Rediscovering 25 Years of Discoveries in Spoken Language Processing: A preliminary ISCA Archive Analysis.

J. Mariani[1,2], P. Paroubek[1], G. Francopoulo[2,3], M. Delaborde[1]
[1]LIMSI-CNRS, [2]IMMI-CNRS, [3]Tagmatica
Joseph.Mariani@limsi.fr, pap@limsi.fr, gil.francopoulo@wanadoo.fr,
delaborde@limsi.fr

Abstract

This paper aims at analyzing the content of the conferences contained in the ISCA Archive over the past 25 years. It follows a similar exercise that has been conducted within the Computational Linguistics community over 50 years of existence at the ACL conference in 2012, and a survey on the IEEE ICASSP conference series from 1976 to 1990, which served in the launching of the ESCA Eurospeech conference. It contains first an analysis of the evolution of the number of papers and authors over time, including their gender and nationality, and of the collaboration among authors. It then studies the references cited in the papers, including their authors and sources. It finally conducts an analysis of the evolution of the research topics within the community over time. The survey shows the present trends in the conference series and in the Spoken Language Processing scientific community. Conducting this survey also demonstrated the importance of a clear and unique identification of authors, papers and other sources to facilitate the analysis. This survey is preliminary, as many other aspects also deserve attention. But we hope it will help better understanding and forging our community in the global village.

Index Terms: ISCA Archive, Spoken Language Processing, Text Analytics, Social Networks, Bibliometrics, Scientometrics.

1. Introduction

1.1 The ISCA community and conference series

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

From 2000 onwards, Eurospeech and ICSLP merged in a single annual Interspeech conference, under the umbrella of the International Speech Communication Association (ISCA), based on ESCA and on the Permanent Council for the organization of the ICSLPs (PC-ICSLP) [1] [2]. On the occasion of Interspeech 2013 in Lyon (France) 24 years after the first Eurospeech conference, which took place in Paris in 1989, it was thought interesting to have a look back at the past years and analyze the steps which resulted in the present situation in spoken language processing science and technology. This analysis aims also at providing a good insight of our community, and may also help building up the next steps for the future.

1.2 The ACL Anthology analysis

A similar inspiring exercise has been conducted by the Association for Computation Linguistics (ACL) on the occasion of their 50th anniversary at the ACL 2012 conference (Jeju, Korea), in the form of a one-day workshop entitled "Rediscovering 50 Years of Discoveries in Natural Language Processing" [3]. This analysis was conducted by 23 authors within 13 papers addressing various aspects, and using technologies developed in the framework of text analytics, a very active area of research in Natural Language Processing nowadays. They used for this the ACL Anthology (http://aclweb.org/anthology/), which contains data coming from the ACL conferences and workshops, but also from other conferences related to Computational Linguistics.

We considered more modestly 25 years of research, given our younger existence. We took the opportunity of the availability of the ISCA Archive (http://www.isca-speech.org/iscaweb/index.php/archive/online-archive), comparable to the ACL Anthology, assembled by Wolfgang Hess, that we deeply thank for his initiative and contribution, which covers the 1987-2012 period. In a first step, we decided to only consider the conferences, starting with the European Conference on Speech Technology (ECST) organized in 1987 in Edinburg, followed by the Eurospeech and ICSLP conference series, and by the Interspeech conference series starting in 2000. We did not take into account the workshops, including the European (then International) Tutorial and Research Workshops organized by ESCA, then ISCA, since 1989, and the other E/ISCA supported events.

1.3 The ICASSP 1976-1990 conference series analysis

A similar, although simpler, analysis was actually conducted by J. Mariani [4] on the IEEE ICASSP conference series (on a 15 years time span from 1976 to 1990), accompanying the launching of the Eurospeech conference in 1989, in his capacity of ESCA president at that time and Technical Chairman of Eurospeech 1989 (Table 1).

