

The average number of co-authors per paper increased over time, from 1.33 in 1965 up to 3.45 in 2015 (i.e. 2 more authors on average) (Fig. 5). It is interesting to note that the number of papers with a single author was 75% in 1965 and decreased to 5%

[6] https://orcid.org

in 2015. This clearly demonstrates the change in the way research is being conducted, going progressively from individual research investigations to large projects conducted within teams or in collaboration within consortia, often in international projects and programs.

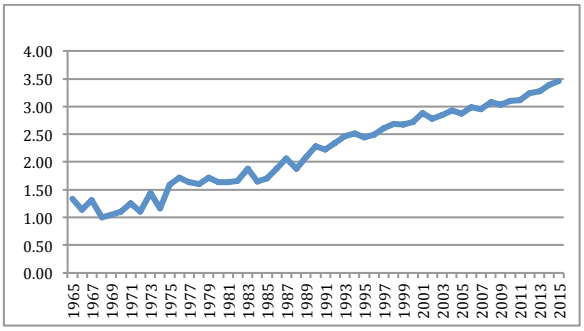

Figure 5. *Average number of authors per paper*

4.3.2. Authors renewal

We then analyzed the authors renewal over time, either the authors who did not publish at the previous conference (*new authors*) or those who had not published at any previous conference (*completely new authors*). It showed (Fig. 6) the percentage of different authors from one year to the next decreased from 100% in 1966 to 61% in 2015, while the number of completely new authors decreased from 100% in 1966 to about 42% in 2015. This suggests a stabilization of the research community over time, but it also reflects a measure of the existence of "new blood" in the field.

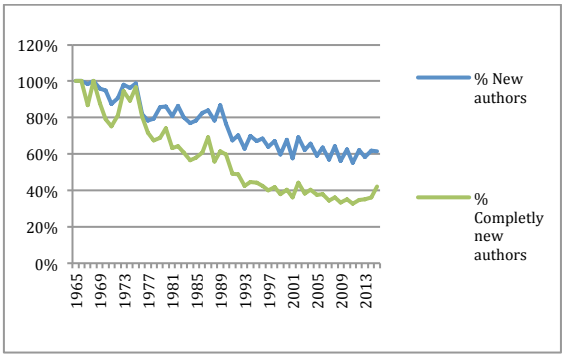

Figure 6. *Percentage of new authors and completely new authors over time.*

4.3.3. Author gender

We conducted a study of the authors gender with the help of a lexicon of 27,509 given names with gender information (66% male, 31% female, 3% epicene [7]). As noted above, 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.) [37], and changes in editorial practices and sharing of the same name by large groups of individuals contribute to make identification by name a difficult problem [35]. In some cases, we only had initials for the first name, which made gender guessing impossible unless the same person appears with his/her first name in full in another publication. 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 should be considered with caution, allowing for an error margin.

The analysis over the thirty four sources shows that 49% of the authors are male (22,858), while 14% of the authors are female (6,746) and 37% are of unknown gender (17,138), either because their given name is epicene, or because we only have the initials of the given name. If we assume that the authors of unknown gender have the same gender distribution as the ones that are categorized, male authors account for 77% and female authors for 23%.

If we consider the situation across the various sources (Fig. 7), we see that the publications related to Signal Processing (*IEEE Transactions on Speech and Language Processing* and ICASSP-S) have the largest participation of male authors (respectively 90 and 88%), while the French conferences and journals, together with LRE and LREC, have the smallest (from 63 to 70%).

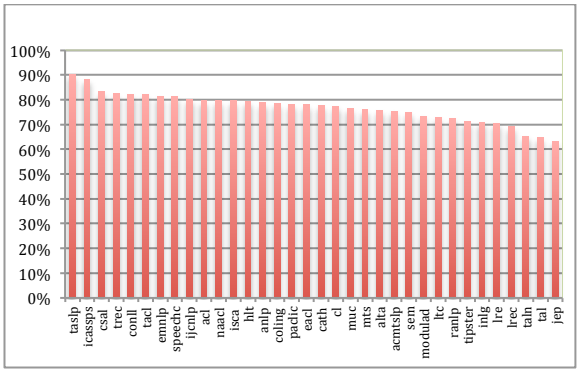

Figure 7. *Percentage of male authors across the sources.*

The analysis of the authors' gender over time (Fig. 8) shows that the ratio of female authors[8] increased over time from 10% to about 20%.

[7] "epicene" means that the given name is gender ambiguous

[8] Those percentages include the number of papers produced

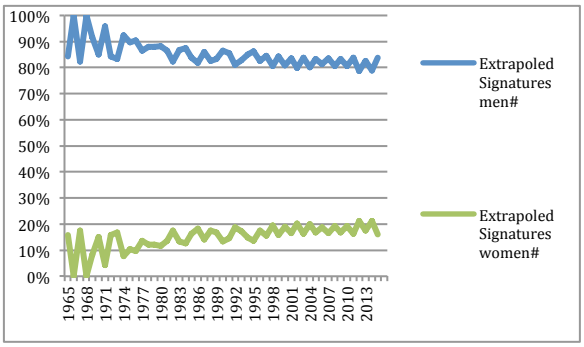

Figure 8. *Gender of the authors' contributions over time.*

5. COLLABORATIONS BETWEEN AUTHORS

5.1. Production and co-production

The most productive author published 358 papers, while 26,870 authors (55% of the 48,894 authors) published only one paper (Fig. 9).

Table 2 gives the list of the 10 most productive authors, accompanied by the number of papers they published as a single author. Table 3 gives the number of authors who published papers as single authors. 42,471 authors (87% of the authors) never published a paper as single author.

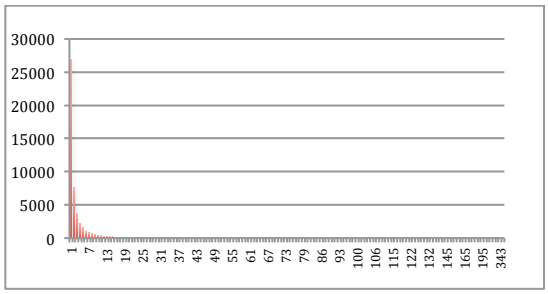

Figure 9. *Number of Papers per Number of Authors*

Name	Number of Papers (= Number of authorships)	Number of Papers as single author[9]
Shrikanth S Narayanan	358	0
Hermann Ney	343	10
John H L Hansen	299	3
Haizhou Li	257	1
Chin-Hui P Lee	218	5
Alex Waibel	207	2
Satoshi Nakamura	205	1
Mark J F Gales	195	9
Lin-Shan Lee	193	0
Li Deng	192	6
Keikichi Hirose	187	1
Kiyohiro Shikano	184	0

Table 2. *10 most productive authors, including the number of papers published as single author*

[9] Keynote papers are not always taken into account if they were not included in the conference programs or proceedings.

# papers	# authors	author name
0	42,471	…
1	4402	…
2	1038	…
3	416	…
4	211	…
5	131	…
6	76	…
7	49	…
8	27	…
9	24	…
10	10	Aravind K Joshi, Eckhard Bick, Hermann Ney, Hugo Van Hamme, Joshua T Goodman, Karen Spärck Jones, Kuldip K Paliwal, Mark Hepple, Raymond S Tomlinson, Roger K Moore
11	10	Dekang Lin, Eduard H Hovy, Jörg Tiedemann, Marius A Pasca, Michael Schiehlen, Olov Engwall, Patrick Saint-Dizier, Philippe Blache, Stephanie Seneff, Tomek Strzalkowski
12	9	David S Pallett, Harvey F Silverman, Jen-Tzung Chien, Kenneth Ward Church, Lynette Hirschman, Martin Kay, Reinhard Rapp, Ted Pedersen, Yorick Wilks
13	4	John Makhoul, Paul S Jacobs, Rens Bod, Robert C Moore
14	2	Dominique Desbois, Sadaoki Furui
15	2	Donna Harman, Takayuki Arai
16	2	Jerry R Hobbs, Steven M Kay
17	2	Beth M Sundheim, Kenneth C Litkowski
18	3	Douglas B Paul, Mark A Johnson, Rathinavelu Chengalvarayan
20	1	Olivier Ferret
21	1	Ralph Grishman
25	1	Ellen M Voorhees
26	1	Jerome R Bellegarda
27	1	W Nick Campbell

Table 3. *Number of single author papers*

The most collaborating author published with 299 different co-authors, while 2,401 authors always published alone (Fig. 10). On average, an author collaborated with 6.6 other authors. Table 4 gives the list of the 12 most co-authoring authors. The two first authors in Table 2 and 4 are the same, but the third one is different.

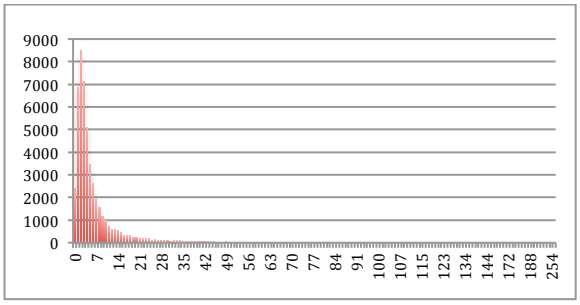

Figure 10. *Number of authors as a function of the number of different co-authors*

Name	# Co-authors
Shrikanth S Narayanan	299
Hermann Ney	254
Haizhou Li	252
Satoshi Nakamura	234
Alex Waibel	212
Mari Ostendorf	199
Chin-Hui P Lee	194
Sanjeev Khudanpur	193
Frank K Soong	188
Lori Lamel	185
Hynek Hermansky	179
Yang Liu	178

Table 4. *The 12 authors with the largest number of co-authors*

5.2. Collaboration graph

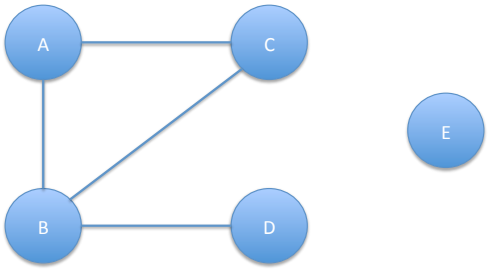

Figure 11. *Collaboration Graph*

A *collaboration graph*[10] (CollG) is a model of a 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. As opposed to a citation graph, a CollG is undirected. It contains no *loop-edge* (an author does not collaborate with him/herself) and no *multiple edges* (there is a single edge between two authors, whatever the number of papers they published together). As it appears in Figure 11, the CollG nodes need not be fully connected, i.e. people who never co-authored a joint paper are represented by isolated nodes (E). Those who are connected constitute a *connected component* (this is the case for A, B, C, D). When a connected component gathers a majority of the nodes, it may be called a *giant component*. *Cliques* are fully connected components where all authors published with one another. The NLP4NLP CollG contains 48,894 nodes corresponding to the 48,894 different authors and 162,497 edges.

