Measuring Innovation in Speech and Language Processing Publications.

Joseph Mariani[1], Gil Francopoulo[2], Patrick Paroubek[1]
[1]LIMSI, CNRS, Université Paris-Saclay, [2]Tagmatica
[1]rue John von Neumann, 91400 ORSAY (France), [2]126 rue de Picpus, 75012 PARIS (France)
Joseph.Mariani@limsi.fr, gil.francopoulo@wanadoo.fr, pap@limsi.fr

Abstract

The goal of this paper is to propose measures of innovation through the study of publications in the field of speech and language processing. It is based on the NLP4NLP corpus, which contains the articles published in major conferences and journals related to speech and language processing over 50 years (1965-2015). It represents 65,003 documents from 34 different sources, conferences and journals, published by 48,894 different authors in 558 events, for a total of more than 270 million words and 324,422 bibliographical references. The data was obtained in textual form or as an image that had to be converted into text. This resulted in a lower quality for the most ancient papers, that we measured through the computation of an unknown word ratio. The multi-word technical terms were automatically extracted after parsing, using a set of general language text corpora. The occurrences, frequencies, existences and presences of the terms were then computed overall, for each year and for each document. It resulted in a list of 3.5 million different terms and 24 million term occurrences. The evolution of the research topics over the year, as reflected by the terms presence, was then computed and we propose a measure of the topic popularity based on this computation. The author(s) who introduced the terms were searched for, together with the year when the term was first introduced and the publication where it was introduced. We then studied the global and evolutional contributions of authors to a given topic. We also studied the global and evolutional contributions of the various publications to a given topic. We finally propose a measure of innovativeness for authors and publications.

Keywords: Speech Processing, Natural Language Processing, Text Analytics, Bibliometrics, Scientometrics.

1. Introduction

1.1. Text Analytics of Scientific Papers

The application of text analytics to bodies of scientific papers has become an active area of research in recent years (see for example (Ding Y. et al., 2014), (Banchs R.E., 2012)[1] or the Saffron[2] project. The authors of this paper were invited to conduct several analyses on various conferences: ISCA-Interspeech (J. Mariani et al., 2013), ELRA-LREC (J. Mariani et al., 2014), L&TC (J. Mariani et al., 2015), that they now enlarge to the Speech and Natural Language Processing (SNLP) field in general. They also investigated various aspects of scholar contributions and their evolution over time, such as the production of papers, the collaborations between authors, the citations of papers and authors, the trends in research topics, paper plagiarism and reuse. They proposed measures of the authors' activity based on **paper production**, various kinds of **centrality in collaboration** networks, and **paper citation**. The present paper makes a link between those analyses and proposes in addition a measure of **innovation** that could be attached to research topics, authors or publications. It is based on the detection of the introduction of new terms and on their use in the SNLP research community, assuming that they correspond to a new research topic and exploring who introduced them, when and where, and how successful has the research topic been since then as reflected by the use of the corresponding term after its introduction.

1.2. The NLP4NLP Corpus

In order to conduct our study, 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). It contains papers from 34 publications, conferences and journals, on SNLP published over 50 years (1965-2015), thereby providing a good picture of research within the international SNLP community. The time span, number and frequency of the events (venues for the conferences, or issues for the journals) and number of papers may strongly vary across the publications. The number of sources globally increased over the year but seems now to be stabilizing at 34. The number of documents also fluctuates over the years, mainly due to the biennial frequency of some conferences. However the total number of papers itself increases steadily reaching a total of more than 65,000 documents as of 2015. In order to study the possible differences across different communities, we considered 3 different research areas, Speech, NLP and Information Retrieval (IR), and we attached the sources to each of those areas, given that some sources may be attached to several areas. The number of documents related to Speech is larger than the one related to NLP, and both are much larger than the one related to IR.

1.3. Data Acquisition

Most of the documents are available in PDF. Those that contain scanned images instead of plain text had to be converted with Tesseract-OC[3] before having their textual content extracted with PDFBox (B. Litchfield, 2005) like the others. A benchmark to estimate the error rate of the extracted content was established based on the ratio of unknown words, using the morphological module of TagParser[4] (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).