Year	Place	Total papers	Total papers on speech	USA	Europe	Japan	Other
1976	Philadelphie	226	119	71	30	10	8
1977	Hartford	239	83	56	18	1	8
1978	Tulsa	227	82	48	18	8	8
1979	Washington	265	116	64	35	3	14
1980	Denver	255	98	73	16	3	6
1981	Atlanta	295	97	63	26	3	5
1982	Paris	540	163	64	70	15	14
1983	Boston	381	123	81	26	12	4
1984	San Diego	576	150	86	40	11	13
1985	Tampa	479	140	85	29	13	13
1986	Tokyo	795	305	110	84	92	19
1987	Dallas	651	182	101	40	19	22
1988	New York	756	193	112	43	17	21
1989	Glasgow	719	214	88	77	33	16
1990	Albuquerque	752	219	113	58	33	15
Total		7156	2284	1215	610	273	186

Table 1. *Analysis of the IEEE ICASSP conference (1976-1990)*

It appeared that the number of papers at ICASSP, in general but also in speech, increased over the years (Fig. 1). The number of speech papers at ICASSP represented overall about 30% of the papers (2284 on 7156), but the ratio of speech papers decreased over time from about 50% in 1976 to 30% in 1990. Looking more precisely, it is striking to notice that, even if the US were the largest providers of speech papers overall (more than 50%) (Fig. 2), whenever the ICASSP conference took place outside the US (Paris (France) in 1982, Tokyo (Japan) in 1986 and Glasgow (UK) in 1989), the total participation increased, the US participation stayed very high, while the European and Asian participation increased a lot (Fig. 3) and even was on a par with the US one (Fig. 4), as it happened typically in Tokyo in 1986. It also resulted in a stronger dynamics of the conference for the following years. This advocated for the launching of truly international conferences more specifically devoted to spoken language processing, while covering all the aspects of this research area, as confirmed by Eurospeech and ICSLP, which immediately obtained a large international success.

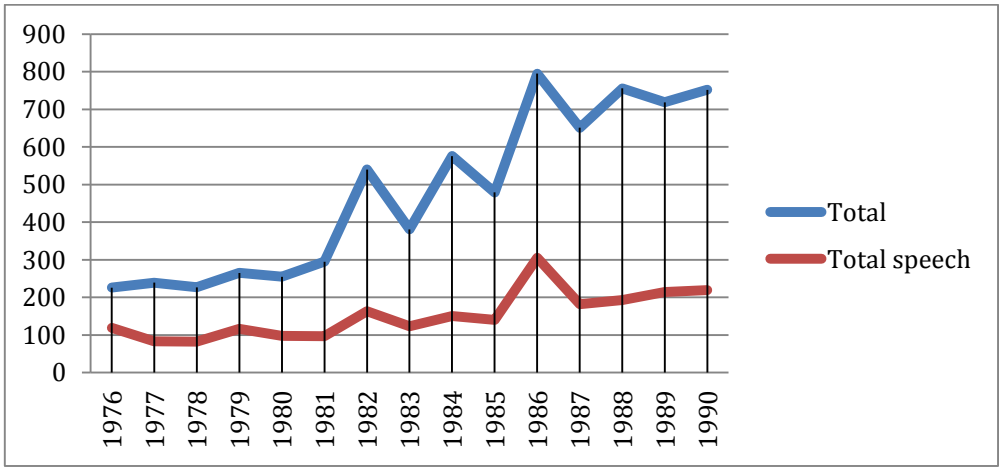


Figure 1. *Evolution of the total number of papers and of the number of speech papers*