Connected Component Size	# of Connected Components	# of authors	% of Authors in the Connected Components	% of Connected Components
39744	1	39744	81%	0%
29	1	29	0%	0%
27	1	27	0%	0%
21	1	21	0%	0%
18	3	54	0%	0%
17	1	17	0%	0%
15	1	15	0%	0%
14	1	14	0%	0%
12	2	24	0%	0%
11	9	99	0%	0%
10	5	50	0%	0%
9	14	126	0%	0%
8	26	208	0%	1%
7	38	266	1%	1%
6	60	360	1%	1%
5	120	600	1%	3%
4	252	1008	2%	5%
3	535	1605	3%	12%
2	1113	2226	5%	24%
1	2401	2401	5%	52%
39963	4585	48894	100%	100%

Table 5. *Connected Components in the Collaboration graph*

[10] http://en.wikipedia.org/wiki/Collaboration_graph

As shown in Table 5, the CollG contains 4,585 connected components. The largest one regroups 39,744 authors, which means that 81% of the 48,894 authors are connected through a collaboration path. The authors of the largest connected component published 58,208 papers (89% of the total number of papers), and the average path length is 5.5. The second connected component regroups 29 authors, who published together, but never with any of the 39,744 previous ones. The remaining connected components contain far fewer authors, each of whom has never published with any of the authors of the larger connected components; these components tend to represent small communities often related to the study of a specific language or a specific topic. As already mentioned, 5% of the authors (2,401) have never published jointly with any other author. As it turned out, in our corpus the largest clique could be identified by simply looking at the paper with the largest number of co-authors (44 authors in NLP4NLP).

Figure 12 gives the percentages of authors in the largest Connected Component for the 34 sources. We see that some conferences international (ISCA, LREC, ICASSP-S, EMNLP, HLT) or national (jep, taln) are more focused than others where the collaboration is more sparse (EACL, ANLP, RANLP).

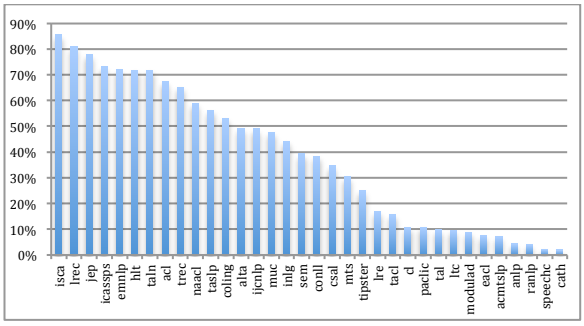

Figure 12. *Percentage of authors in the largest Connected Component of the CollG for the 34 sources*

5.3. Measures of Centrality

We explored the role of each author in the CollG in order to assess his/her centrality. In graph theory, there exist several types of centrality measures [13]. The *Closeness distance* has been introduced in Human Sciences to measure the efficiency of a Communication Network ([2], [3]). It is based on the shortest geodesic distance between two authors regardless of the number of collaborations between the two authors. The *Closeness centrality* is computed as the average closeness distance of an author with all other authors belonging to the same connected component. More precisely, we use the *harmonic centrality* which is a refinement

introduced recently by Y. Rochat [32] of the original formula to take into account the whole graph in one step instead of each connected component separately. The *degree centrality* is simply the number of different co-authors of each author, i.e. the number of edges attached to the corresponding node. The *betweenness centrality* is based on the number of paths crossing a node and reflects the importance of an author as a bridge across different sets of authors (or sub-communities).

Looking at Table 6, we see that some authors who appear in the Top 10 according to the Closeness Centrality also appear in the other two types of centrality, eventually with a different ranking, while others do not.

Closeness centrality			Degree centrality		Betweenness centrality		
Author's name	Harmonic Centrality	Norm on First	Author's name	Index & Norm on First	Author's name	Index	Norm on First
Mari Ostendorf	11958.271	1	Shrikanth S Narayanan	1	Shrikanth S Narayanan	23492104	1
Shrikanth S Narayanan	11890.931	0.994	Hermann Ney	0.854	Haizhou Li	21312971	0.907
Chin Hui P Lee	11869.656	0.993	Haizhou Li	0.854	Satoshi Nakamura	20451472	0.871
Hermann Ney	11824.125	0.989	Satoshi Nakamura	0.784	Chin Hui P Lee	18488513	0.787
Haizhou Li	11803.879	0.987	Alex Waibel	0.714	Hermann Ney	16131472	0.687
Julia B Hirschberg	11756.034	0.983	Mari Ostendorf	0.671	Frank K Soong	15473696	0.659
Nelson Morgan	11700.633	0.978	Sanjeev Khudanpur	0.648	Alex Waibel	14639035	0.623
Sanjeev Khudanpur	11659.186	0.975	Chin Hui P Lee	0.645	Yang Liu	13433061	0.572
Satoshi Nakamura	11657.86	0.975	Frank K Soong	0.635	Lori Lamel	13160473	0.56
Alex Waibel	11655.467	0.975	Lori Lamel	0.625	Khalid Choukri	13150169	0.56

Table 6. *Computation and comparison of the Closeness Centrality, Degree Centrality and Betweenness Centrality for the 10 most central authors.*

6. CITATIONS

6.1. Citation graphs

Unlike the CollG, a *citation graph* (CitG) is directed. (Figure 13). In an *authors citation graph* (ACG), nodes (or vertices) represent individual authors. We may consider the *citing authors graph (CgAG)*, in which a citing author is linked to all the authors of the papers that he/she cites by an edge directed towards those authors; and the *cited authors graph (CdAG)*, where each cited author is linked to the authors who cite him/her by an edge directed towards this author. These graphs may have *loop-edges*, as an author may cite and be cited by him/herself, but they have no *multiple edges*: there is only one edge between two authors, whatever the number of times an author cites or is being cited by another author.

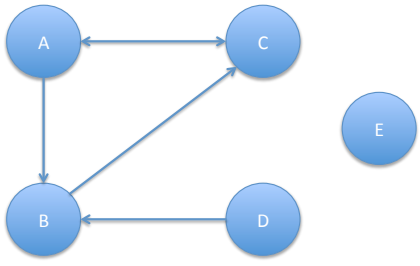

Figure 13. Citation Graph

In a *papers citation graph* (PCG), nodes represent individual papers. Here also, we may consider the *citing papers graph (CgPG)*, in which a paper is linked to all the papers it cites by an edge directed towards those papers; and the *cited papers graph (CdPG)*, where each paper is linked to all the papers that cite it by an edge directed towards those papers. These graphs contain *no loop-edge*, as a paper does not cite itself, and no *multiple edges*: there is only one edge between two papers, whatever the number of times a paper cite or is being cited by another paper.

The citation graphs need not be connected, as an author may not cite any author and may not be cited by any author, not even him/herself, or a paper may not cite any paper and may not be cited by any other paper; in these cases, corresponding authors or papers appear as isolated nodes in the citation graphs (E). The nodes that are connected through a directed path (as it is the case for A, B, C, D in Figure 13 where Author A cites Authors B and C, Author B cites Author C, Author C cites Author A and Author D cites Author B) constitute a *strongly connected component*. The nodes that are connected in both directions constitute a *symmetric strongly connected component*; they are common in ACGs (Author A cites Author B and Author B cites Author A, for example), but uncommon in PCGs, (if Paper M cites Paper N, it is very unlikely that Paper N will cite Paper M, as papers typically reference papers that have been already published. It may however happen in case of simultaneous publications).

We studied those different citation graphs using the full NLP4NLP corpus, and for each of the 34 sources, individually or within the NLP4NLP corpus. We provide some elements of comparison between the publications, keeping in mind that the time scale and frequency are different, for conferences (e.g. 9 venues over 17 years for LREC, 28 venues over 27 years for ISCA, and 36 venues over 35 years for ACL), and for journals. We

considered the 65,003 papers we have in NLP4NLP, which include 324,422 references.

6.2. Citations over time

We studied citations in papers that are accessible in digital form (not the scanned ones, given the poorer quality). 58,204 papers contain a list of references. If we consider the average number of references in papers, we see that it increased over time from close to 0 in 1965 to 8.5 in 2015 (Fig. 14). This is a general trend that goes together with the citing habits and the number of published papers in the literature[11].

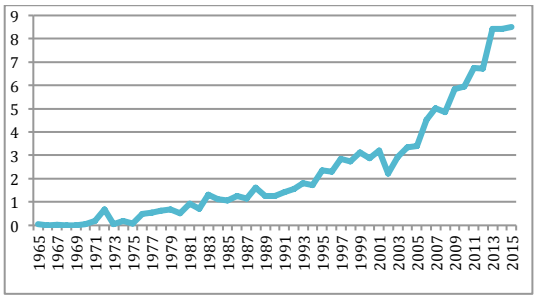

Figure 14. *Average number of references per paper over the years.*

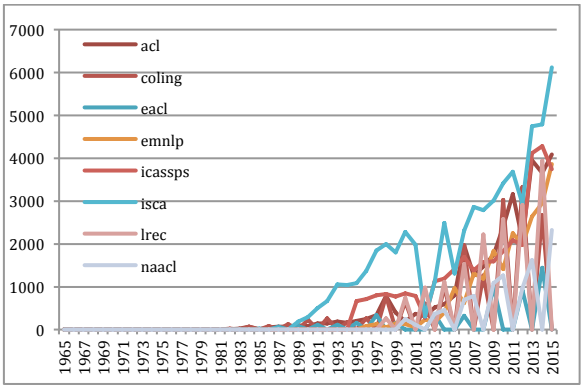

Figure 15. *Number of references in papers over the years for the 8 most important conferences.*

The comparative study of the number of references and of the number of citations over the years for the 34 sources is difficult to handle. If we limit this study to the 8 most important conferences (ACL, COLING, EACL, EMNLP, ICASSP, ISCA, LREC, NAACL) we see that the number of references strongly increased over time in the ISCA conference series (Figure 15). This is directly in agreement with the ISCA Board policy which decided in 2005 to enlarge the number of pages in the yearly conference papers from 6 to 7, with the rule than the allowed extra page should only consist of references, in order to encourage authors to better cite the others' work. The saw tooth aspect of LREC, EACL and NAACL is due to the fact that those conferences are biennial.

Similarly, it is difficult to analyze the variation of cited papers over time due to the different conference frequency. In order to solve this problem, we may integrate the number of number of papers being cited up to the given year. In this case, we see (Fig. 16) that the number of ISCA papers being cited grows at a high rate over time. The same appears for ACL with some delay which is now overcomed. ICASSP comes in the third position. We then find a group of two with COLING and EMNLP, followed by LREC and NAACL. Then comes EACL.

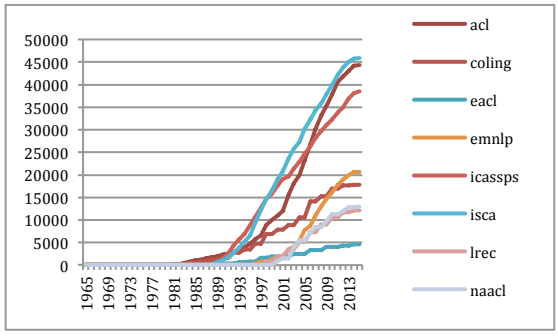

Figure 16. *Number of papers that have been referenced over the years for the 8 most important conferences.*

6.3. Authors citations

We then studied the Authors Citation Graphs (Figure 17) and compared the number of authors in the largest Connected Component for each of the 34 publications. It appears that the authors publishing in a set of north-American publications (*Computational Linguistics*, EMNLP, CONLL, HLT, NAACL, ACL, TACL) are used to cite each other.

