Rediscovering 10 to 20 Years of Discoveries in Language & Technology

J. Mariani[1,2], G. Francopoulo[2,3], P. Paroubek[1], Z. Vetulani[4]
[1]LIMSI-CNRS, [2]IMMI-CNRS, [3]Tagmatica, [4]Adam Mickiewicz University, Poznan
E-mail : Joseph.Mariani@limsi.fr, pap@limsi.fr, gil.francopoulo@wanadoo.fr, vetulani@amu.edu.pl

Abstract

This paper analyzes the content of the proceedings of the Language and Technology Conference (L&TC) since its first edition in 1995, with the goal of gaining a picture of the L&TC community and the topics that are most relevant to the field. We follow the methodology used in similar studies, including the survey of the IEEE ICASSP conference proceedings from 1976 to 1990, the survey of Association of Computational Linguistics (ACL) conference proceedings over 50 years, the survey of the proceedings of the conferences contained in the ISCA Archive over 25 years (1987-2012) and the survey of the LREC conference over 16 years (1998-2014). We gathered the NLP4NLP corpus, which contains a large number of papers published by the Speech and Natural Language Processing community in 34 conferences and journals that we use as a reference. The NLP methods used in our analyses have actually been described in papers contained in this corpus, hence the name we gave it. The L&TC conference was first organized in 1995, reactivated in 2005 and it took place every odd year since then. We first verified the quality of the proceedings. We show the evolution over time of the number of papers and authors, the renewal of the authors, their distribution by gender, the continuity of their participation and their productivity, as well as the collaborations among them through the study of the collaboration graph. We then analyze citations of papers and authors, through the study of citations graphs. We also consider the evolution of research topics over time and identify the authors who introduced key terms, as a mark of innovation. Finally, we propose a measure of a researcher's notability based on production, collaboration, citation and innovation, and provide the results for L&TC authors. In addition to providing a revealing characterization of the L&TC community, the study also demonstrates the need for establishing a system for unique identification of authors, papers, and other sources to facilitate this type of analysis. This study may provide insights for future directions of the L&TC, on the occasion of its 20th birthday.

Keywords: Language Technology, Language Resources, Language Processing Systems Evaluation, Text Analytics, Social Networks, Bibliometrics, Scientometrics.

1. Introduction

1.1. Text analytics of scientific paper

The application of text analytics to bodies of scientific papers has become an active area of research in recent years Studies of research publication data mine conference and workshop proceedings to determine trends in publications within a given area or field, such as networks of collaboration and author and paper citation, author/topic pairings, topic shifts over time, and author and participant demographics, with the goal of better understanding research trends, collaborations, participation and publication data, etc. In the field of Speech and Natural Language Processing (SNLP), several studies of this type have recently been conducted, including the following:

- ACL Anthology[1] (D. Radev et al., 2013) analysis, presented in several papers at the Association for Computational Linguistics (ACL) workshop entitled “Rediscovering 50 Years of Discoveries in Natural Language Processing” on the occasion of ACL’s 50th anniversary in 2012[2]. The workshop included the contributions of 23 authors through 13 papers (ACL, 2012).
- Analysis of 25 years of research contained in the International Speech Communication Association (ISCA) Archive[3] (assembled by Wolfgang Hess) published in proceedings of various conferences in the ISCA series (e.g., ECST, Eurospeech, ICSLP, Interspeech) between 1987 and 2012 (J. Mariani et al., 2013).
- Analysis of the publications presented at the biennial Language Resources and Evaluation Conference (LREC) over the past 16 years, from its inception in 1998 through 2014, which was presented on the occasion of LREC 15th anniversary during the Opening session at LREC’2014 (Reykjavik, Iceland) and is based on the LREC Anthology (J. Mariani et al., 2014).

1.2. The L&TC community and conference series analysis

Activity in the area of Language Technology increased enormously over the past 30 years, due to the necessity to process the information conveyed through speech and text, and to allow for a natural interaction between humans and machines. The first Language and technology Conference (L&TC) was held in 1995 in Poznan (Poland) and was organized and chaired by Zygmunt Vetulani, following an incentive from the European Commission. Foreign people such as Antonio Zampolli, Dafydd Gibbon, Jan Roukens, Dan Tufis, Bente Maegaard, participated in this first conference. The second one took place 10 years later, in 2005. Following its success, L&TC has since been held each odd year in Poznan. 2015 is therefore the 20th anniversary of L&TC, or the 10th anniversary if we start from its renewal in 2005.

We will first present here an analysis of the number of papers and the authors over time, including study of their gender; collaboration among authors; the citation among authors and papers; the evolution of topics and those who introduced them. We then propose a measure of a researcher’s notability in the L&TC scientific community based on this analysis.

1.3. The NLP4NLP Speech and Natural Language Processing Analysis

short name	# docs	type	long name	Language	access to content	Period	# venues[4]
acl	4262	conference	Association for Computational Linguistics conference	English	open access*	1979-2014	36
alta	262	conference	Australasian Language Technology Association	English	open access*	2003-2014	12
anlp	329	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	777	journal	American Journal of Computational Linguistics	English	open access*	1980-2014	35
coling	3833	conference	Conference on Computational Linguistics	English	open access*	1965-2014	21
conll	789	conference	Computational Natural Language Learning	English	open access*	1997-2014	17
csal	718	journal	Computer Speech and Language	English	private access	1986-2015	29
eacl	900	conference	European Chapter of the ACL conference	English	open access*	1983-2014	14
emnlp	1708	conference	Empirical methods in natural language processing	English	open access*	1996-2014	19
hlt	2080	conference	Human Language Technology	English	open access*	1986-2013	18
icassps	9023	conference	IEEE International Conference on Acoustics, Speech and Signal Processing - Speech Track	English	private access	1990-2014	25
ijcnlp	899	conference	International Joint Conference on NLP	English	open access*	2005-2013	5
inlg	199	conference	International Conference on Natural Language Generation	English	open access*	1996-2012	6
isca	17592	conference	International Speech Communication Association conferences (ECST, Eurospeech, ICSLP, Interspeech)	English	open access	1987-2014	27
jep	507	conference	Journées d'Etudes sur la Parole	French	open access*	2002-2014	5
lre	276	journal	Language Resources and Evaluation	English	private access	2005-2014	10
lrec	4552	conference	Language Resources and Evaluation Conference	English	open access*	1998-2014	9
ltc	299	conference	Language and Technology Conference	English	private access	2009-2013	3
modulad	232	journal	Le Monde des Utilisateurs de L'Analyse des Données	French	open access	1988-2010	23
muc	149	conference	Message Understanding Conference	English	open access*	1991-1998	5
naacl	1000	conference	North American Chapter of ACL conference	English	open access*	2000-2013	10
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

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

[2] Results of these analyses together with corresponding data and tools are available on-line at the University of Michigan http://clair.eecs.umich.edu/aan/index.php.