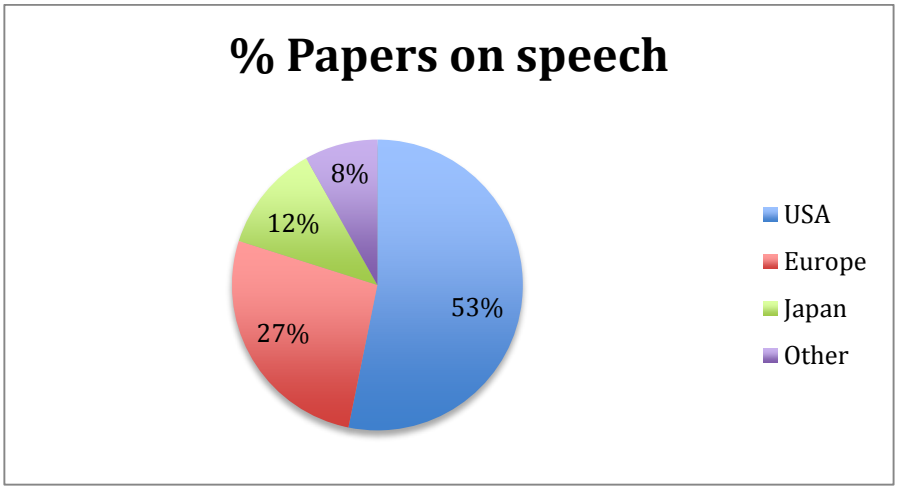


Figure 2. *Percentages of speech papers*

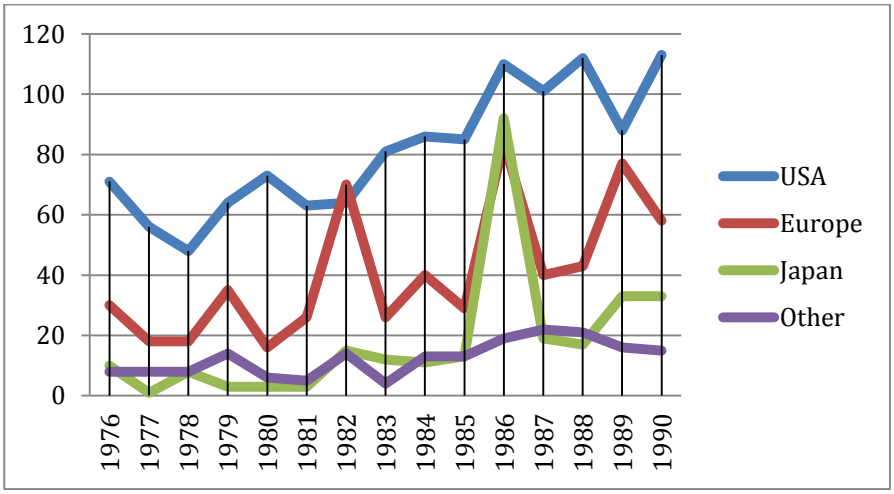


Figure 3. *Evolution of the number of speech papers per geographic origin*

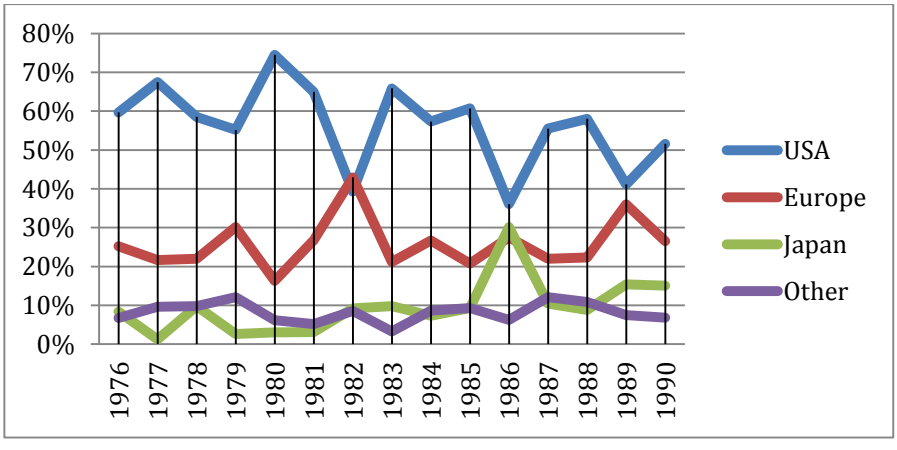


Figure 4. *Evolution of the percentage of speech papers per geographic origin*

2. Analysis of the series of ECST, Eurospeech, ICSLP and Interspeech conferences

Year	Conference	Place	# papers	# authors	# authors/paper
1987	ECST	Edinburgh	252	578	2.29
1989	Eurospeech	Paris	360	854	2.37
1990	ICSLP	Kobe	350	914	2.61
1991	Eurospeech	Genoa	335	858	2.56
1992	ICSLP	Banff	413	1,076	2.61
1993	Eurospeech	Berlin	527	1,367	2.59
1994	ICSLP	Yokohama	560	1,542	2.75
1995	Eurospeech	Madrid	519	1,376	2.65
1996	ICSLP	Philadelphia	635	1,737	2.74
1997	Eurospeech	Rhodes	722	1,946	2.70
1998	ICSLP	Sydney	850	2,361	2.78
1999	Eurospeech	Budapest	723	2,124	2.94
2000	Interspeech	Beijing	924	2,711	2.93
2001	Interspeech	Aalborg	672	1,988	2.96
2002	Interspeech	Denver	679	1,932	2.85
2003	Interspeech	Geneva	798	2,330	2.92
2004	Interspeech	Jeju	774	2,239	2.89
2005	Interspeech	Lisbon	869	2,606	3.00
2006	Interspeech	Pittsburgh	659	2,037	3.09
2007	Interspeech	Antwerp	751	2,346	3.12
2008	Interspeech	Brisbane	762	2,442	3.20
2009	Interspeech	Brighton	765	2,455	3.21
2010	Interspeech	Makuhari	781	2,525	3.23
2011	Interspeech	Firenze	846	2,748	3.25
2012	Interspeech	Portland	680	2,218	3.26
			16,206	47,310	2.92

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

The study covers the series of conferences contained in the ISCA Archive, starting with the ECST conference (Edinburg, 1987), followed by the biennial Eurospeech conference series starting in Paris (France) in 1989 and ICSLP conference series starting in Kobe (Japan) in 1990, which merged into the Interspeech conference series since 2000 in Beijing until 2012 (see Table 2). This covers a series of 25 events and a time span of 25 years (1987-2012). We did not consider for the time being in this study the workshops, and especially the E/ISCA Tutorial and Research Workshops (E/ITRW) organized since 1989, and the E/ISCA supported events, which are also contained in the Archive.

2.1. The resources: data and tools