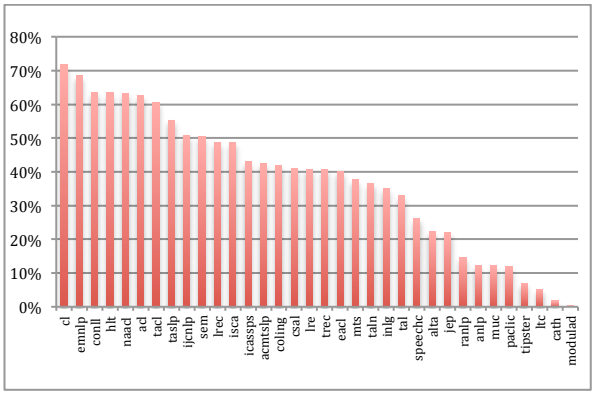

Figure 17. *Percentage of authors in the largest Strongly Connected Component*

Table 7 gives the list of the 10 most cited authors, with the number of citations and the number of

[11] We should however remind that we only consider here the NLP4NLP data

papers written by the author. We see that this ratio may largely vary, some people having few papers but a large audience for this limited set of papers. We also provide the ratio of self-citation (citation of the author in a paper written by the author), which also show various habits.

Name	# References	Nb of papers written by the author	Ratio #references / nb of papers written by the author	Percentage of self-citations
Hermann Ney	5200	343	15.160	17.538
Franz Josef Och	4098	42	97.571	2.221
Christopher D Manning	3972	116	34.241	5.060
Philipp Koehn	3121	39	80.026	2.435
Dan Klein	3080	99	31.111	7.532
Michael John Collins	3077	53	58.057	3.640
Andreas Stolcke	3053	130	23.485	7.141
Mark J F Gales	2540	195	13.026	18.858
Salim Roukos	2505	67	37.388	2.236
Chin-Hui P Lee	2450	218	11.239	18.245

Table 7. *10 most cited authors*

6.4. Papers citations

Figure 18 gives the average number of papers (*mean degree*) of each publication being cited in the complete set of 34 publications. We see that papers published in *Computational Linguistics* are by far the most cited, with more than 20 citations on average. It is followed by NAACL, ACL and EMNLP, then HLT and CONLL. Speech journals (CSAL, TASLP, *Speech Communication*) and especially conferences show lower scores. This is in agreement also with the citation habits of the corresponding communities. Papers are obviously less cited if they are published in languages other than English, as it appears for TAL, TALN, JEP and Modulad.

It is striking to see (Table 8) that 42% of the articles are never coted and that 40% of the authors are never cited. After further investigations in Google Scholar, it appears that some of those authors belong to a different scientific community from neighboring research domains (machine learning, medical engineering, phonetics, general linguistics), in which they are cited, while they rarely published in NLP4NLP.

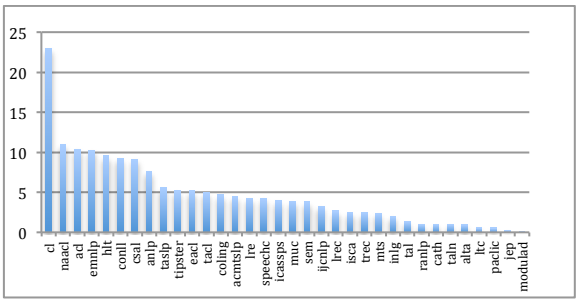

Figure 18. *Mean Degree of papers being cited for the 34 sources*

	Number	%
Never Cited Articles	27,183	42%
Never Cited Authors	19,740	40%

Table 8. *Articles and authors that are never cited*

6.5. H-index

A publication as an H-Index of N if N is the largest number of articles published in that publication that are cited at least N times in NLP4NLP. The computation of the H-Index for the 34 publications (figure 19) shows that the ACL conference has the largest H-Index, with 75 articles cited 75 times or more. It is followed by TASLP (66), Computational Linguistics (58), HLT (56), EMNLP (55), ICASSP-S (54) and ISCA conference series (51). However, it should be stressed that both ACL and ISCA conferences for example cover a much longer time period than LREC.

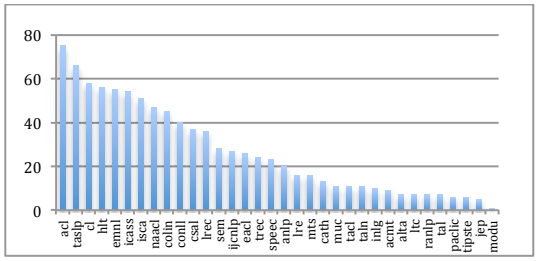

Figure 19. *General H-Index of the 34 sources*

Rank	Source	H-5 Index	H-5 Median
1	Meeting of the Association for Computational Linguistics (ACL)	65	99
2	Conference on Empirical Methods in Natural Language Processing (EMNLP)	56	81
3	IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)	54	73
4	IEEE Transactions on Audio, Speech, and Language Processing (TASLP)	51	78
5	North American Chapter of the Association for Computational Linguistics (NAACL)	48	71
6	International Conference on Spoken Language Processing (INTERSPEECH)	39	70
7	International Conference on Language Resources and Evaluation (LREC)	38	64
8	International Conference on Computational Linguistics (COLING)	38	59
9	arXiv Computation and Language (cs.CL)	37	70
10	Computer Speech & Language (CSL)	32	51
11	Speech Communication (SpeCom)	32	49
12	Computational Linguistics (CL)	31	40
13	Conference on Computational Natural Language Learning (CONLL)	24	36
14	Language Resources and Evaluation (LRE)	23	42
15	International Workshop on Semantic Evaluation (SEMEVAL)	23	41
16	Conference of the European Chapter of the Association for Computational Linguistics (EACL)	21	34
17	International Joint Conference on Natural Language Processing (IJCNLP)	20	27
18	IEEE Spoken Language Technology Workshop (SLT)	18	28
19	Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)	18	27
20	Workshop on Statistical Machine Translation	18	24

Table 9. *Ranking of 20 top sources according to Google Scholar H-Index over 5 last years (2011-2015)*[12]

[12] h5-index is the h-index for articles published in the last 5 complete years. It is the largest number h such that h articles published in 2010-2014 have at least h citations each. h5-median for a publication is the median number of citations for the articles that make up its h5-index.

This analysis on NLP4NLP covers 50 years, but only considers the NLP4NLP publications. It is possible to compare with the Google Scholar[13] H-Index as of March 2016, which considers all the scientific literature, but only within the last 5 years (Table 9). ACL also appears first in the ranking of computational linguistics conferences and journals with an H-index of 65 and an h5-median mean of 99, followed by EMNLP (56), IEEE ICASSP (54), IEEE TASLP (51) and NAACL (48), while one may note the strong upraising of LREC (38) over the 5 last years.

7. USE OF LANGUAGE RESOURCES

We have conducted an analysis of the mention of Language Resources in the corpus. Language Resources are bricks that are being used by researchers to conduct their research investigations and develop their system. We consider here Language Resources in the broad sense embracing data (corpus, lexicons, dictionaries, terminological databases, etc.), tools (morpho-syntactic taggers, prosodic analyzers, annotation tools, etc.), system evaluation resources (metrics, software, training, dry run or test corpus, evaluation package, etc.) and meta-resources (best practices, guidelines, norms, standards, etc.). We considered the Language Resources that are mentioned in the LRE Map [4]. This database was produced in the FlaReNet European project and is constituted by the authors of papers at various conferences of the domain that are invited when submitting their paper to fill in a questionnaire which provides the main characteristics of the Language Resources produced or used in the research investigations that they report in their paper. The LRE Map that we used contains information harvested in 10 conferences from 2010 to 2012, for a total of 4,396 resources. After cleaning those entries (correcting the name of the resources, eliminating the duplicates, regrouping the various versions of resources from the same family, etc.), we ended up with 1,301 different resources that we searched in the NLP4NLP corpus.

Table 10 provides the ranking of Language Resources according to the number of articles where they are mentioned (what we call "*existence*"). It also gives for each resource its type (corpus, lexicon, tool, etc.), the number of mentions in the papers ("occurrences"), the first authors who mentioned it as well as the first publications, and the first and final year when it was mentioned. We see that "WordNet" comes first, followed by "Timit", "Wikipedia", "Penn Treebank" and the "Praat" speech analysis tool.

We studied the evolution of the number of resources compared with the evolution of the number of papers over the years (Figure 20). It appears that the corresponding curves cross in 2005, date since which more than one Language Resource is mentioned on average in a paper. This may reflect the shift from *Knowledge-based* approaches to *Data-driven* approaches.

One may also track the propagation of a Language Resource in the corpus. Figure 21 gives the propagation of the "WordNet" resource, which initially appeared in the HLT conference in 1991, and then propagated on the following years, first in computational linguistics conferences, then also in speech processing conferences.

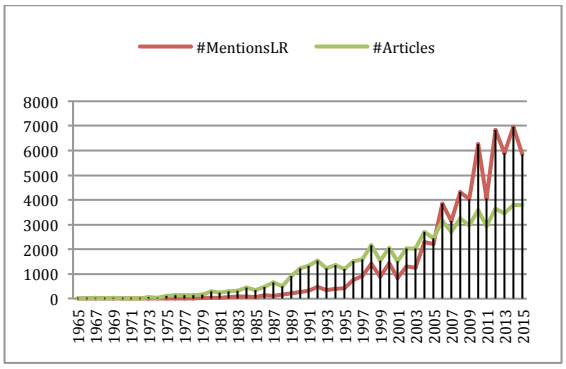

Figure 20. Evolution of the number of mentions of Language Resources in papers over the years

We may attribute an Impact Factor to Language Resources according to the number of articles that mention the resource as it appears in Table 10. Table 11 provides the Impact Factors for the Language Resources of the "data" and "tools" types.

Data	Impact Factor	Tools	Impact Factor
Wordnet	4203	Praat	1254
Timit	3005	SRI Language Modeling Toolkit	1029
Wikipedia	2824	Weka	957
Penn Treebank	1993	GIZA++	758
Europarl	855		
FrameNet	824		

Table 11. *Language Resources Impact factor (data and tools)*

13 http://scholar.google.com/citations?view_op=top_venues&hl=en&vq=eng_computationallinguistics

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

Table 10. *Presence of the LRE Map Language Resources in the NLP4NLP articles*

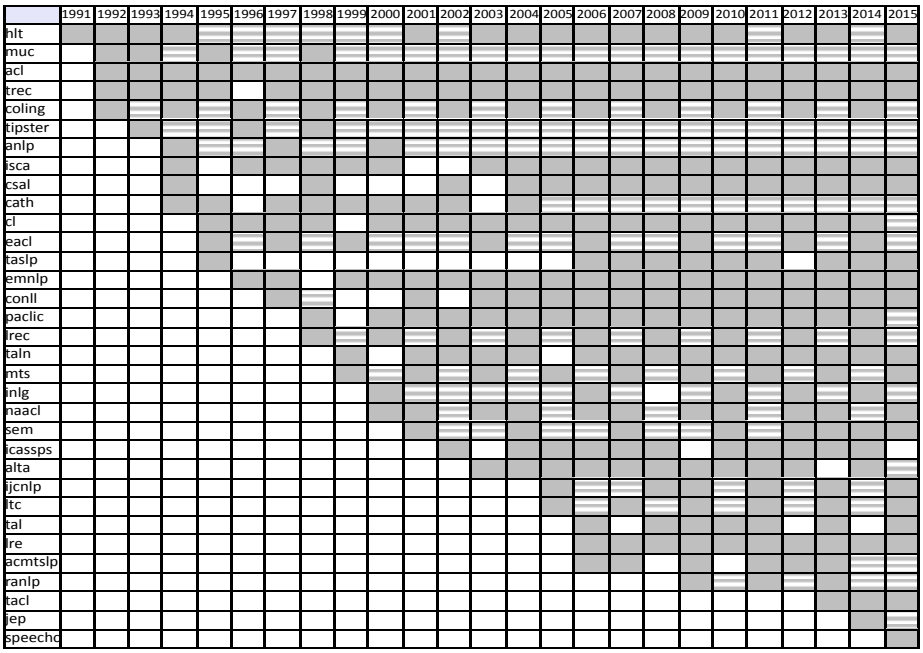

Figure 21. *Propagation of the mention of the "Wordnet" resource in NLP4NLP[14] conferences and journals.*

[14] Hatched slots correspond to years where the conference didn't occurred or the journal wasn't published

8. RESEARCH TOPICS

8.1. Term frequency and presence

Modeling the topics of a research field is a challenge in NLP (see for example (M. Paul et al. 2009), (D. Hall et al., 2008)). Here, our objectives were twofold: i) to compute the most frequent terms used in the domain, ii) to study their variation over time. We start from the NLP4NLPcorpus, which contains a grand total of 271,934,391 words, mostly in English.