[1] The results of these analyses together with corresponding data and tools are also available on-line at the University of Michigan. http://clair.eecs.umich.edu/aan/index.php.
[2] http://saffron.deri.ie
[3] https://code.google.com/p/tesseract-ocr/
[4] www.tagmatica.com

Following this content extraction, another step in our preprocessing was dedicated to split the content into abstract, body and references sections, when they exist. It resulted in a corpus that contains about 270 million words, the quality of which got improved over time. The study of authors is problematic due to variations in the rendering of names (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), which resulted in a list of 48,894 different authors. The number of authors also varies across the sources. The most productive author published 358 papers, while 26,870 authors (55% of the authors) published only one paper.

2. Terms and Topics

2.1. Term Extraction

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 extracted from the original electronic documents. Over these 50 years, the archives contain 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 (P. Drouin, 2004). 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, using the TagParser deep syntactic parser applied on a corpus containing the British National Corpus (aka BNC), the Open American National Corpus (aka OANC), the Suzanne corpus release-5, the English EuroParl archives (years 1999 until 2009), plus a small collection of newspapers in the domain of sports, politics and economy, taking care of avoiding any texts dealing with SNLP. In a second step, we parsed the NLP4NLP corpus with the same filters and used our language model to distinguish SNLP-specific terms from common ones. We worked from the hypothesis that when a sequence of words is *inside* the NLP4NLP corpus and *not inside* the general language profile, the term is specific to the field of SNLP. The 65,003 documents written by 48,894 authors reduce to 61,661 documents written by 42,278 authors when considering only the papers written in English. They include 3,485,408 different terms (unigrams, bigrams and trigrams) and 23,803,462 term occurrences, provided that this number counts all the occurrences of all the sizes.

Rank	Term	Variants of all sorts	Date when the term appeared	Authors who introduced the term	Documents	Archive #Occurrences	Archive frequency	Archive #Existences	Archive Presence	Archive Rank Occurrence	Archive Rank Presence	Archive Ratio occurrences / existences	# occurrences in the last year	# existences in the last year	Frequency in the last year	Presence in the last year
1	dataset	data-set, data-sets, datasets	1966	Laurence Urdang	cath1966-3	65250	0.003	9940	0.16	11	18	6.6	14039	1472	0.0092	0.44
2	metric	metrics	1965	A Andreyewsky	C65-1002	50679	0.002	11335	0.18	19	10	4.5	5425	1108	0.0036	0.34
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	45616	0.002	16939	0.27	22	2	2.7	3463	1095	0.0023	0.33
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	54790	0.002	8885	0.14	16	27	6.2	8024	1037	0.0053	0.31
5	classifier	classifiers	1967	Aravind K Joshi, Danuta Hiz	C67-1007	98229	0.004	11546	0.18	7	9	8.5	8202	1000	0.0054	0.30
6	SR	ASR, ASRs, Automatic Speech Recognition, SRs, Speech Recognition, automatic speech recognition, speech recognition	1970	Josse De Kock	cath1970-9	129979	0.006	20382	0.32	2	1	6.4	8524	1000	0.0056	0.30
7	optimization	optimisation, optimisations, optimizations	1967	Ellis B Page	C67-1032	35257	0.002	10196	0.16	35	16	3.5	3331	903	0.0022	0.27
8	annotation	annotations	1967	Kenneth Janda, Martin Kay	cath1967-12 cath1967-8	111084	0.005	11975	0.19	4	7	9.3	7515	896	0.0049	0.27
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	102057	0.005	13823	0.22	5	4	7.4	7489	860	0.0049	0.26
10	LM	LMs, Language Model, Language Models, language model, language models	1965	Sheldon Klein	C65-1014	116684	0.005	13117	0.21	3	5	8.9	8522	851	0.0056	0.26