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

[4] This is the number of venues where data was obtainable; there may have been other venues in addition.

sem	752	conference	Lexical and Computational Semantics / Semantic Evaluation	English	open access*	2001-2014	7
speechc	549	journal	Speech Communication	English	private access	1982-2015	34
tacl	92	journal	Transactions of the Association of Computational Linguistics	English	open access*	2013-2015	3
tal	156	journal	Revue Traitement Automatique du Langage	French	open access	2006-2013	8
taln	976	conference	Traitement Automatique du Langage Naturel	French	open access*	1997-2014	18
taslp	2659	journal	IEEE/ACM Transactions on Audio, Speech and Language Processing	English	private access	1993-2015	23
tipster	105	conference	Tipster DARPA text program	English	open access*	1993-1998	3
trec	1756	conference	Text Retrieval Conference	English	open access	1992-2014	23
Total	59766					1965-2015	515 506**

Table 1. The NLP4NLP Corpus of Conferences (23) and Journals (9)
(*: included in the ACL Anthology, **: joint conferences are counted once)

We produced a corpus containing research papers on spoken and written language processing, called the NLP4NLP corpus, a name chosen to reflect the fact that the study uses NLP methods that are the subject of the corpus content itself (G. Francopoulo et al., 2015a, G. Francopoulo et al., 2015b). The NLP4NLP corpus contains papers from thirty-two conferences and journals on natural language processing (NLP) and spoken language processing (SLP) published over 50 years (1965-2015) and including the L&TC series (Table 1), thereby providing a good picture of research within the international SNLP community. We included material from conferences and journals only, as workshops may have widely varying ways of reviewing papers. The comparative analysis of the data contained in this corpus is presently ongoing and will be presented in a future paper. In the present paper, we used the entire corpus to study citations to and from L&TC papers: it gives an analysis on how the L&TC community globally considers and is being considered by its general scientific environment.

2. Analysis of the series of L&TC conferences

As a convention, we refer to the conference publication as a *document*. A *paper* or *article* corresponds to a *document* that may have been published in one or several conference series when presented at a joint conference. We refer to individual *authors* and mention their *authorships* or *contributions* to a publication where they act as *contributors*. The same author may sign several papers at a given conference, as a single author or together with one or several co-authors.

2.1. The L&TC conference series

This study covers the series of L&TC conferences, which contains the proceedings of all six L&TC conferences (see Table 2), covering a time span of 18 years (1995-2013).

Year	#Papers	#Authorships	#Authorships/paper
1995	36	49	1.361
2005	105	215	2.048
2007	115	297	2.583
2009	104	261	2.510
2011	107	276	2.579
2013	88	226	2.568
Total	555	1324	2.386

Table 2. *List of conferences with number of papers and of authorships.*

2.2. Data and tools

Over the years, 555 papers have been published in the six L&TC proceedings. All the documents are available in PDF, except the 1995 proceedings, which are only available on paper, and in Polish. We used for this first conference a translation of the titles and a short abstract of the content in English. Following the publication in the proceedings, a selection of revised papers was published as a book, in the Archives of Control Sciences for L&TC 2005, in the Lecture Notes on Artificial Intelligence (Springer) for the subsequent ones.

A benchmark to estimate the error rate of the extracted content was established based on a simple heuristics, which is that "rubbish" character strings are not entries in lexicons. This estimation is computed as the number of unknown words divided by the number of words. The number of errors was computed from the result of the morphological module of TagParser (G. Francopoulo, 2007), a deep industrial parser based on a broad English lexicon and Global Atlas (a knowledge base containing more than one million words from 18 Wikipedias) (G. Francopoulo, 2013). Variations in performance quality measures were used to control the parameterization of the content preprocessing tools.

Following this content extraction, another step in our preprocessing was dedicated to split the content into abstract, body and references sections. We created a small set of rules in Java to extract the abstract and body of the papers and compute their quality.

The result of the preprocessing is summarized in the following table, and it can be noticed that the corpus contains about 1.8 million words, and that the overall quality is good (better than 98%).

year	nb of papers from the metadata	nb of papers in PDF	nb of papers in XML (= output of PDFBox)	nb of non empty papers as extraction result	nb of papers with an abstract (from extraction)	nb of papers with references (from extraction)	nb of unknown words	nb of known words	nb of words of the content	evaluation of noise = pourcentag e of nb of known words / nb of words of the content	evaluation of silence = pourcentag e of non empty papers as extraction result / PDF docs	combined evaluation of noise and silence	nb of English papers	nb of French papers	nb of papers in another language (es+de+ru)
1995	36	36	36	36	0	0	19	802	821	97.686	100.000	98.829	36	0	0
2005	105	105	103	103	84	101	8565	339893	348458	97.542	98.095	97.818	103	0	0
2007	115	115	115	115	98	111	11395	404380	415775	97.259	100.000	98.611	115	0	0
2009	104	104	103	103	71	82	7141	319211	326352	97.812	99.038	98.421	103	0	0
2011	107	107	107	107	91	106	11798	398968	410766	97.128	100.000	98.543	107	0	0
2013	88	88	86	86	70	81	10012	288615	298627	96.647	97.727	97.184	86	0	0
total	555	555	550	550	414	481	48930	1751869	1800799	97.283	99.099	98.183	550	0	0

Table 3. *Quality of the preprocessing*

2.3. Overall analysis: papers and authors

The study of authors is problematic due to variations of the same name (family name and given name, initials, middle initials, ordering, married name, etc.). It therefore required a tedious semi-automatic cleaning process (J. Mariani et al., 2014b). This suggests a need to determine ways to uniquely identify researchers.

The total number of papers published in the conference series is 555 (Table 2). The number of authorships is more than 1,300. Those numbers increase almost linearly over time (Fig. 1).

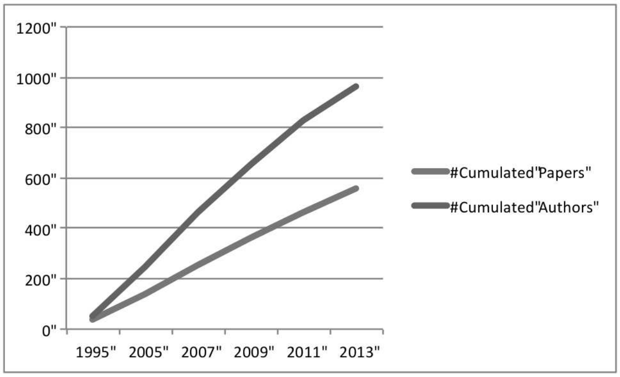


Figure 1. *Number of papers and authorships over time*

The average number of co-authors per paper increased over time, from 1.5 in 1995 up to 2.5 in 2013 (i.e. one more co-author on average) (Fig. 2). 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. The largest number of co-authors for a paper is 12, in a paper published at L&TC 2011.

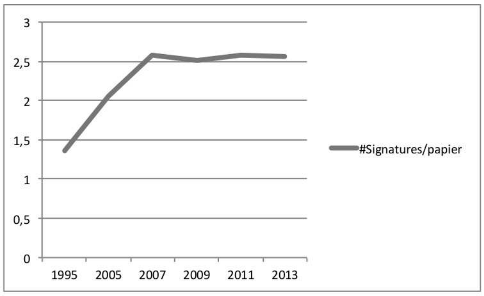


Figure 2. *Average number of authors per paper*

We studied the number of repeat authors at successive conferences (Table 4). For each conference, we identified the authors who did not publish at the previous conference (*new authors*). We also studied those who had not published at any previous L&TC conference (*completely new authors*).

Year	#New authors	#Different Authors	#New Authors/#authors	#Completely new authors	#Completely New Authors/#Authors
1995	48	48	1.000	48	1.000
2005	195	200	0.975	195	0.975
2007	220	263	0.837	217	0.825
2009	199	241	0.826	189	0.784
2011	194	247	0.785	176	0.713
2013	159	204	0.779	134	0.657
Total				959	

Table 4. *Author renewal and redundancy*

We then studied the authors' renewal. It clearly showed (Fig. 3) that the ratio of the different authors between one conference and the next, and the ratio of authors who never published in L&TC beforehand stay very high over time (resp. 80% and 70%), showing a regular participation of fresh blood.