Because our aim is to study the terms of the NLP domain, it was necessary to avoid noise from phrases that are used in other senses in the English language. We therefore adopted a contrastive approach, using the same strategy implemented in TermoStat [6]. For this purpose, as a first step, we processed a vast number of English texts that were not research papers in order to compute a statistical language profile. To accomplish this, we applied a deep syntactic parser called TagParser[15] to produce the noun phrases in each text. For each sentence, we kept only the noun phrases with a regular noun as a head, thus excluding the situations where a pronoun, date, or number is the head. We retained the various combinations of sequence of adjectives, prepositions and nouns excluding initial determiners using unigrams, bigrams and trigrams sequences and stored the resulting statistical language model. This process was applied on a corpus containing the British National Corpus (aka BNC)[16] [34], the Open American National Corpus (aka OANC[17]) [16], the Suzanne corpus release-5[18], the English EuroParl archives (years 1999 until 2009)[19], plus a small collection of newspapers in the domain of sports, politics and economy, comprising a total of 200M words. It should be noted that, in selecting this corpus, we took care to avoid any texts dealing with Natural Language Processing.

In a second step, we parsed the NLP4NLP corpus with the same filters and used our language model to distinguish technically specific terms from common ones. We explored 61,661 documents when considering only the papers written in English. They include 3,485,408 different terms (unigrams, bigrams and trigrams) and 23,871,856 term occurrences, that we gathered into synsets, regrouping variation in upper/lower case, singular/plural number, US/UK difference, abbreviation/expanded form and absence/presence of a semantically neutral adjective.

Table 12 gives the ranking of the 10 most frequent terms in the corpus, with the number of occurrences and the frequency. It also includes their variants, the number of articles where they appear ("*Existence*") and its ratio with the number of papers ("*Presence*"). We also computed the average number of occurrences of the terms in the documents where they exist. This ratio varies a lot. In Table 12, it varies from 6.38 for *Speech Recognition* to 10.11 for *Signal to Noise Ratio*.

8.2. Change in Topics

We then studied the evolution of those terms over the years. A visualization software[20] was designed in order to provide the yearly term ranking according to various parameters: time period, number of terms, selection of a set of terms, ranking according to frequency or presence [30]. Figure 22 shows this evolution within the ISCA-Interspeech conference for the terms "HMM" (Hidden Markov Models), "GMM" (Gaussian Mixtures Models), "Annotation", "Neural Networks", "DNN" (Deep Neural Network) and "Dataset". We see the popularity of HMMs, which stayed at the first rank for many years, got rejoined by GMMs, and are now slightly behind. The saw tooth evolution of "Annotation" is due to the biennial frequency of the LREC conference, where this term is frequently used given that the conference is related to Language Resources. Neural Networks got first a high ranking, then declined and are now back to the forefront with the "*Deep Neural Networks"* (DNN) and the accompanying *Datasets* that feed them.

8.3. Tag Clouds for frequent terms

The aim of this section is to provide a global estimation of the main terms used in over the years as well as an indication of the stability of the terms over the years. For this purpose, we use TagCrowd[21] to generate tag clouds[22]. Figure 23 shows the tag clouds in 10 years intervals from 1965 to 2015.

[15] www.tagmatica.com
[16] www.natcorp.ox.ac.uk
[17] www.americannationalcorpus.org
[18] www.grsampson.net/Resources.html
[19] www.statmt.org/europarl

[20] Gapchart: http://vernier.frederic.free.fr/Infovis/rankVis4/
[21] www.tagcrowd.com. Our thanks to Daniel Steinbock for providing access to this web service.

Rank	Term	Variants of all sorts	Archive #Occurrences	Archive frequency	Archive #Existences	Archive Presence	#Occurrences / #Existences
1	HMM	HMMs, Hidden Markov Model, Hidden Markov Models, Hidden Markov model, Hidden Markov models, hidden Markov Model, hidden Markov Models, hidden Markov model, hidden Markov models	135828	0.00618	14362	0.22673	9.46
2	SR	ASR, ASRs, Automatic Speech Recognition, SRs, Speech Recognition, automatic speech recognition, speech recognition	130028	0.00591	20383	0.32178	6.38
3	LM	LMs, Language Model, Language Models, language model, language models	116684	0.00531	13117	0.20707	8.90
4	annotation	annotations	111084	0.00505	11975	0.18904	9.28
5	POS	POSs, Part Of Speech, Part of Speech, Part-Of-Speech, Part-of-Speech, Parts Of Speech, Parts of Speech, Pos, part of speech, part-of-speech, parts of speech, parts-of-speech	102079	0.00464	13834	0.21839	7.38
6	NP	NPs, noun phrase, noun phrases	99074	0.00451	9937	0.15687	9.97
7	classifier	classifiers	98138	0.00446	11545	0.18226	8.50
8	parser	parsers	86137	0.00392	9533	0.15049	9.04
9	segmentation	segmentations	76290	0.00347	10872	0.17163	7.02
10	SNR	SNRs, Signal Noise Ratio, Signal Noise Ratios, signal noise ratio, signal noise ratios	69319	0.00315	6859	0.10828	10.11

Table 12. *10 most frequent terms in the corpus, with number of occurrences, frequency, number of existences and presence.*

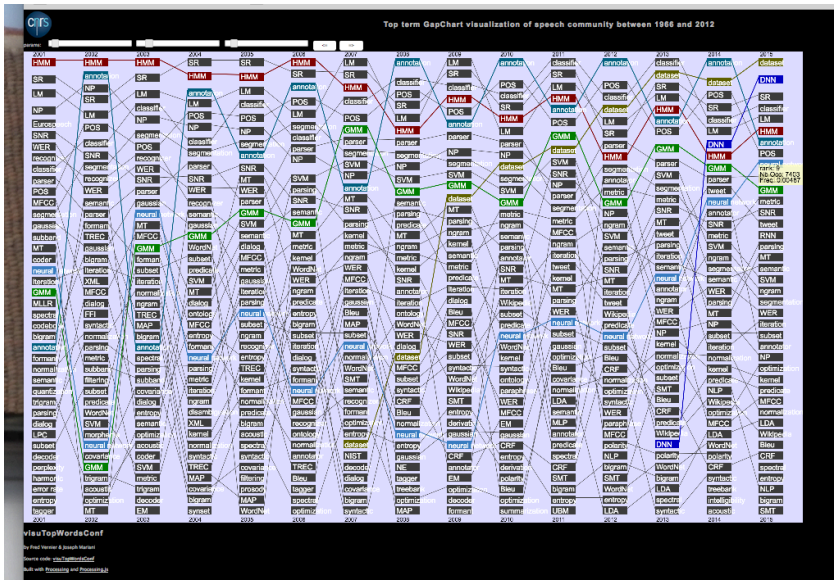

Figure 22. *Evolution over the years of the ranking of the terms according to their frequency for the ISCA-Interspeech conference (2001-2015).*

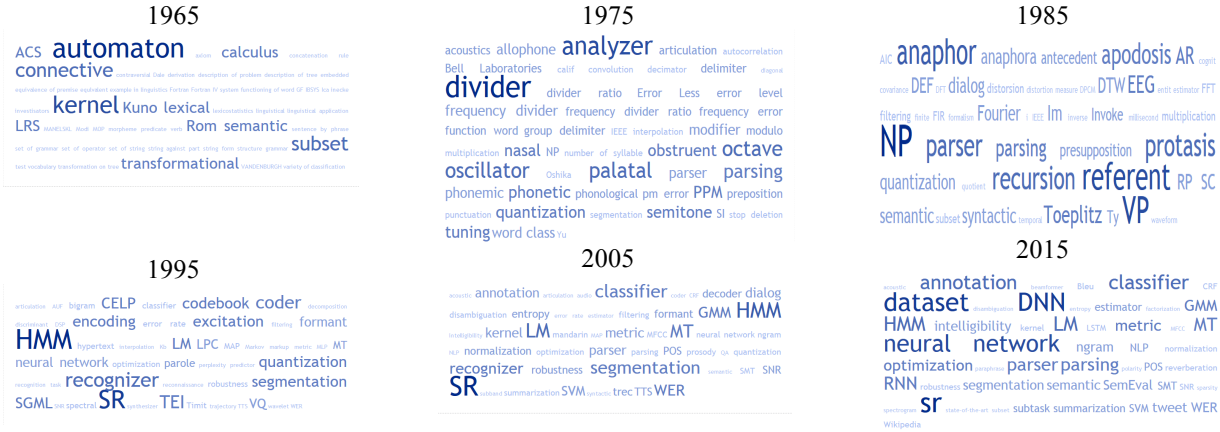

Figure 23. *Tag Cloud based on the abstracts from 1965 to 2015*

Globally, it appears that the most frequent terms changed over the years. In 1965, only COLING is considered. Most of the terms concern computation. In 1975, only *Computer and the Humanities* and the *IEEE Transactions on Acoustics, Speech and Signal Processing* are considered. The Tag Cloud still show a large presence of generic terms, but also of terms attached to audio processing. In 1985, the number of sources is larger and more diversified. The interest for parsing is clear. HMM, and especially discrete models, appear neatly in 1995 together with speech recognition and quantization, while in NLP, *TEI, SGML*, and *MT* are mentioned. The year 2005 shows the interest for Language Resources (*Annotation)* and for evaluation *(metric, WER)*, while *MT* is increasing and *GMM* stands next to *HMM*. 2015 is the year of *neural networks (DNN, RNN)* together with data *(Dataset)*. *Speech Recognition (SR)* stayed popular since 1995 and *Parsing* comes back to the forefront.

8.4. New terms introduced by the authors and by the publications

We studied who introduced new technical terms, when and in which publication, as a mark of the ability of the various authors or publications to bring innovative ideas in the scientific domain. We considered the 61,661 documents written in English and the 42,278 authors who used the 3,485,408 terms contained in those documents.