Table 1: 10 most present terms in 2015, with variants, date, authors and publications where they were first introduced, number of occurrences and existences in 2015, number of occurrences, frequency, number of existences and presence in the 50 year archive, with ranking and average number of occurrences of the terms in the documents

The 500 most frequent terms were computed over the period of 50 years, according to the following strategy. First, the most frequent terms were computed from raw occurrence counts, and secondly the synonyms sets (aka *synsets*) for the most 200 frequent terms of each year were manually declared in the lexicon of TagParser. We gather in the term *synset*, 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. We will call "*existence*"[5] the fact that a term exists in a document and "*presence*" the percentage of documents where the term exists. We computed in that way the occurrences, frequencies, existences and presences of the terms globally and over time (1965-2015), and the average number of occurrences of the terms in the documents where they exist (see Table 1). The ranking of the terms slightly differs if we consider the frequency or the presence. The most frequent term overall is "HMM" (*Hidden Markov Models*), which doesn't appear on Table

[5] sometimes called "Boolean frequency" or "binary frequency"

1 as it is ranked 16th in 2015, while the most present term is "*Speech Recognition*". The average number of occurrences of the terms in the documents where they exist varies a lot (from 9.3 for "*annotation*" to 2.7 for "*subset*" in Table 1).

2.2. New Terms Introduced by the Authors

We then 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 then take into account the terms that 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 1965, 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 on a given year.

Table 1 provides the ranked list of the 10 most popular terms in 2015 based on their presence. For example, the term *Dataset* appeared first in the year 1966, when it was mentioned in a single paper authored by L. Urdang[6] while it was mentioned 14,039 times in 1,472 papers in 2015, and 65,250 times in 9,940 papers overall (i.e. in 16% of the papers!). From its first mention in the introduction of a panel session by Bonnie Lynn Webber at ACL[7] 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 other examples of terms that are presently most popular.

3. Measuring Innovation

3.1. Measuring the Importance of Topics

We then considered the possibility to measure the importance of a term. Fig. 1 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. 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).

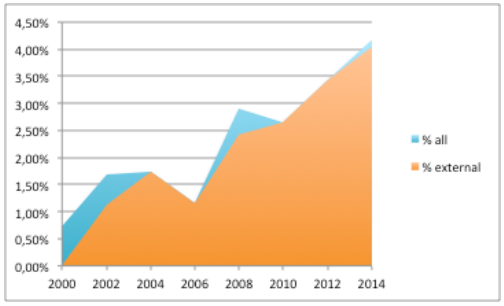

Fig. 1: Presence of the term "*cross validation*" over the years (% of all papers)

We propose to consider as the annual innovation score the presence of the term on that year. In this example, 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 here 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 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 2).

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 2: 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

[6] Laurence Urdang (1966), The Systems Designs and Devices Used to Process The Random House Dictionary of the English Language. Computer and the Humanities. 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.

[7] 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?"*, which may appear as a premonition in 1980!

(percentage of papers containing a given term **up to** a given year) (Fig. 2).

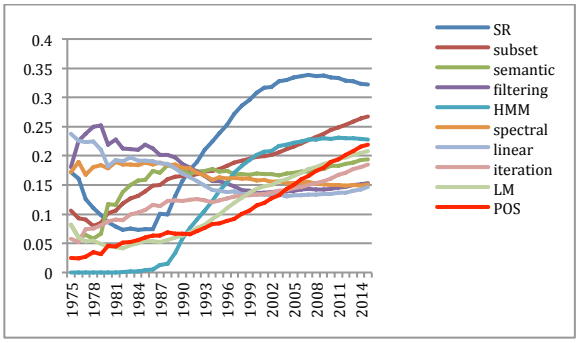

Fig. 2: Cumulative presence of the 10 most important terms over time (% of all papers)

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 up to 2008. Its shape coincides with *Hidden Markov Models* that accompanied the effort on *Speech Recognition* as the most successful method over a long period and had then been mentioned in close to 25% of the papers. *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.