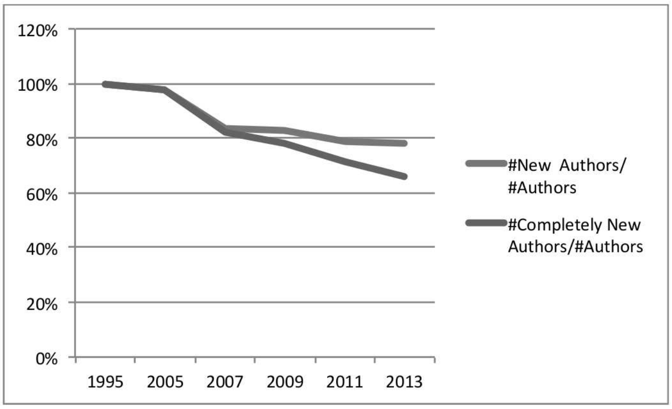


Figure 3. *Percentage of new authors and completely new authors over time.*

Author gender. An author gender study was performed with the help of a lexicon of 27,509 given names with gender information (66% male, 31% female, 3% epicene[5]). 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.) (Yu Fu et al., 2010), 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. In some cases, we only had an initial for the first name, which made gender guessing impossible unless the same

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

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 six conferences shows that 62% of the authors are male, while 22% of the authors are female, 2% are of indeterminate gender, and 14% are of unknown gender. If we assume that the authors of indeterminate and unknown gender have the same gender distribution as the ones that are categorized, male authors account for 74% and female authors for 26%, compared with 70%/30% for LREC and 80%/20% for ACL and ISCA (Fig. 4).

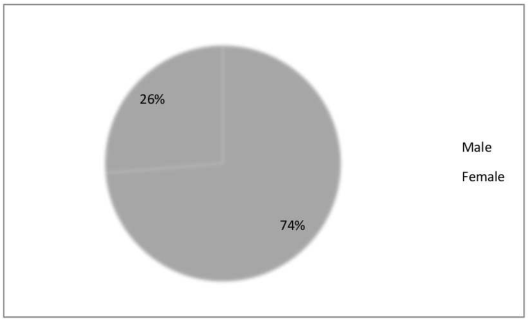


Figure 4. *Authors gende*r

Author production. Eight authors published in all five conferences, if we exclude 1995 (Fumiyo Fukumoto, Filip Graliński, Cvetana Krstev, Yves Lepage, Jacek Marciniak, Yoshimi Suzuki, Zygmunt Vetulani, Duško Vitas). About 800 authors (more than 80% of the 959 authors) published at a single conference (Fig. 5).

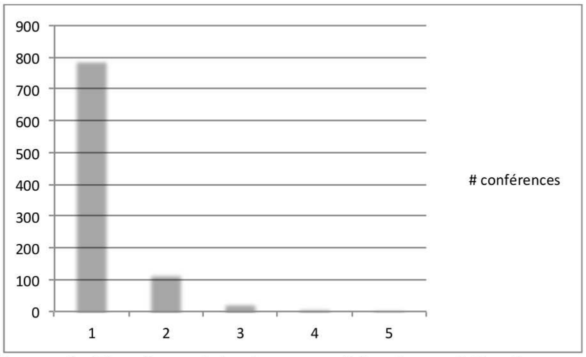


Figure 5. *Number of Authors per Number of Conferences*

The most productive author published 12 papers, while about 750 authors (78% of the 959 authors) published only one paper (Fig. 6). The author who published the largest number of papers as single author is Elżbieta Hajnicz, while 416 authors (43% of the authors) never published a paper as single author.

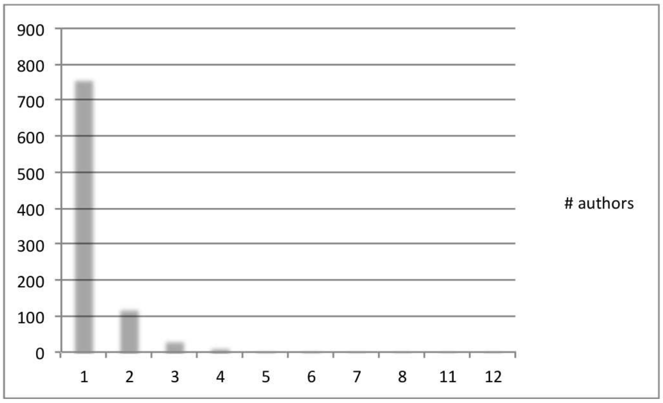


Figure 14. *Number of Papers per Number of Authors*

Table 5 gives the list of the 10 most productive authors, with number of papers they published.

Yves Lepage	12
Yoshimi Suzuki	12
Fumiyo Fukumoto	11
Zygmunt Vetulani	9
Adam Przepiórkowski	8
Duško Vitas	7
Krzysztof Jassem	6
Filip Graliński	6
Jacek Marciniak	6
Cvetana Krstev	6

Table 5. *10 most productive authors*

2.4. Collaborations

The most collaborating authors published with 15 different co-authors, while close to 100 authors always published alone (Fig. 7). Six authors published with 13 or more different co-authors (Table 6).

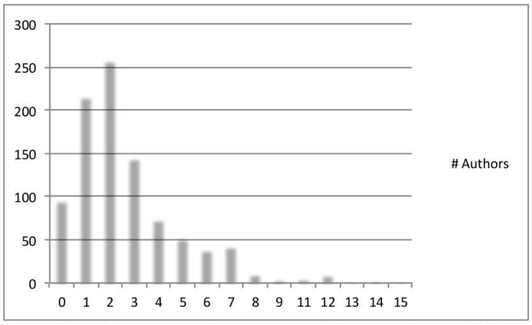


Figure 7. *Number of authors as a function of the number of different co-authors*

Justyna Walkowska	15
Zygmunt Vetulani	15
Marián Trnka	14
Milan Rusko	14
Tomasz Obrębski	13
Yves Lepage	13

Table 6. *The 6 authors with the largest number of co-authors*

A *collaboration graph*[6] (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). The CollG need not be fully connected, that is, people who never co-authored a joint paper are represented by isolated nodes. Those who are connected constitute a *connected component*. *Cliques* are fully connected components where all authors published with one another. The *collaboration distance* is the geodesic distance, or path-length, between two nodes in a CollG, which is equal to the smallest number of edges in an edge-path, or *collaboration path*, connecting them. The *diameter* of the CollG is the longest collaboration path in that graph. If no path connecting two nodes in a CollG exists, the collaboration distance between them is considered to be infinite. The *degree* of a node (number of edges attached to the node) reflects the number of co-authors

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

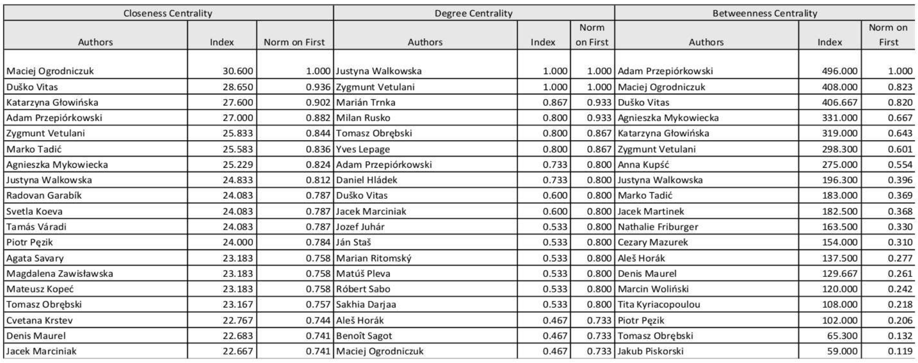
associated with each author, as an absolute measure of his/her collaboration ability. The *clustering coefficient* of a node is a measure of the degree to which its neighboring nodes tend to cluster together: i.e. how close they are to form a clique. The *density* of a graph is the fraction of all possible edges that actually exist in the CollG, thus providing a measure of the density of collaboration: if all authors have published at least one paper with all the other authors, the density of collaboration of the graph would be equal to 1.