Rank	Term	Variants of all sorts	Year when the term appeared	Authors who introduced the term	Documents	Number of occurrences of the term in 2015	Number of existences of the term in 2015
1	dataset	data-set, data-sets, datasets	1966	Laurence Urdang	cath1966-3	14039	1472
2	metric	metrics	1965	A Andreyewsky	C65-1002	5425	1108
3	subset	sub set, sub sets, sub-set, sub-sets, subsets	1965	Denis M Manelski, E D Pendergraft, Gilbert K Krulee, Itiroo Sakai, N Dale, Wojciech Skalmowski	C65-1006 C65-1018 C65-1021 C65-1025	3463	1095
4	neural network	ANN, ANNs, Artificial Neural Network, Artificial Neural Networks, NN, NNs, Neural Network, Neural Networks, NeuralNet, NeuralNets, neural net, neural nets, neural networks	1980	Bonnie Lynn Webber	P80-1032	8024	1037
5	classifier	classifiers	1967	Aravind K Joshi, Danuta Hiz	C67-1007	8202	1000
6	SR	ASR, ASRs, Automatic Speech Recognition, SRs, Speech Recognition, automatic speech recognition, speech recognition	1970	Josse De Kock	cath1970-9	8524	1000
7	optimization	optimisation, optimisations, optimizations	1967	Ellis B Page	C67-1032	3331	903
8	annotation	annotations	1967	Kenneth Janda, Martin Kay	cath1967-12 cath1967-8	7515	896
9	POS	POSs, Part Of Speech, Part of Speech, Part-Of-Speech, Part-of-Speech, Parts Of Speech, Parts of Speech, Pos, part of speech, part-of-speech, parts of speech, parts-of-speech	1965	Denis M Manelski, Dániel Várga, Gilbert K Krulee, Makoto Nagao, Toshiyuki Sakai	C65-1018 C65-1022 C65-1029	7489	860
10	LM	LMs, Language Model, Language Models, language model, language models	1965	Sheldon Klein	C65-1014	8522	851

Table 13. *List of the 10 most popular terms in 2015 ranked according to their presence in papers.*

Table 13 provides the list of the 10 most popular terms ranked according to the presence of the term in 2015, which is the final year that we took into consideration and which may reflect their present "success", with the first year when the term appeared, the authors who mentioned it for the first time and the publication where it was mentioned, as well as the number of occurrences and presence in 2015. We see for example that "dataset" was voluntarily introduced by Laurence Urdang [23] in 1966 in *Computer and the Humanities*, that it was mentioned only once on that year, while it appears 14,039 times in 1474 papers in 2015! From its first mention in the introduction of a panel session by Bonnie Lynn Webber at ACL[24] in 1980 to 2015, the number of papers mentioning *Neural Networks* increased from 1 to 1037, and the number of occurrences reached 8,024. *Metric, Subset, Classifier, Speech Recognition, Optimization, Annotation, Part-of-Speech* and *Language Model* are also examples of terms that became very popular over time. Starting from this information, we investigated the possibility to compute an innovation measure that could be attached to an author or a publication.

8.5. A measure of innovation of the terms, authors and publications

8.5.1. Measuring the importance of topics

[23] Laurence Urdang, The Systems Designs and Devices Used to Process The Random House Dictionary of the English Language. Computer and the Humanities, 1966. Interestingly, the author writes: "Each unit of information--regardless of length--was called a dataset, a name which we coined at the time. (For various reasons, this word does not happen to be an entry in The Random House Dictionary of the English Language, our new book, which I shall refer to as the RHD).", a statement which witnesses her authorship of the term.

[24] Interestingly, she mentions the Arthur Clarke's "2001, Space Odyssey" movie: " Barring Clarke's reliance on the triumph of automatic neural network generation, what are the major hurdles that still need to be overcome before Natural Language Interactive Systems become practical?" A premonition in 1980!

We considered the possibility to measure the importance of a term. Figure 24 gives the annual presence (percentage of papers containing the term) for the term "*cross validation*", which was encountered for the first time in 2 papers in 2000[25]. In order to measure the success of the term over time, we compute the sum of the annual presences. We may choose to consider all papers or only those ("external papers" marked in orange) that are written by authors who are different than those who introduced the term (marked in blue).

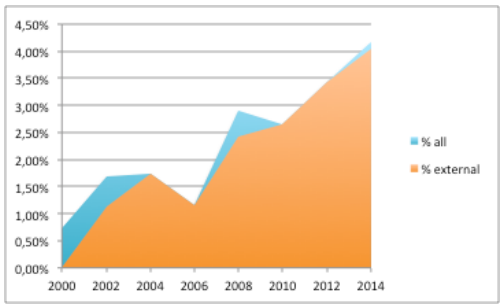

Fig. 24. Presence of the term "cross validation"

We proposed to consider as the annual innovation score the presence of the term on that year. It went from 0.75% of the papers in 2000 to 4% of the papers in 2014. We propose to consider as the global innovation score of the term the corresponding surface, taking into account the inventors' papers in the year of introduction and all the papers in the subsequent years. We see in our example that it takes into account the periods when the term gets more present (2000 to 2004, 2006 to 2008 and 2010 to 2014), as well as those when it looses popularity (2004 to 2006 and 2008 to 2010). The innovation score for the term is the sum of the yearly presences of the term and amounts to 0.17 (17%). This approach emphasizes the importance of the term in the first years when it is mentioned, as the total number of papers is then lower. Some non-scientific terms may not have been filtered out, but their influence will be small as their presence is limited and random. We considered the 1,000 most frequent terms over the 50-year period, as we believe they contain most of the important scientific advances in the field of SNLP. Given the poor quality and low number of different sources and papers in the first years, we decided to only consider for the time being the period from 1975 to 2015. This innovation measure provides an overall ranking of the terms. We also computed separate rankings for NLP and for Speech (Table 14).

[25] "Van Eynde, F.; Zavrel, J. and Daelemans W. (2000), Part of Speech Tagging and Lemmatisation for the Spoken Dutch Corpus" and "Džeroski, S. ; Erjavec, T. and Zavrel J. (2000), Morphosyntactic Tagging of Slovene: Evaluating Taggers and Tagsets"

Rank	Terms		
	Overall	NLP	Speech
1	Speech Recognition	semantic	Speech Recognition
2	Subset	syntactic	Spectral
3	Semantic	NP	Acoustics
4	Filtering	POS	Gaussian
5	HMM	parser	HMM
6	Spectral	parsing	Filtering
7	Linear	subset	Linear
8	iteration	lexical	Fourier
9	Language Model	Machine Translation	Subset
10	POS	predicate	Acoustic

Table 14. Global ranking of the importance of the terms overall and separately for Speech and NLP.

We studied the evolution of the presence of the terms over the years, in order to check the changes in paradigm. However, the fact that some conferences are annual, while others are biennial brings noise. Instead of considering the annual presence of the terms (percentage of papers containing a given term **on** a given year), we therefore considered the cumulative presence of the terms (percentage of papers containing a given term **up to** a given year) (Fig. 25).

We see that *Speech Recognition* has been a very popular topic over the years, reaching a presence in close to 35% of the papers published until 2008. Its shape coincides with *Hidden Markov Models* that accompanied the effort on *Speech Recognition* as the most successful method over a long period. *Semantic* processing was a hot topic of research by the end of the 80's, and regained interest recently. *Language Models* and *Part-of-Speech* received continuing marks of interest over the years.

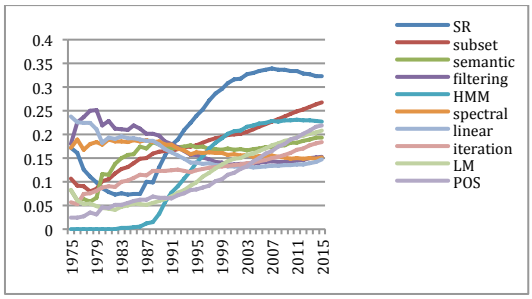

Fig. 25. Cumulative presence of the 10 most important terms over time

8.5.2. Measuring authors' innovation

We also computed in a similar way an *innovation score* for each author, illustrating his or her contribution in the introduction of new terms that subsequently became popular. The score is computed as the sum over the years of the annual presence of the terms in papers published by the authors (percentage of papers containing the term and signed by the author on a given year). This innovation measure provided an overall ranking of

the authors. We also computed separate rankings for NLP and for Speech Processing (Table 15).

Authors		
Overall	NLP	Speech
Lawrence R Rabiner	Ralph Grishman	Lawrence R Rabiner
Hermann Ney	Kathleen R Mckeown	John H L Hansen
John H L Hansen	Jun'Ichi Tsujii	Shrikanth S Narayanan
Shrikanth S Narayanan	Aravind K Joshi	Hermann Ney
Chin Hui P Lee	Jaime G Carbonell	Chin Hui P Lee
Li Deng	Ralph M Weischedel	Li Deng
Mari Ostendorf	Mark A Johnson	Mark J F Gales
Alex Waibel	Fernando C N Pereira	Frank K Soong
Haizhou Li	Christopher D Manning	Haizhou Li
John Makhoul	Ted Briscoe	Thomas Kailath

Table 15. Global ranking of authors overall and separately for Speech and NLP.

We should stress that this measure doesn't place on the forefront the "inventors" of a new topic. It rather helps identifying the early adopters who published a lot after the topic was initially introduced. We studied several cases, such as F. Jelinek and S. Levinson regarding *Hidden Markov Models*, where renowned authors don't appear within the 10 top authors contributing to those terms. We often see that they initially published in a different research field than SNLP (*Information Theory* in the case of F. Jelinek, for example) that we don't consider in our corpus. This measure also reflects the size of the production of papers from the authors on emerging topics, with an emphasis on the pioneering most ancient authors, such as L. Rabiner and J. Makhoul, at a time when the total number of papers was low. The global ranking favors those who published both in Speech and Language Processing, such as H. Ney or A. Waibel.

We may study the domains where the authors brought their main contributions, and how it evolves over time. We faced the same problem due to the noise brought by the different frequency of the conferences as we did when studying the evolution of the terms, and we rather considered the cumulative contribution of the author specific to that term (percentage of papers signed by the author among the papers containing a given term **up to** a given year). We see for example that L. Rabiner brought important early contributions to the fields of *Acoustics*, *Signal Processing* and *Speech Recognition* in general, and specifically to *Linear Prediction Coding (LPC)* and *filtering* (Fig. 26). He even authored 30% of the papers dealing with *LPC* which were published up to 1976 and the only paper mentioning *endpoint detection* in 1975.

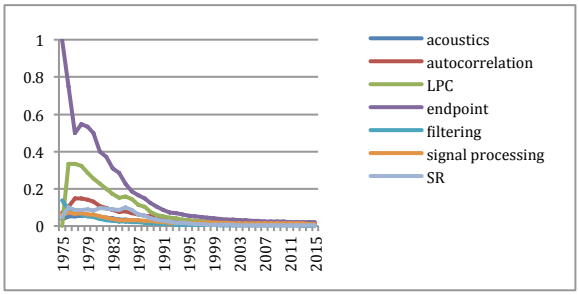

Fig. 26. Main contributions for L. Rabiner

We may also wish to study the contributions of authors on a specific topic, using the same cumulative score. Fig. 27 provides the cumulative percentage of papers containing the term HMM published up to a given year by the 10 most contributing authors. We also added F. Jelinek as a well-known pioneer in that field and S. Levinson as the author of the first article containing that term in our corpus, which represented 0.4% of the papers published in 1982. We see the contributions of pioneers such as F. Soong, of important contributors in an early stage such as C. H. Lee, S. Furui or K. Shikano or later stage such as M. Gales.