3.2. Measuring Authors' Innovation

We also computed in a similar way an *innovation score* for each author, illustrating his or her contribution in the introduction and early use 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 3).

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

Table 3: Global ranking of authors overall and separately for Speech and NLP

We should stress that this measure doesn't place on the forefront uniquely the "inventors" of a new topic, as it is difficult to identify them given that we only consider a subset of the scientific literature over a limited period. 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 (the *IEEE Transactions on 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 overall ranking also 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 (that we will call "*topical papers*") **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. 3). 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.

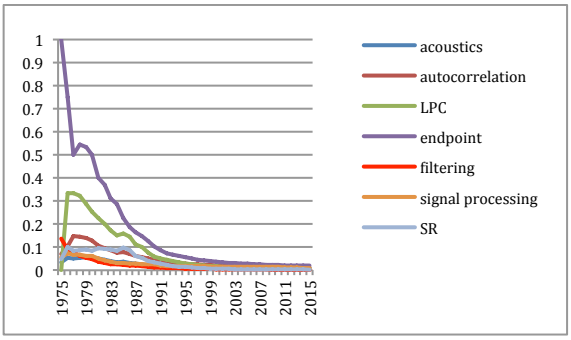

Fig. 3: Main contributions areas for L. Rabiner (% of topical papers)

H. Ney brought important contributions to the study of *perplexity* (authoring 10% of the papers which were published on that topic up to 1988) in *Language Models* (LM) using trigrams and bigrams (Figure 4).

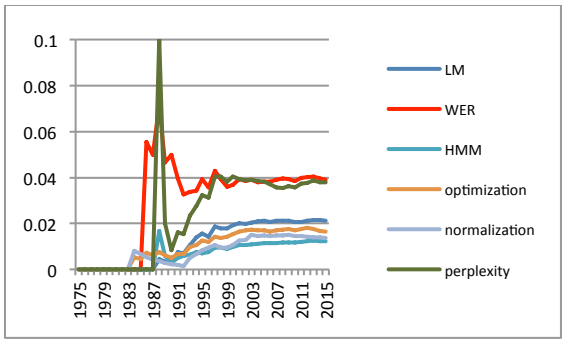

Fig. 4: Main contribution areas for H. Ney (% of topical papers)

A. Waibel brought important contributions in the use of *HMM* and even more of *Neural Networks* for speech and language processing already in the early 90s (Figure 5).

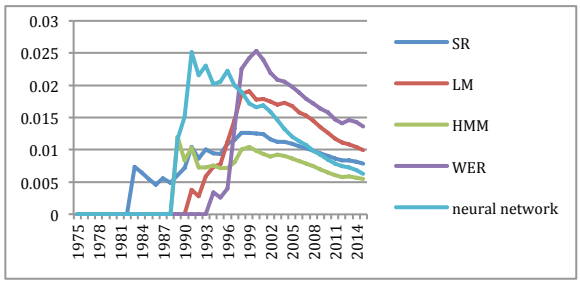

Fig. 5: Main contribution areas for A. Waibel (% of topical papers)

We may also wish to study the contributions of authors on a specific topic, using the same cumulative score. Fig. 6 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 a later stage such as M. Gales.

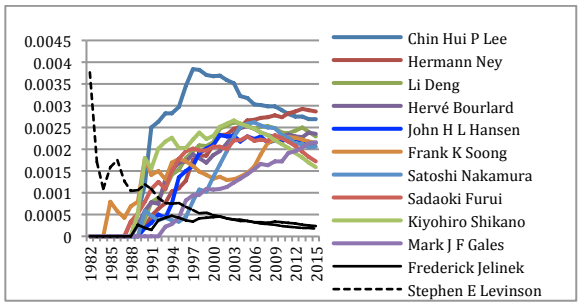

Fig. 6: Authors' contributions to *HMM* in SNLP (% of all papers)

Similarly, we studied the authors' contributions to *Deep Neural Networks (DNN*) which recently gained a large audience (Figure 7). 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 finally places him at the second rank globally.