The L&TC CollG contains 959 nodes corresponding to the 959 different authors who have published at L&TC.

The largest connected component groups 62 authors, which means that 6% of the 959 L&TC authors are connected through a collaboration path. The authors of the largest connected component published 41 papers (7% of the total number of papers). The second largest connected component has 34 authors.

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 (L. Freeman, 1978). The *Closeness distance* has been introduced in Human Sciences to measure the efficiency of a Communication Network (A. Bavelas, 1948 and A. Bavelas, 1950). 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, 2009) 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 7, 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		
Authors	Index	Norm on First	Authors	Index	Norm on First	Authors	Index	Norm on First
Maciej Ogrodniczuk	30.600	1.000	Justyna Walkowska	1.000	1.000	Adam Przepiórkowski	496.000	1.000
Duško Vitas	28.650	0.936	Zygmunt Vetulani	1.000	1.000	Maciej Ogrodniczuk	408.000	0.823
Katarzyna Głowińska	27.600	0.902	Marián Trnka	0.867	0.933	Duško Vitas	406.667	0.820
Adam Przepiórkowski	27.000	0.882	Milan Rusko	0.800	0.933	Agnieszka Mykowiecka	331.000	0.667
Zygmunt Vetulani	25.833	0.844	Tomasz Obrębski	0.800	0.867	Katarzyna Głowińska	319.000	0.643
Marko Tadić	25.583	0.836	Yves Lepage	0.800	0.867	Zygmunt Vetulani	298.300	0.601
Agnieszka Mykowiecka	25.229	0.824	Adam Przepiórkowski	0.733	0.800	Anna Kupść	275.000	0.554
Justyna Walkowska	24.833	0.812	Daniel Hládek	0.733	0.800	Justyna Walkowska	196.300	0.396
Radovan Garabík	24.083	0.787	Duško Vitas	0.600	0.800	Marko Tadić	183.000	0.369
Svetla Koeva	24.083	0.787	Jacek Marciniak	0.600	0.800	Jacek Martinek	182.500	0.368
Tamás Váradi	24.083	0.787	Jozef Juhár	0.533	0.800	Nathalie Friburger	163.500	0.330
Piotr Pęzik	24.000	0.784	Ján Staš	0.533	0.800	Cezary Mazurek	154.000	0.310
Agata Savary	23.183	0.758	Marian Ritomský	0.533	0.800	Aleš Horák	137.500	0.277
Magdalena Zawisławska	23.183	0.758	Matúš Pleva	0.533	0.800	Denis Maurel	129.667	0.261
Mateusz Kopeć	23.183	0.758	Róbert Sabo	0.533	0.800	Marcin Woliński	120.000	0.242
Tomasz Obrębski	23.167	0.757	Sakhia Darjaa	0.533	0.800	Tita Kyriacopoulou	108.000	0.218
Cvetana Krstev	22.767	0.744	Aleš Horák	0.467	0.733	Piotr Pęzik	102.000	0.206
Denis Maurel	22.683	0.741	Benoît Sagot	0.467	0.733	Tomasz Obrębski	65.300	0.132
Jacek Marciniak	22.667	0.741	Maciej Ogrodniczuk	0.467	0.733	Jakub Piskorski	59.000	0.119

Table 7. *Computation and comparison of the Closeness Centrality, Degree Centrality and Betweenness Centrality for the 10 most central author .*

2.5. Citations

Unlike the CollG, a *citation graph* (CitG) is directed. 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.

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 this cases, corresponding authors or papers appear as isolated nodes in the citation graphs. The nodes that are connected through a directed path (Author A cites Author B and Author B cites Author C and Author C cites Author A, for example), 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, (for example, 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).

The *citation distance* between two nodes is the smallest number of directed edges in an edge-path connecting them. The *diameter* of a citation graph is the longest path in the graph, which will be identical in both the citing and cited graphs. If no path connecting two nodes in a citation graph exists, the citation distance between them is said to be infinite. In a citing graph, the degree of a node (the number of directed edges issued from that node) reflects the absolute number of authors (or papers) cited by each author (or paper). In a cited graph, the degree of a node reflects the absolute number of authors (or papers) citing each author (or paper). As in the CollG, the *clustering coefficient* of a node is a measure of the degree to which its neighbors tend to cluster together. The *density* of a citation graph, which is the fraction of possible edges that exist in the graph, provides a measure of the density of citation: if all authors (or papers) cite at least once each other author (or paper), the density of citation of the graph would be equal to 1.

We studied citations in papers from 2005 to 2013. 481 of the 555 papers do not contain a list of references. We studied the four Citing and Cited Authors and Papers Graphs, using the L&TC conference series to represent the L&TC community and the NLP4NLP corpus[7], which also includes L&TC, to represent the general Speech and Natural Language Processing scientific community (SNLP).

We studied:

- the citation in L&TC papers of other L&TC papers (*Internal Papers Citations*: the citations within L&TC),
- the citation in L&TC papers of NLP4NLP papers (*Outgoing Global Papers Citations:* how L&TC cites its scientific environment),
- the citation in NLP4NLP papers of L&TC papers (*Ingoing Global Papers Citations:* how L&TC is being cited by its scientific environment).

Similarly, we also studied:

- the citation by L&TC authors of L&TC authors (*Internal Authors Citations*),
- the citation by L&TC authors of SNLP authors (*Outgoing Global Authors Citations*),
- the citation by SNLP authors of L&TC authors (*Ingoing Global Authors Citations*:).

2.5.1. Authors citations

We first consider *internal authors citations*: the citation by authors in their L&TC papers of authors for their L&TC papers.

Internal renown of L&TC authors (CdAG): Table 8 gives the list of the 10 most cited L&TC authors in L&TC papers, with the number of citations.

[7] See Table 1

Adam%Przepiórkowski	15
Jacek%Marciniak	11
Justyna%Walkowska	11
Zygmunt%Vetulani	11
Tomasz%Obrębski	10
Barbara%LewandowskaF Tomaszczyk	8
Marek%Łaziński	8
Mirosław%Bańko	8
Piotr%Pęzik	8
Rafał%%Górski	8

Table 8. *10 most cited L&TC authors in L&TC papers*

We now consider *global authors citations*: citation by L&TC authors of SNLP authors and by SNLP authors of L&TC authors.

Global renown of L&TC authors: Table 9 gives the list of the 10 most cited L&TC authors in NLP4NLP papers.

Adam Przepiórkowski	57
Barbara Lewandowska-Tomaszczyk	39
Marek Łaziński	39
Mirosław Bańko	39
Piotr Pęzik	39
Rafał L Górski	39
Benoît Sagot	35
Eric De La Clergerie	16
Zygmunt Vetulani	15
Jacek Marciniak	13

Table 9. *10 most cited L&TC authors in* NLP4NLP *papers*

Global renown of authors in L&TC papers: Table 10 gives the list of the 10 most cited SNLP authors.

Philipp Koehn	36
Adam Przepiórkowski	34
Franz Josef Och	26
Hermann Ney	22
Andreas Stolcke	19
Marek Łaziński	19
Rafał L Górski	19
Tomaž Erjavec	17
Christopher D Manning	16
Martha Palmer	15

Table 10. *10 most cited SNLP authors in L&TC papers*

2.5.2. Papers citations

Here also, we first consider *internal papers citations*: the citation in L&TC papers of L&TC papers.