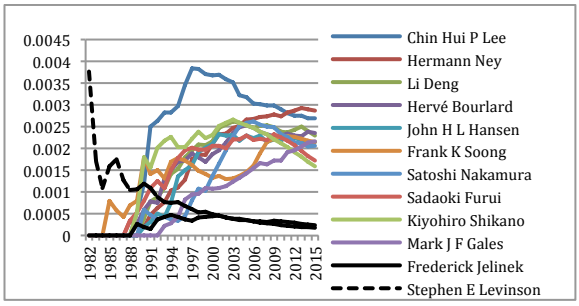

Fig. 27. Authors' contributions to HMM in SNLP

Similarly, we studied the authors' contributions to *Deep Neural Networks (DNN)* which recently gained a large audience (Fig. 28). We see the strong contribution of Asian authors on this topic, with the pioneering contributions of Dong Yu and Li Deng up to 2012 where they represented altogether about 50% of the papers mentioning DNN since 2009, while Deliang Wang published later but with a large productivity which places him at the second rank globally.

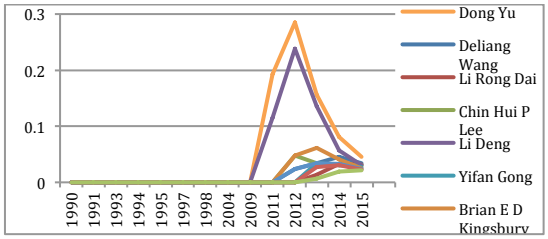

Figure 28. Authors' contributions to the study of DNN in SNLP

8.5.3. Measuring the innovation in publications

We finally computed with the same approach an *innovation score* for each publication. The score is similarly computed as the sum over the years of the annual presence of the terms in papers published in the source, conference or journal (percentage of papers containing the term which were published in the publication on a given year). This innovation measure provided an overall ranking of the publication. We also computed separate rankings for NLP and for Speech Processing (Table 16).

Rank	Sources		
	Overall	NLP	Speech
1	taslp	acl	taslp
2	isca	coling	isca
3	icassps	cath	icassps
4	acl	lrec	lrec
5	coling	cl	csal
6	lrec	hlt	speechc
7	hlt	eacl	mts
8	emnlp	emnlp	ltc
9	cl	trec	lre
10	cath	mts	acmtslp

Table 16. Global ranking of the importance of the sources overall and separately for Speech and NLP.

Just as in the case of authors, the measure also reflects here the productivity, which favors the Speech Processing field where more papers have been published, and the pioneering activities, as reflected by the ranking of *IEEE TASLP*. In the overall ranking, publications that concern both Speech and Language Processing (LREC, HLT) get a bonus.

We may study the domains where the publications brought their main contributions, and how it evolves over time. We faced the same problem due to the noise brought by the different frequency of the conferences as we did when studying the evolution of the terms and authors, and we rather considered the cumulative contribution of the publication specific to that term (percentage of papers published in the source among the papers containing the term **up to** a given year). We see for example (Fig. 29) that ACL showed a strong activity and represented 40% of papers published about *parsing*, 35% of papers published about *semantic, syntactic* and *lexical* and 25% of papers published about *Machine Translation* up to 1985. Its share in those areas then globally decreases to about 15% of the total number of publications, due to the launching of new conferences and journals, while the share of publications on *Machine Translation* within ACL increased.

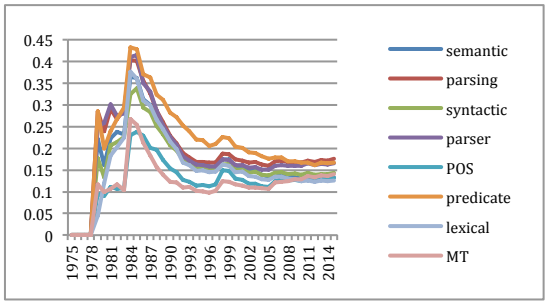

Fig. 29. Main domains within the ACL conference series

We may also wish to study the contributions of publications to a specific term, using the same cumulative score. Fig. 30 provides the cumulative percentage of papers containing the term HMM published up to a given year by the 10 most contributing publications. We see that all papers were initially published in the *IEEE Transactions on Speech and Audio Processing*. Other publications took a share of those contributions when they were created (*Computer Speech and Language* starting in 1986, *ISCA Conference series* starting in 1987) or when we start having access to them (*IEEE-ICASSP*, starting in 1990). We see that *ISCA Conference series* represents 45% of the papers published on HMM up to 2015, while *IEEE-ICASSP* represents 25%. We also see that HMMs were first used in speech processing related publications, then in NLP publications as well (ACL, EMNLP), while publications that are placed in-between (CSL, HLT, LREC) helped spreading the approach from speech to NLP.

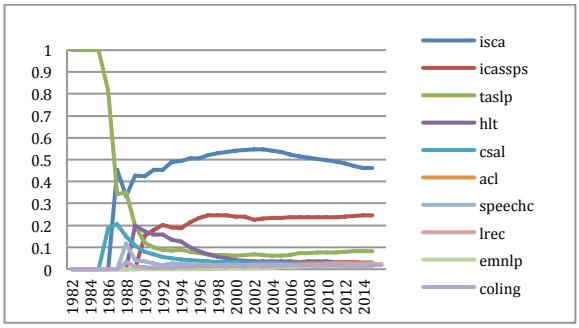

Fig. 30. Sources' contributions to the study of HMM

8.6. Research Topic Prediction

We also explored the feasibility of predicting the research topics for the coming years based on the past. We used for this the Weka[26] machine learning software environment [36]. We applied each software contained in Weka to the time series of terms ordered according to their frequency and

[26] www.cs.waikato.ac.nz/ml/weka

retained the software which provided the best results with the corresponding set of optimal parameters (especially the history time length), after a-posteriori verification. We then applied this software to the full set of the NLP4NLP corpus. Table 17 gives the ranking of the most frequent terms in 2013 and 2014 with their frequency, the topic predicted for 2015 on the basis of the past rankings and the ranking actually observed in 2015. We see that the prediction is correct for the top term ("dataset"). The next predicted term was "annotation" which only appears at the 9th rank, probably due to the fact that LREC didn't take place in 2015. It is followed by "POS", which actually appears at the 4th rank.

Observed in 2013	Observed in 2014	Predicted for 2015	Observed in 2015	Rank
classifier (0.00576)	annotation (0.00792)	dataset (0.00653)	dataset (0.00886)	1
LM (0.00565)	dataset (0.00639)	annotation (0.00626)	DNN (0.00613)	2
dataset (0.00548)	POS (0.00600)	POS (0.00549)	classifier (0.00491)	3
POS (0.00536)	LM (0.00513)	LM (0.00479)	POS (0.00485)	4
annotation (0.00509)	classifier (0.00507)	classifier (0.00466)	neural network (0.00455)	5
SR (0.00507)	SR (0.00449)	DNN (0.00437)	LM (0.00454)	6
HMM (0.00478)	parser (0.00388)	SR (0.00429)	SR (0.00439)	7
parser (0.00404)	DNN (0.00369)	HMM (0.00365)	parser (0.00436)	8
GMM (0.00367)	HMM (0.00352)	neural network (0.00345)	annotation (0.00414)	9
segmentation (0.00298)	neural network (0.00326)	tweet (0.00312)	HMM (0.00384)	10

Table 17. *Research topics prediction using the Weka software environment.*

As we have the information on the actual observations in the annual rankings, it is possible to measure the reliability of the predictions by measuring the distance between the predicted frequencies and the observed frequencies. Figure 31 gives this distance for the predictions in year 2011 to 2015 based on time series until 2010. We see the distance largely increases in 2013, that is three years after the year of prediction. We may therefore think that it is not reasonable to predict the future of a research domain beyond a 2-year horizon (unless a major discovery happens in the meanwhile…).

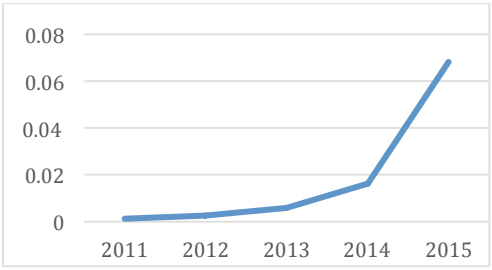

Figure 31. *Reliability of the predictions: prediction error over the years from 2011*

It is possible to measure the difference between the prediction and the observation in each year. It provides a measure of the "surprise" between what we were expecting and what actually occurred. The years where this "surprise" is the largest may correspond to epistemological ruptures. Figure 32 gives the evolution of thos distance between 2011 and 2015. We see that 2012 was a year of big changes.

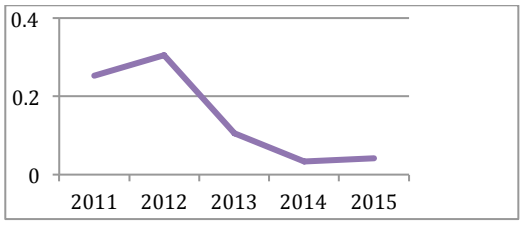

Figure 32. *Evolution of the distance between prediction and observation over the years as a measure of "surprise" that may correspond to an epistemological rupture.*

We may also compute this distance for a specific topic, in order to analyze the way this term evolves compared with what was expected. Figure 33 shows the evolution of the *"Deep Neural Network" (DNN)* topic. We see that up to 2014, we didn't expect the success of this approach, while starting in 2014, it became part of the usual set of tools for automatic language processing.

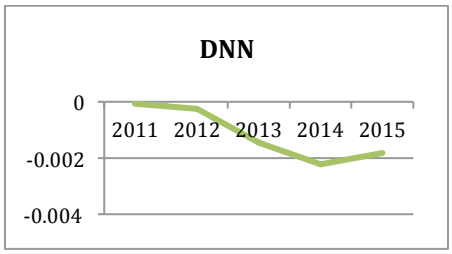

Figure 33. *Measure of the expectation of an emerging research topic: Deep Neural Networks*

Table 18 provides the predictions for the next five years starting in 2016 : not surprisingly, it is expected that neural networks, more or less deep and more or less recurrent, will keep on attracting the researchers' attention.