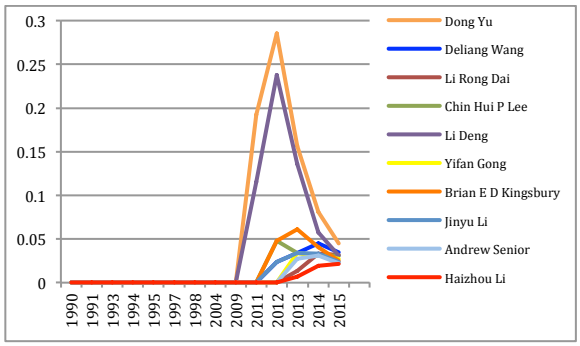

Fig. 7: Authors' contributions to the study of *DNN* in speech and language processing (% of topical papers)

3.3. 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 4).

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 4 : 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) also get a bonus here.

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. 8) 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 recently increased.

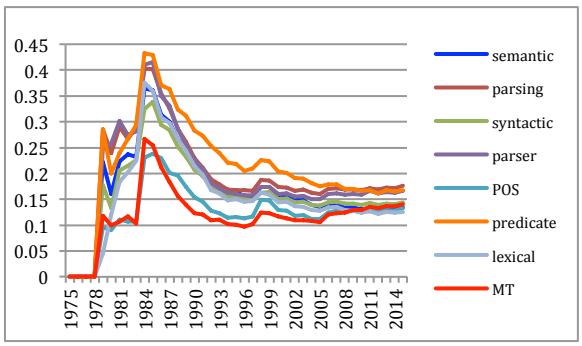

Fig. 8: Main domains within the ACL conference series (% of topical papers)

We may also wish to study the contributions of publications to a specific term, using the same cumulative score. Fig. 9 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 both (CSL, HLT, LREC) helped spreading the approach from speech to NLP.

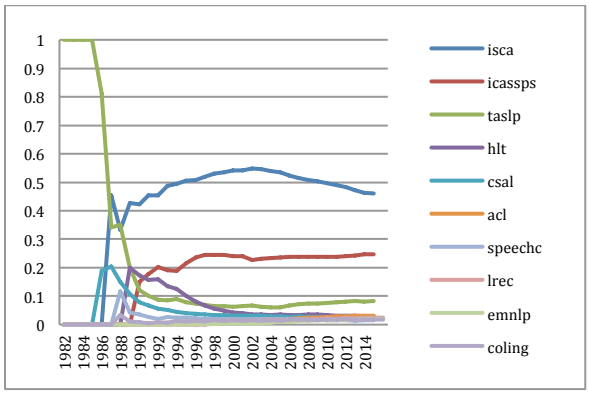

Fig. 9: Sources' contributions to the study of *HMM*. *(% of topical papers)*

4. Perspectives and Conclusions

We proposed in this analysis a measure of innovation for terms, authors and sources. This measure gives an image of the scientific community that seems acceptable. However, it emphasizes the eldest contributions and the productivity. We plan to further refine this measure. We already experimented some variants of the algorithm, such as only considering the periods when the popularity of a term is increasing, without getting very different results. In this analysis, we faced the problem of the lack of quality of the most ancient data that was obtained through OCR from the paper version of the proceedings, which sometimes even contain handwritten comments! For that reason, we focused the study on the period starting in 1975 and we still had to carry out some manual corrections. We plan to develop an automatic term extraction process taking into account the context in which the term is identified. This would allow making the distinction between real and false occurrences of the terms, especially when they have acronyms as variants. It would avoid the tedious manual checking that we presently conduct and would improve the overall process.

5. Bibliographical References

Banchs, Rafael E. (2012), Proceedings of the *ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries* Association for Computational Linguistics 2012 Jeju, Korea. http://aclweb.org/anthology/W12-32

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

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

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.

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

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.

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

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.

6. Language Resource References

Francopoulo, Gil; Mariani, Joseph and Paroubek, Patrick (2016), NLP4NLP: NLP scientific papers for processing with NLP technology http://www.nlp4nlp.org/