Internal renown of L&TC papers (CdPG): Table 11 gives the list of the 10 most cited L&TC papers in L&TC papers, with the list of authors, the title and the number of citations.

Adam Przepiórkowski, Mirosław Bańko, Rafał L Górski, Barbara Lewandowska-Tomaszczyk, Marek Łaziński, Piotr Pęzik	National Corpus of Polish	8
Zygmunt Vetulani, Justyna Walkowska, Tomasz Obrębski, Pawel Konieczka, Przemyslaw Rzepecki, Jacek Marciniak	PolNet - Polish WordNet project algorithm	4
Karel Pala, Aleš Horák, Adam Rambousek, Zygmunt Vetulani, Pawel Konieczka, Jacek Marciniak, Tomasz Obrębski, Przemyslaw Rzepecki, Justyna Walkowska	DEB Platform tools for effective development of WordNets in application to PolNet	3
Zygmunt Vetulani, Jacek Marciniak, Tomasz Obrębski, Marek Kubis, Jędrzej Osiński, Justyna Walkowska, Piotr Kubacki, Krzysztof Witalewski	POLINT-112-SMS: Beta Prototype	3
Jakub Fast, Adam Przepiórkowski	Automatic Extraction of Polish Verb Subcategorization An Evaluation of Common Statistics	2
Marcin Woliński	An efficient implementation of a large grammar of Polish	2
Adam Przepiórkowski, Piotr Bański	Which XML standards for multilevel corpus annotation?	2
Rafal Mlodzki, Adam Przepiórkowski	The WSD Development Environment	2
Marek Kubis	An access layer to PolNet in POLINT-112-SMS	2
Lars Hellan, Mary Esther, Kropp Dakubu	A methodology for enhancing argument structure specification	2

Table 11. *The 10 L&TC papers most cited by other L&TC papers*

We now consider *global papers citations*: citation in L&TC papers of NLP4NLP papers and of L&TC papers in NLP4NLP papers.

Global renown of L&TC papers: Table 12 gives the list of the 10 most cited L&TC papers in NLP4NLP papers.

Adam Przepiórkowski, Mirosław Bańko, Rafał L Górski, Barbara Lewandowska-Tomaszczyk, Marek Łaziński, Piotr Pęzik	National Corpus of Polish	39
Benoît Sagot, Pierre Boullier	From raw corpus to word lattices: robust pre-parsing processing	12
Paweł Mazur, Robert Dale	The DANTE Temporal Expression Tagger	10
Claire Gardent, Bruno Guillaume, Guy Perrier, Ingrid Falk	Maurice Gross' grammar lexicon and Natural Language Processing	9
Eric De La Clergerie, Lionel Clément	MAF: a Morphosyntactic Annotation Framework	9
Tomža Erjavec, Camelia Ignat, Bruno Pouliquen, Ralf Steinberger	Massive multi lingual corpus compilation: Acquis Communautaire and totale	8
Kais Dukes	Semantic Annotation of Robotic Spatial Commands	8
Cvetana Krstev, Dusko Vitasz, Denis Maurel, Mickaël Tran	Multilingual Ontology of Proper Names	7
Benoît Sagot	Building a morphosyntactic lexicon and a pre-syntactic processing chain for Polish	7
Caroline Brun, Maud Ehrmann, Guillaume Jacquet	A Hybrid System for Named Entity Metonymy Resolution	7

Table 12. *The 10 L&TC most cited papers in NLP4NLP papers*

Global renown of papers in L&TC papers. Table 13 gives the list of the 10 most cited NLP4NLP papers in L&TC papers, with the conference or journal where they have been published. It includes one L&TC paper.

Philipp Koehn	Europarl: A Parallel Corpus for Statistical Machine Translation	MT Summit 2005	12
Franz Josef Och, Hermann Ney	A Systematic Comparison of Various Statistical Alignment Models	Computational Linguistics 2003	11
Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Christopher Dyer, Ondřej Bojar, Alexandra Constantin, Evan Herbst	Moses: Open Source Toolkit for Statistical Machine Translation	ACL 2007	10
Andreas Stolcke	SRILM - an extensible language modeling toolkit	Interspeech 2002	10
Dekang Lin	Automatic Retrieval and Clustering of Similar Words	ACL 1998	9
Adam Przepiórkowski, Mirosław Bańko, Rafał L Górski, Barbara Lewandowska-Tomaszczyk, Marek Laziński, Piotr Pęzik	National Corpus of Polish	L&TC 2011	8
Peter E Brown, Stephen A Della Pietra, Vincent J Della Pietra, Robert L Mercer	The Mathematics of Statistical Machine Translation: Parameter Estimation	Computational Linguistics 1993	7
Philipp Koehn, Franz Josef Och, Daniel Marcu	Statistical Phrase-Based Translation	NAACL 2003	7
Mitchell P Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz	Building a Large Annotated Corpus of English: The Penn Treebank	Computational Linguistics 1993	6
Adam Przepiórkowski, Rafał L Górski, Marek Laziński, Piotr Pęzik	Recent Developments in the National Corpus of Polish	LREC 2010	6

Table 13. *The 10 NLP4NLP most cited papers in L&TC papers*

2.6. Topics

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. Like the study of citations, our initial input is the textual content of the papers available in a digital format apart from the proceedings of 1995 (519 documents). Over these 18 years, the archives contain a grand total of about 1,800,000 words, as shown in Table 3.

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 (P. Drouin, 2004). 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[8] 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)[9], the Open American National Corpus (aka OANC[10]), the Suzanne corpus release-5[11], the English EuroParl archives (years 1999 until 2009)[12], 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 L&TC content with the same filters and used our language model to distinguish L&TC-specific terms from common ones. We worked from the hypothesis that when a sequence of words is *inside* the Anthology and *not inside* the general language profile, the term is specific to the field of language resources and evaluation. The 1,800,799 word content in 519 documents include 56,923 different terms (unigrams, bigrams and trigrams) and 150,529 term occurrences, provided that this number counts all the occurrences of all the sizes and does not restrict to the longest terms, thus counting a great number of overlapping situations between fragments of texts.

The twenty most frequent terms in the field of language resources and evaluation were computed over the period of 8 years, according to the following strategy. First, the most frequent terms were computed in a raw manner, and secondly the synonyms sets (aka synsets) for all most 50 frequent terms of each year (which are frequently the same from one year to another) were manually declared in the lexicon of TagParser. Around the term synset, we gathered the variation in upper/lower case, singular/plural number, US/UK difference, abbreviation/expanded form and absence/presence of a semantically neutral adjective, like "artificial" in "artificial neural network". Thirdly, the most frequent terms were recomputed with the amended lexicon. The 20 most frequent terms over time (2005-2013) are the following (Table 14):