Observed 2014	Observed 2015	Prediction 2016	Prediction 2017	Prediction 2018	Prediction 2019	Prediction 2020	Rank
annotation	dataset	dataset	dataset	dataset	dataset	dataset	1
dataset	DNN	DNN	DNN	DNN	DNN	DNN	2
POS	classifier	annotation	neural network	neural network	neural network	neural network	3
LM	POS	POS	SR	RNN	RNN	RNN	4
classifier	neural network	neural network	classifier	POS	parser	parser	5
SR	LM	classifier	LM	parser	SR	SR	6
parser	SR	parser	POS	annotation	LM	metric	7
DNN	parser	SR	RNN	classifier	classifier	POS	8
HMM	annotation	LM	parser	SR	metric	parsing	9
neural network	HMM	HMM	HMM	metric	POS	classifier	10

Table 18. *Predictions for the next five years 2016-2020*

9. Text reuse and plagiarism

We finally studied text reuse and plagiarism within NLP4NLP papers. In order to do so, we compared one by one the 65,003 NLP4NLP articles written by the 48,894 authors, after conducting a deep syntactic analysis using TagParser [8] in order to reduce the influence of the style variants and to exclude general language expressions. The comparison between an article and all the articles which were published beforehand or on the same year is then conducted by comparing windows of seven lexical entities using the *Jaccard distance* and, after several experiments, we retained the couples of papers that have a similarity of 4% or more. We then consider four different cases: in the cases where two articles have at least one author in common, if the source paper is cited, we will name it "self-reuse", else "self-plagiarism". If the two articles have no author in common, if the source paper is cited, we will name it "reuse", else "plagiarism" (Table 19).

>4% similarity	Source is quoted	Source is not quoted
At least one author in both papers	Self-Reuse	Self-Plagiarism
No author in common	Reuse	Plagiarism

Table 19. *Definitions of (self-)reuse and (self-) plagiarism*

The results show that the number of self-reuse and self-plagiarism is very important (about 18% of the articles) (Figure 34). This number is too important for conducting a manual verification. 205 articles have the same title and 130 articles have the same title and exactly the same list of authors! Table 20 gives the number of self-reused or self-plagiarized papers for each publication pairs. We see that the flow of articles is especially large between the IEEE-ICASSP and ISCA-Interspeech conferences, as well as between the conferences and journals from the same domain, such as IEEE-ICASSP or ISCA-Interspeech and TASLP, CSAL or Speech Com, which seems quite normal.

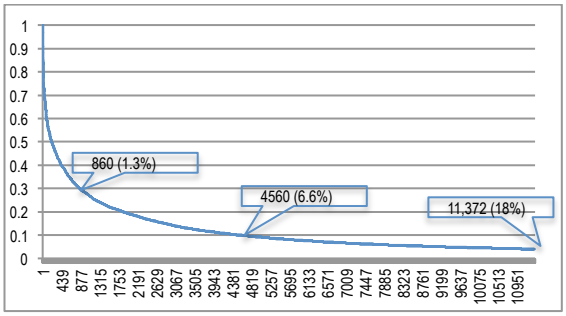

Figure 34. *Similarity scores of the couples detected as self-reuse / self-plagiarism*

On the contrary, the number of reuses and plagiarisms is very low and concerns only 0.3% of the articles (Figure 35). Table 21 provides the number of articles being identified as reused or plagiarized. Here, a manual checking was possible as the number of cases is low and showed that almost all the detected cases were not real plagiarism (wrong spelling of the names of the cited authors or of the title of the cited paper, correct referencing of another article, reference to a third paper, etc.).

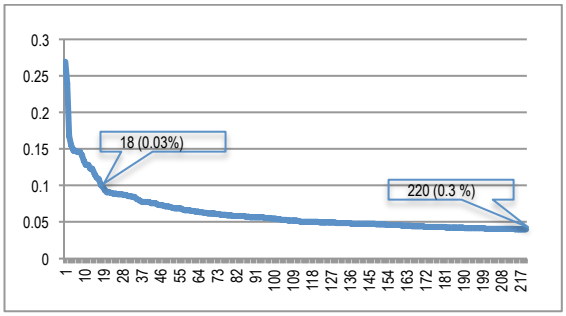

Figure 35. *Similarity scores of the couples detected as reuse / plagiarism*

Used \ Using	acl	acmtslp	alta	anlp	cath	cl	coling	conll	csal	eacl	emnlp	hlt	icassps	ijcnlp	inlg	isca	jep	lre	lrec	ltc	modulad	mts	muc	naacl	paclic	ranlp	sem	speechc	tacl	tal	taln	taslp	tipster	trec	Total used	Total using	Difference	
acl	22	8	1	4	8	136	78	25	31	22	83	85	29	31	7	48	0	20	71	4	0	19	1	51	8	5	26	1	2	0	0	24	4	9	863	625	238	acl
acmtslp	1	0	0	0	0	0	0	0	2	0	0	2	3	2	0	6	0	1	1	0	0	0	0	2	0	0	1	0	1	0	0	2	0	0	24	93	-69	acmtslp
alta	3	0	2	0	0	1	5	0	1	2	5	0	0	1	0	4	0	0	4	0	0	0	0	0	0	1	0	0	0	0	0	0	0	4	33	14	19	alta
anlp	7	0	0	1	3	5	8	1	1	2	1	4	0	0	0	1	0	0	5	0	0	1	0	2	1	0	0	0	0	0	0	0	2	5	50	50	0	anlp
cath	1	0	0	1	7	2	0	0	0	1	0	1	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	18	50	-32	cath
cl	9	0	0	4	3	0	4	0	2	4	3	1	0	0	0	0	0	2	5	0	0	0	0	0	1	0	4	0	0	0	0	0	0	0	42	433	-391	cl
coling	74	10	3	8	7	62	19	24	17	15	43	49	8	24	7	42	0	14	90	4	0	9	2	33	12	5	25	3	0	0	0	12	6	5	632	500	132	coling
conll	26	1	1	1	1	20	18	8	5	6	16	11	2	14	2	2	0	2	10	1	0	3	0	7	0	5	13	0	1	0	0	3	0	0	179	151	28	conll
csal	3	0	0	0	0	4	4	2	7	0	3	2	20	1	0	35	0	2	7	0	0	0	0	0	0	0	2	6	0	0	0	13	0	0	111	643	-532	csal
eacl	16	2	0	2	5	31	12	6	3	1	8	13	3	1	2	9	0	0	21	1	0	1	0	13	1	1	4	0	0	0	0	5	0	1	162	130	32	eacl
emnlp	103	2	2	1	2	44	52	26	18	9	16	30	14	47	1	27	0	5	29	0	0	7	0	22	2	1	19	0	3	0	0	20	1	5	508	355	153	emnlp
hlt	83	12	0	5	3	48	48	11	42	14	33	22	29	30	2	104	0	4	26	1	0	13	2	6	1	0	9	8	0	0	0	25	7	19	607	476	131	hlt
icassps	16	5	0	0	0	3	4	1	130	4	7	21	262	2	0	1005	0	0	19	0	0	2	0	14	2	0	0	65	0	0	0	746	0	3	2311	2160	151	icassps
ijcnlp	27	6	1	0	0	3	29	10	7	2	34	18	2	4	3	7	0	5	19	3	0	9	0	13	4	8	3	0	0	0	0	4	0	1	222	237	-15	ijcnlp
inlg	7	0	0	1	1	6	5	2	0	3	1	3	0	1	2	4	0	1	6	0	0	1	0	4	0	0	0	0	0	0	0	1	0	0	49	35	14	inlg
isca	56	23	0	2	0	13	45	0	317	10	25	116	1531	10	4	879	0	10	133	19	0	12	0	38	6	0	1	233	0	0	0	669	0	5	4157	2460	1697	isca
jep	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	16	18	-2	jep
lre	2	1	0	0	0	2	3	0	0	0	0	1	0	0	0	2	0	2	6	0	0	0	0	1	1	0	0	0	0	0	0	1	0	0	22	146	-124	lre
lrec	58	3	0	2	6	16	80	6	13	15	16	17	16	10	2	72	0	52	67	12	0	6	0	11	11	4	12	5	2	0	0	6	1	3	524	660	-136	lrec
ltc	4	0	0	0	0	0	0	0	0	0	0	0	0	2	0	15	0	1	35	10	0	2	0	0	6	6	1	4	0	0	0	0	0	0	86	71	15	ltc
modulad	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	modulad
mts	13	0	0	0	0	2	9	2	0	2	9	10	3	9	0	9	0	2	20	2	0	8	0	8	5	2	1	1	0	0	0	2	0	0	119	109	10	mts
muc	2	0	0	2	0	2	3	0	0	1	0	7	0	0	0	0	0	0	0	0	0	0	10	1	0	0	0	0	0	0	0	0	18	1	47	28	19	muc
naacl	46	10	0	2	1	24	30	7	12	11	22	5	15	22	3	30	0	3	16	1	0	9	0	3	0	0	9	1	0	0	0	8	0	3	293	251	42	naacl
paclic	4	0	0	0	1	0	12	1	1	1	1	0	2	8	0	3	0	5	18	7	0	3	0	0	21	7	1	0	0	0	0	1	0	0	97	85	12	paclic
ranlp	3	2	0	0	0	0	2	4	4	2	2	1	0	7	0	0	0	2	19	5	0	2	0	1	2	4	2	1	0	0	0	0	0	1	66	54	12	ranlp
sem	25	2	0	0	0	7	16	14	4	1	12	12	0	8	0	0	0	13	12	1	0	1	0	8	1	4	53	0	0	0	0	0	0	1	195	188	7	sem
speechc	0	0	0	0	0	1	1	0	11	0	0	4	17	0	0	48	0	0	2	0	0	0	0	0	0	0	0	1	0	0	0	17	0	0	102	344	-242	speechc
tacl	1	1	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	7	9	-2	tacl
tal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	13	0	0	0	18	59	-41	tal
taln	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	53	9	0	0	0	65	22	43	taln
taslp	0	5	0	0	0	0	1	1	13	0	1	4	197	0	0	103	0	0	2	0	0	1	0	2	0	0	0	15	0	0	0	49	0	0	394	1610	-1216	taslp
tipster	3	0	0	3	0	0	6	0	0	0	1	5	0	0	0	0	0	0	2	0	0	0	13	1	0	0	0	0	0	0	0	0	2	7	43	65	-22	tipster
trec	10	0	4	11	2	1	6	0	2	2	11	32	7	3	0	5	0	0	10	0	0	0	0	10	0	1	1	0	0	0	0	2	24	287	431	362	69	trec
Total using	625	93	14	50	50	433	500	151	643	130	355	476	2160	237	35	2460	18	146	660	71	0	109	28	251	85	54	188	344	9	59	22	1610	65	362	12493	12493	0	

Table 20. *Self-reuse and Self-Plagiarism Matrix, with indication of the 7 most using and used sources.*