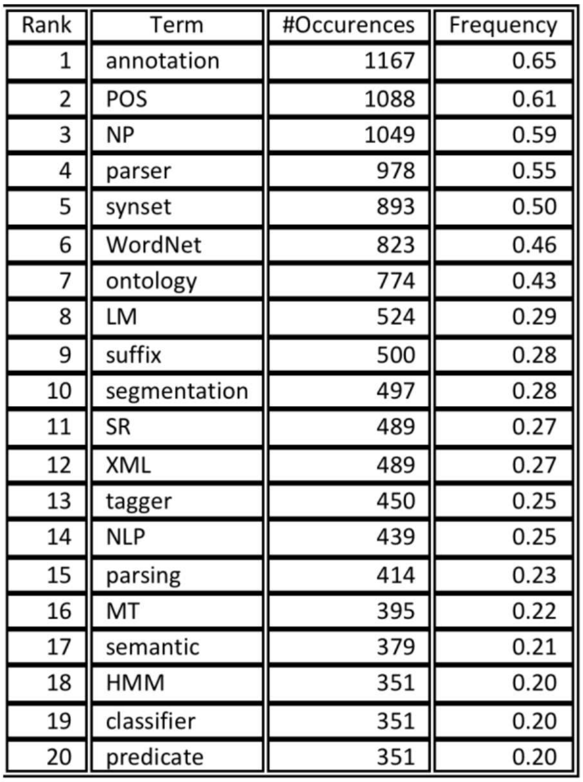
[8] www.tagmatica.com

[9] www.natcorp.ox.ac.uk

[10] www.americannationalcorpus.org

[11] www.grsampson.net/Resources.html

[12] www.statmt.org/europarl



Rank	Term	#Occurences	Frequency
1	annotation	1167	0.65
2	POS	1088	0.61
3	NP	1049	0.59
4	parser	978	0.55
5	synset	893	0.50
6	WordNet	823	0.46
7	ontology	774	0.43
8	LM	524	0.29
9	suffix	500	0.28
10	segmentation	497	0.28
11	SR	489	0.27
12	XML	489	0.27
13	tagger	450	0.25
14	NLP	439	0.25
15	parsing	414	0.23
16	MT	395	0.22
17	semantic	379	0.21
18	HMM	351	0.20
19	classifier	351	0.20
20	predicate	351	0.20

Table 14. *20 most frequent terms overall*

2.6.1. Change in Topics.

We studied the ranking among the 50 most popular terms (mixing unigrams, bigrams and trigrams) representing several topics of interest. We first studied the following terms, which stayed in the top 20 over 18 years: *Annotation*, *Ontology, Parser*, *Synset, Wordnet* and *Part Of Speech (POS)* (Fig. 8).

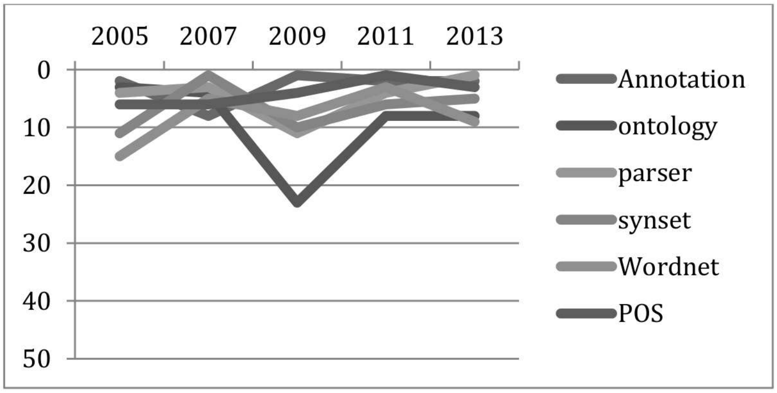


Figure 8. *Terms remaining popular*

We studied several terms that became more popular over time: *Machine Translation (MT), Language Model (LM), dataset,* and, more recently, *Named Entity (NE)* and *Polarity* (Fig. 9).

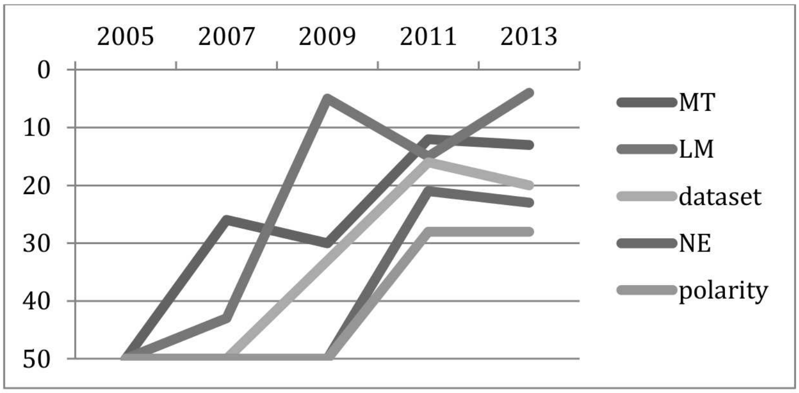


Figure 9. *Terms becoming popular*

We also studied terms that ad momentary success over time: *MSegmentation, Speech recognition (SR) and dialog* (Fig. 10).

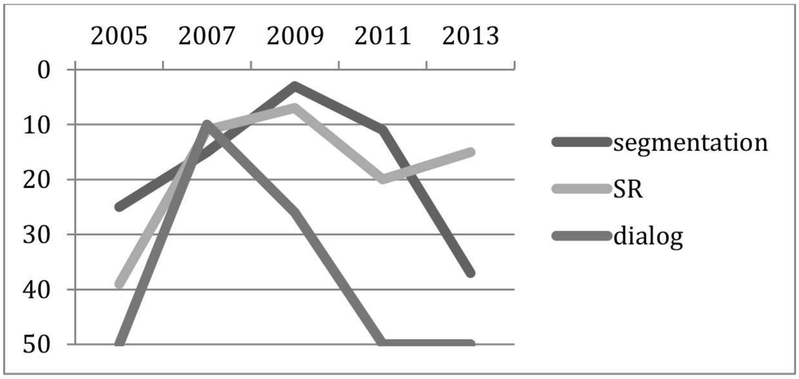
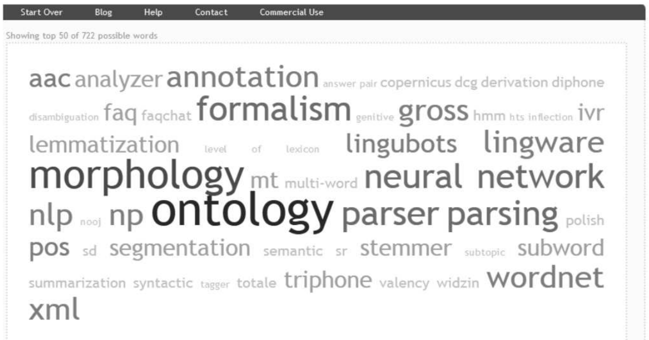


Figure 10. Terms with momentary success

2.6.2. Tag Clouds for frequent terms.

The aim of this section is to provide a global estimation of the main terms used in specific years as well as an indication of the stability of the terms over the years. For this purpose, we use TagCrowd[13] to generate tag clouds[14]. Figures 11-12 show the tag clouds for L&TC 2005 and 2013.


Start Over
Blog
Help
Contact
Commercial Use

Showing top 50 of 722 possible words

aac analyzer annotation answer pair copernicus dcg derivation diphone
disambiguation faq faqchat formalism genitive gross hmm hts inflection ivr
lemmatization level of lexicon lingubots lingware
morphology mt multi-word neural network
nlp nooj np ontology parser parsing polish
pos sd segmentation semantic sr stemmer subtopic subword
summarization syntactic tagger totale triphone valency widzin wordnet
xml

Figure 11. *Tag Cloud based on the 2005 abstracts*

[13] www.tagcrowd.com. Our thanks to Daniel Steinbock for providing access to this web service.

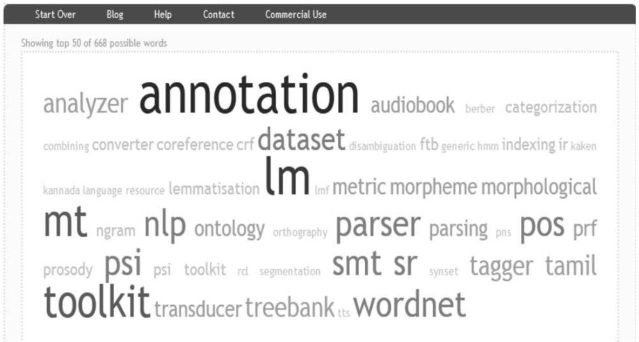
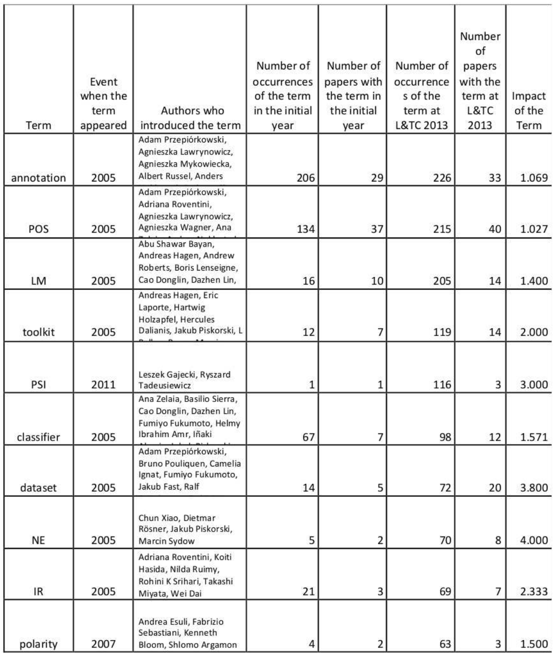

Start Over
Blog
Help
Contact
Commercial Use
Showing top 50 of 668 possible words
analyzer annotation audiobook berber categorization
combining converter coreference crf dataset disambiguation ftb generic hmm indexing ir kaken
kannada language resource lemmatisation lm lmf metric morpheme morphological
mt ngram nlp ontology orthography parser parsing pns pos prf
prosody psi psi toolkit rd segmentation smt sr synset tagger tamil
toolkit transducer treebank tts wordnet

Figure 12. *Tag Cloud based on the 2013 abstracts*

Globally, it appears that most frequent terms remained constant across the years, such as *Annotation* or *Wordnet*. *AAC (Acoustic Audio Coding)*, *Formalism*, *lingubots*, *lingware*, *stemmer*, *subword* or *triphone* disappeared, while *audiobook*, *categorization*, *dataset*, *language model*, *Statistical Machine Translation*, *toolkit*, *transducer* or t*reebank* went to the forefront. We may also notice the presence of less-resourced languages such as *berber*, *kannada* or *tamil*, due to the Less-Resourced Languages special session which is organized within L&TC since 2009.

2.6.3. New terms introduced by the authors.

We studied when and who introduced new terms, as a mark of the innovative ability of various authors, which may also provide an estimate of their contribution to the advances of the scientific domain. We make the hypothesis that an innovation is induced by the introduction of a term which was previously unused in the community and then became popular. We consider the 555 documents and the 959 authors who used the 56,923 terms contained in those documents. We consider the first proceedings (1995) as the starting point for the introduction of new terms. We then take into account the terms which are present in 2013 but not in 1995, and which are of scientific interest (excluding author's names, unless they correspond to a specific algorithm or method, city names, laboratory names, etc.). For each of these terms, starting from the second L&TC (L&TC 2005) proceedings, we determine the author(s) who introduced the term, referred to as the "inventor(s)" of the term. This may yield several names, as the papers could be co-authored or the term could be mentioned in more than one paper in a given year. We compute the *overall impact factor* (OIF) of a term as the ratio between the number of papers mentioning it (its "presence" in papers) in 2014 and the number of papers that mentioned it in the year when it first appeared.



Term	Event when the term appeared	Authors who introduced the term	Number of occurrences of the term in the initial year	Number of papers with the term in the initial year	Number of occurrence s of the term at L&TC 2013	Number of papers with the term at L&TC 2013	Impact of the Term
annotation	2005	Adam Przepiórkowski, Agnieszka Lawrynowicz, Agnieszka Mykowiecka, Albert Russel, Anders	206	29	226	33	1.069
POS	2005	Adam Przepiórkowski, Adriana Roventini, Agnieszka Lawrynowicz, Agnieszka Wagner, Ana	134	37	215	40	1.027
LM	2005	Abu Shawar Bayan, Andreas Hagen, Andrew Roberts, Boris Lenseigne, Cao Donglin, Dazhen Lin,	16	10	205	14	1.400
toolkit	2005	Andreas Hagen, Eric Laporte, Hartwig Holzapfel, Hercules Dalianis, Jakub Piskorski, L	12	7	119	14	2.000
PSI	2011	Leszek Gajecki, Ryszard Tadeusiewicz	1	1	116	3	3.000
classifier	2005	Ana Zelaia, Basilio Sierra, Cao Donglin, Dazhen Lin, Fumiyo Fukumoto, Helmy Ibrahim Amr, Iñaki	67	7	98	12	1.571
dataset	2005	Adam Przepiórkowski, Bruno Pouliquen, Camelia Ignat, Fumiyo Fukumoto, Jakub Fast, Ralf	14	5	72	20	3.800
NE	2005	Chun Xiao, Dietmar Rösner, Jakub Piskorski, Marcin Sydow	5	2	70	8	4.000
IR	2005	Adriana Roventini, Koiti Hasida, Nilda Ruimy, Rohini K Srihari, Takashi Miyata, Wei Dai	21	3	69	7	2.333
polarity	2007	Andrea Esuli, Fabrizio Sebastiani, Kenneth Bloom, Shlomo Argamon	4	2	63	3	1.500

Table 15. *List of the 10 most popular terms at L&TC 2013 ranked according to the greatest presence in papers: date of introduction, authors and Overall Impact Factor.*

Table 15 provides the ranked list of the 10 most popular terms based on the occurrence of the term in 2013. For example, the term *Named Entity (NE)* appeared first in the year 2005, when it was mentioned five times in two papers. In 2014, *NE* was mentioned 70 times in 8 papers, yielding an OIF of 8/2=4. Some terms, such as *Annotation*, were already widely used in 2005, and therefore get a low OIF.

From this analysis, we compute an *innovation score* for each author, illustrating his or her ability to introduce new terms that subsequently became popular, obtained as follows: for each term, we first compute the percentage of papers that contain the term at each conference ("relative presence" of the term) (Fig. 13). We only consider papers written by authors that are different from those who "invented" the term, in order to avoid self citation, i.e. an excessive weight for the overuse of non-propagated terms, typically program or system names.

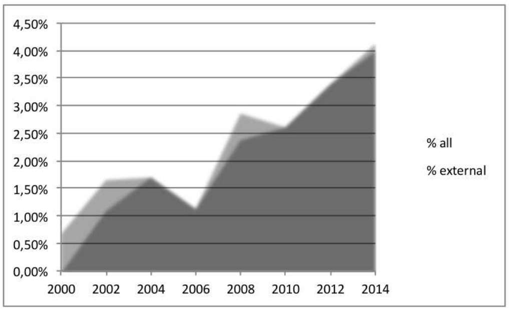


Figure 13. *Relative presence of a term over the years, considering either "all" papers or only those written by authors who are different than those who introduced the term ("external papers").*

The total innovation score of a term is the corresponding surface, taking into account the inventors' papers in the year of introduction and the external papers in the subsequent years (Fig. 14). The innovation score is the sum of the yearly relative presences of the term. Some non-scientific terms may not have been filtered out, but their influence will be small as their presence is limited, while terms that became popular at some point in the past but lost popularity afterwards will remain in consideration.