Used \ Using	acl	acmtslp	alta	anlp	cath	cl	coling	conll	csal	eacl	emnlp	hlt	icassps	ijcnlp	inlg	isca	jep	lre	lrec	ltc	modulad	mts	muc	naacl	paclic	ranlp	sem	speechc	tacl	tal	taln	taslp	tipster	trec	Total used	Total using	Difference	
acl	1	0	0	0	1	1	2	2	0	0	4	3	0	3	0	2	0	0	1	1	0	0	1	1	1	1	3	0	0	0	0	0	0	0	28	7	21	acl
acmtslp	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	acmtslp
alta	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	alta
anlp	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	anlp
cath	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	-2	cath
cl	0	0	0	0	0	0	1	0	0	0	1	1	0	1	0	0	0	0	4	0	0	1	0	1	2	0	0	0	0	0	0	0	0	0	12	5	7	cl
coling	0	0	0	0	1	0	0	0	0	0	0	2	1	1	0	2	0	0	2	0	0	0	1	1	1	0	2	0	0	0	0	1	0	0	15	7	8	coling
conll	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	3	5	-2	conll
csal	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	3	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	7	6	1	csal
eacl	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	0	eacl
emnlp	0	0	0	0	0	2	0	2	0	1	1	2	0	0	0	0	0	0	0	0	0	1	0	2	0	0	0	0	0	0	0	0	0	2	13	15	-2	emnlp
hlt	2	0	0	0	0	1	0	1	1	0	2	1	1	1	0	2	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	2	17	17	0	hlt
icassps	0	0	0	0	0	0	0	0	1	0	1	2	3	0	0	32	0	0	0	0	0	0	0	2	0	0	0	2	0	0	0	5	0	0	48	37	11	icassps
ijcnlp	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	2	9	-7	ijcnlp
inlg	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	inlg
isca	0	0	0	0	0	1	1	0	1	0	0	1	18	1	0	7	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	3	0	0	36	70	-34	isca
jep	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	jep
lre	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	-1	lre
lrec	0	0	0	0	0	0	0	0	1	0	2	0	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	1	8	8	0	lrec
ltc	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	-4	ltc
modulad	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	modulad
mts	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	4	3	1	mts
muc	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	3	3	0	muc
naacl	1	0	0	0	0	0	0	0	1	0	1	1	1	0	0	1	0	0	0	0	0	0	0	0	1	2	0	0	0	0	0	0	0	0	9	10	-1	naacl
paclic	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	2	10	-8	paclic
ranlp	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	-3	ranlp
sem	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	3	7	-4	sem
speechc	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	4	5	-1	speechc
tacl	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	tacl
tal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	tal
taln	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	taln
taslp	0	0	0	0	0	0	0	0	1	1	0	0	10	0	0	16	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	30	10	20	taslp
tipster	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	2	2	0	tipster
trec	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	13	13	0	trec
Total using	7	0	0	0	2	5	7	5	6	2	15	17	37	9	0	70	0	1	8	4	0	3	3	10	10	3	7	5	0	0	0	10	2	13	261	261	0	

Table 21. *Reuse and Plagiarism Matrix, with indication of the 7 most using and used sources*

We then studied the time delay between a first publication and its reuse (Figure 36). It appears that 38% of the reuse are done on the same year, 71% on the following year, 83% within the next two years and 93% within the next three years. 30% of the similar papers published on the same year concern the couple of conferences ISCA-ICASSP.

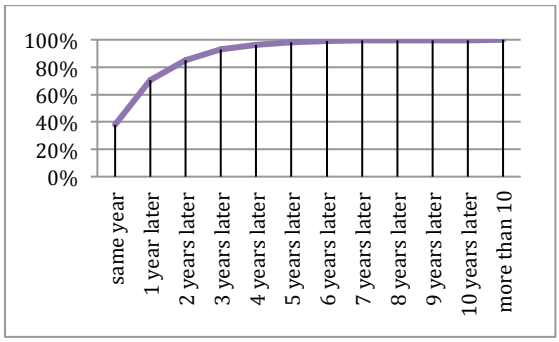

Figure 36. *Time delay between publication and reuse (in %)*

We now consider the reuse of conference papers in journal papers (Figure 37). We observe here a similar time schedule, with a delay of one year: 12% of the reused papers were published on the same year, 41% within the next year, 68% over 2 years, 85% over 3 years and 93% over 4 years.

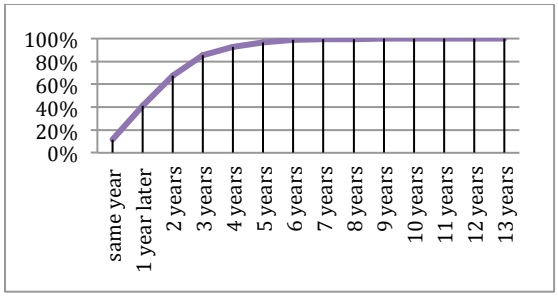

Figure 37. *Time delay between publication in conferences and reuse in journals (in %)*

10. CONCLUSIONS AND PERSPECTIVES

We have presented here an overall survey of the main results of an analysis of the large NLP4NLP corpus which covers a large part of the publications related to Natural Language Processing over a long and recent period of 50 years (1965-2015), where major advances have been achieved thanks to continuous and constant research efforts benefiting from the existence of an infrastructure gathering incentive research programs, language resources availability and regular organization of evaluation campaigns.

We struggled in this analysis with the lack of a consistent and uniform identification of entities (such as authors names, gender, affiliations, paper language, conference and journal titles, funding agencies, etc.). Establishing standards for such identification would considerably help, but will demand an international effort in order to ensure that the identifiers are unique and persistent, which appears as a challenge for the scientific community.

We still have to refine our innovation measure and we would like to better automatize the extraction of terms and authors' names while reducing the error rate by considering the context in which they appear, analyze citations polarity and better identify weak signals which may indicate the raise of a new scientific paradigm that may come from sources that are far away from the domain we study.

11. 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 a 14 GBytes ISCA Archive, Emmanuelle Foxonet who provided a list of authors given names with genre, Florian Boudin, who made available the TALN Anthology, Helen van der Stelt and Jolanda Voogd (Springer) who provided the LRE data and Douglas O'Shaughnessy, Denise Hurley, Rebecca Wollman and Casey Schwartz (IEEE) who provided the IEEE ICASSP and TASLP data. They also thank Khalid Choukri, Alexandre Sicard and Nicoletta Calzolari, who provided information about the past LREC conferences, Nicoletta Calzolari, Riccardo del Gratta, Khalid Choukri, Irene Russo, Francesco Rubino et Claudia Soria for producing and distributing the LRE Map, 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 17 years conferences without whom this analysis could not have been conducted!

12. APOLOGIES

This survey has been made on textual data, which cover a 50-year period, including scanned content. The analysis uses tools that automatically process the content of the scientific papers and may make errors. Therefore, the results should be regarded as reflecting a large margin of error. The authors wish to apologize for any errors the reader may detect, and they will gladly rectify any such errors take in future releases of the survey results.

13. REFERENCES

[1] 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

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

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

[4] Calzolari, Nicoletta; Del Gratta, Riccardo; Francopoulo, Gil; Mariani, Joseph; Rubino, Francesco; Russo, Irene and Soria, Claudia (2012), The LRE Map. Harmonising Community Descriptions of Resources, In Proceedings of the Language Resources and Evaluation Conference (LREC 2012), Istanbul, Turkey, 23-25 May 2012.

[5] Ding, Ying; Rousseau, Ronald and Wolfram, Dietmar ed. (2014), *Measuring Scholarly Impact*, Springer. 2014, ISBN: 978-3-319-10376-1.

[6] 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.

[7] Dunne, C.; Shneiderman, B.; Gove, R.; Klavans, J. and Dorr, B. (2012), Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization. Journal of the American Society for Information Science and Technology, 63(12), 2351-2369.

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

[9] 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.

[10] Francopoulo, Gil; Mariani, Joseph and Paroubek, Patrick (2015a) NLP4NLP: The Cobbler's Children Won't Go Unshod, 4th International Workshop on Mining Scientific Publications (WOSP2015), Joint Conference on Digital Libraries 2015 (JCDL 2015), Knoxville (USA), June 24, 2015.

[11] Francopoulo, Gil; Mariani, Joseph and Paroubek, Patrick (2015b) NLP4NLP: Applying NLP to written and spoken scientific NLP corpora, Workshop on Mining Scientific Papers: Computational Linguistics and Bibliometrics, 15th International Society of Scientometrics and Informetrics Conference (ISSI 2015), Istanbul (Turkey), June 29, 2015.

[12] Francopoulo Gil, Mariani Joseph, Paroubek Patrick (2015c). NLP4NLP: the cobbler's children won't go unshod, in D-Lib Vol. 21, N° 11/12, Nov./Dec. 2015[27].

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

[14] Gollapalli, Sujatha Das and Li, Xiao-li (2015) EMNLP versus ACL: Analyzing NLP Research Over Time, EMNLP 2015, Lisbon (Portugal), September 17-21, 2015

[15] Hall, David Leo Wright ; Jurafsky, Daniel and Manning, Christopher (2008) Studying the History of Ideas Using Topic Models, In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '08), 363–371.

[16] 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.

[17] Jha, Rahul; Jbara, Amjad-Abu; Qazvinian, Vahed and Radev, Dragomir R. (2016) NLP-driven citation analysis for scientometrics. Natural Language Engineering, Available on CJO 2016, doi:10.1017/S1351324915000443

[18] 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.

[19] Li, H.; Councill, I.; Lee, W.C. and Giles, C.L. (2006) CiteSeerx: an architecture and web service design for an academic document search engine, In: Proceedings of the 15th Int. Conference on the World Wide Web.

[20] 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.

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

[22] 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.

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

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

[25] Mariani, Joseph; Francopoulo, Gil; Paroubek, Patrick and Vetulani, Zygmunt (2015), Rediscovering 10 to 20 Years of Discoveries in Language & Technology, Proceedings of L&TC 2015, 27-29 November 2015, Poznan, Poland.

[26] Mariani, Joseph; Paroubek, Patrick; Francopoulo, Gil and Hamon, Olivier (2016), Rediscovering 15+2 Years of Discoveries in Language Resources and Evaluation, Language Resources and Evaluation Journal, 2016, pp. 1-56, ISSN: 1574-0218, doi: 10.1007/s10579-016-9352-9

[27] Mariani, Joseph; Francopoulo, Gil and Paroubek, Patrick (2017), Reuse and Plagiarism in Speech and Natural Language Processing Publications, P. Int J Digit

[27] www.dlib.org/dlib/november15/francopoulo/11francopoulo.html

Libr (2017). doi:10.1007/s00799-017-0211-0Moro Andrea, Raganato Alessandro, Navigli Roberto (2014). Entity Linking meets Word Sense Disambiguation: a Unified Approach, Transactions of the ACL.

[28] Osborne, F.; Motta, E. and Mulholland, P. (2013), Exploring Scholarly Data with Rexplore, International Semantic Web Conference, Sydney, Australia.

[29] Paul, Michael and Roxana Girju (2009) Topic Modeling of Research Fields: An Interdisciplinary Perspective, In Recent Advances in Natural Language Processing (RANLP 2009), Borovets, Bulgaria.

[30] Perin, Charles ; Boy, Jeremy and Vernier, Frédéric (2016), GapChart : a Gap Strategy to Visualize the Temporal Evolution of both Ranks and Scores, IEEE Computer Graphics and Applications, Special issue on Sports Data Visualization, September/October 2016

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

[32] Rochat, Yannick (2009), Closeness centrality extended to unconnected graphs: The harmonic centrality index. Applications of Social Network Analysis (ASNA), 2009, Zurich, Switzerland.

[33] Tang, J.; Zhang, J.; Yao, L.; Li, J.; Zhang, L. and Su, Z. (2008), ArnetMiner: extraction and mining of academic social networks. In: Proceeding of the 14th Int. Conference on Knowledge Discovery and Data Mining .

[34] 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/

[35] Vogel, Adam and Jurafsky, Dan (2012). He said, she said: gender in the ACL anthology. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries* (ACL'12). Association for Computational Linguistics, Stroudsburg, PA, USA, 33-41.

[36] Witten, Ian H.; Eibe, Frank and Hall, Mark A. (2011), Data Mining: practical machine learning tools and techniques. Third Edition. Morgan Kaufmann, Burlington, USA

[37] 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.