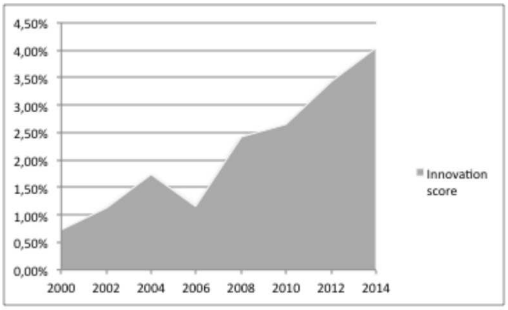
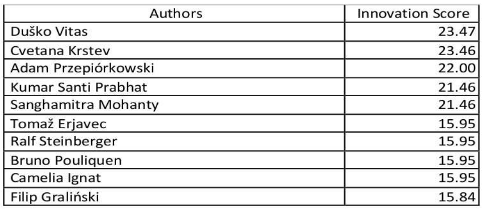


Fig. 14 *Innovation Score of a term*

The innovation score for an author is the sum of the innovation scores of the terms he/she invented (Table 16).



Authors	Innovation Score
Duško Vitas	23.47
Cvetana Krstev	23.46
Adam Przepiórkowski	22.00
Kumar Santi Prabhat	21.46
Sanghamitra Mohanty	21.46
Tomaž Erjavec	15.95
Ralf Steinberger	15.95
Bruno Pouliquen	15.95
Camelia Ignat	15.95
Filip Graliński	15.84

Table 16. *10 most innovative authors according to the introduction of new terms.*

2.7. A Composite Hybrid Measure of Authors Notability

The study of social networks often uses the collaboration network centrality measures that we described above. As already mentioned, centrality measures reflect different roles of the authors and do not take into account several important criteria, in particular the productivity of the authors (number of published papers), their audience (number of citations), and their ability to introduce novelty in research.

Authors	Notability		collaboration		production		citation		innovation	
	rank	Norm on first	rank	Norm on first	rank	Norm on first	rank	Norm on first	rank	Norm on first
Adam Przepiórkowski	1	1.0	4	0.882	5	0.667	1	1.000	3	0.936
Zygmunt Vetulani	2	0.8	5	0.844	4	0.750	2	0.733	47	0.506
Duško Vitas	3	0.7	2	0.936	6	0.583	28	0.067	1	1.000
Tomasz Obrębski	3	0.7	16	0.757	14	0.417	5	0.667	46	0.512
Cvetana Krstev	3	0.7	17	0.744	7	0.500	28	0.067	2	1.000
Jacek Marciniak	6	0.6	19	0.741	7	0.500	2	0.733	207	0.025
Justyna Walkowska	6	0.6	8	0.812	14	0.417	2	0.733	205	0.026
Marcin Woliński	8	0.5	32	0.664	14	0.417	15	0.200	25	0.590
Aleš Horák	8	0.5	20	0.735	21	0.333	14	0.267	69	0.461
Yves Lepage	8	0.5	99	0.425	1	1.000	28	0.067	109	0.293
Filip Graliński	8	0.5	54	0.596	7	0.500	93	0.000	10	0.675
Maciej Ogrodniczuk	8	0.5	1	1.000	14	0.417	28	0.067	117	0.259
Yoshimi Suzuki	8	0.5	428	0.098	1	1.000	93	0.000	15	0.629
Fumiyo Fukumoto	8	0.5	428	0.098	3	0.917	93	0.000	17	0.628
Denis Maurel	8	0.5	18	0.741	38	0.250	28	0.067	34	0.540
Marek Swidziński	8	0.5	43	0.631	38	0.250	21	0.133	27	0.580
Benoît Sagot	8	0.5	94	0.452	7	0.500	28	0.067	30	0.549

Table 17. *17 most notable authors in the L&TC community according to a composition of 4 criteria (Collaboration (closeness centrality), Production, Citation and Innovation).*

We therefore propose (Table 17) as a measure of notability a Composite Hybrid Measure based on the arithmetic mean of the normalized ranking of authors according to those four criteria: Collaboration (see Table 7), Production (see Table 5), Citation (see Table 18) and Innovation (see Table 16). Given the approximations in the various measures we use, we clustered the ranking. It is followed by a large list of authors with a notability score of 0.4. This ranking is not intended to be a hit parade of the "best" L&TC authors, but is rather intended to provide a picture of the L&TC ecosystem and acknowledge the contributions of the members of its community, while stressing that those contributions may have various aspects.

3. Future Work

Our next step is now to conduct a study of the whole NLP4NLP corpus, with a comparison across the various conferences and journals it contains over a 50-year time scale (1965-2015). We plan to investigate more deeply the structure of the corresponding research community through the graph of collaboration and the graph of citations among authors, as a social network. This process will help identifying factions of people who publish together or cite each other. We will also refine the study of the polarity of the citations, extend the bottom up term analysis already begun, and deepen the potential detection of weak signals and emerging trends. Establishing links among authors, citations and topics will allow us to study the changes in the topics of interest for authors or factions.

We will also study the mutual influence of the conferences and journals, and their respective contribution in the advances of the research field, while identifying possible cultural differences among them. We plan to consider the relationship among language resources, as registered in the LRE Map (N. Calzolari et al., 2012), and scientific papers. Researchers in other disciplines, e.g. biology (E. Bravo et al., 2015), face the same problems as in speech and NLP, such as identifying resources in a persistent and unique way, computing Resource Impact Factors, etc. Therefore different scientific communities could benefit from mutual experience and methodologies.

Finally, we plan to produce a RDF version of the corpus and make the results available over the web as Linked Open Data. The raw data that we gathered and the information we extracted after substantial cleaning could provide data for evaluation campaigns (such as automatic Name Extraction, or Multimedia Gender Detection).

4. Conclusions

In this analysis, we faced some difficulty in the use of the available data. The information for L&TC 1995 was not fully available in English in an electronic format. We struggled with the lack of a consistent and uniform identification of entities (authors names, gender, affiliations, paper language, conference and journal titles, funding agencies, etc.). Establishing standards for such identification will demand an international effort in order to ensure that the identifiers are unique, which appears as a challenge for the scientific community.

Research in Language Technology for spoken, written and signed languages has made major advances over the past fifteen years through constant and steady scientific effort that was fostered thanks to the availability of a necessary infrastructure made up of publicly funded programs, largely available language resources, and regularly organized evaluation campaigns.

5. Acknowledgements

The authors wish to thank the ACL colleagues, Ken Church, Sanjeev Khudanpur, Amjbad Abu Jbara, Dragomir Radev and Simone Teufel, who helped them in the starting phase, Isabel Trancoso, who gave her ISCA Archive analysis on the use of assessment and corpora, Wolfgang Hess, who produced and provided 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, Nancy Ide and Christopher Cieri who largely improved the readability of the paper. They also thank Khalid Choukri, Alexandre Sicard, Nicoletta Calzolari, Victoria Arranz, Ioanna Giannopoulou, Johann Gorlier, Jérémy Leixa, Valérie Mapelli and Hélène Mazo, who provided information about the LREC conferences, Marta Witkowska who summarized and translated L&TC 1995, and all the organizers, reviewers and authors over the 17 years L&TC conferences without whom this analysis could not have been conducted!

6. Apologies

This survey has been made on textual data, which cover a 18-year period, including incomplete data for 1995. 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.

7. References

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

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

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

Bravo, E.; Calzolari, A.; De Castro, P.; Mabile, L.; Napolitani, F.; Rossi, A.M. and Cambon-Thomsen, A. (2015), Developing a guideline to standardize the citation of bioresources in journal articles (CoBRA). BMC Med. 2015; 13:33.

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.

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

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

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

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.

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.

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

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.

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

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

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

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.

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.

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.

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

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

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/

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.