Regarding the conference series, a part of the ISCA Archive, that we may call metadata, is available online (List of authors and sessions, Content of the sessions and, for each article, Titles, Authors, Affiliations, Abstract and Bibliographic Reference of the paper), while the full content of the articles is only available for the ISCA members. In this study, we used the metadata for the chapters 2.2. (Papers) and 2.3. (Authors), and the full content for the chapters 2.4. (Citations) and 2.5. (Topics).

The metadata were processed with MS Excel, OpenOffice spreadsheet Calc, the R statistical suite [5], the search engine swish-e [6], RankChart and various scripts written in bash shell and C++. The linguistic processing was limited to the use of G. Grefenstette awk implementation of Porter's stemmer [7] and local grammars compiled either with the Unitex toolkit [8]

or flex [9]. The large graph visualization and analysis platform Tulip [10] was used to browse the co-author and publication graphs.

For analyzing the articles content, we re-used the toolkit developed for the processing of the ACL Anthology [11], extracting (when possible) the text from the pdf version of the articles with pdfbox [12] and parsing it with ParsCit [13] to identify the various sections and citation elements. The text was then parsed with the syntactic analyzer TagParser [14].

Along with the previous toolkits, we have used the following language resources: the British National Corpus (BNC) [15], the Open American National Corpus (OANC) [16], Europarl [17], Tagmatica Named Entity database extracted from Wikipedia and various journalistic sources, and a lexicon of 59,850 first names with gender information.

2.2. The papers

The total number of papers published in the conference series amounts to 16,206 (Table 2), with a steadily increase over time from 252 in 1987 to 924 at Interspeech 2000, followed by a relative stability with lower numbers (less than 700) in Aalborg (2001), Denver (2002), Pittsburgh (2006) and Portland (2012), and higher numbers (more than 800) in Lisbon (2005) and Firenze (2011) (Fig. 5). It should be noted that it was decided to rise up the rejection rate at the Interspeech 2012 conference in Portland, thus resulting in a decrease of the number of papers.

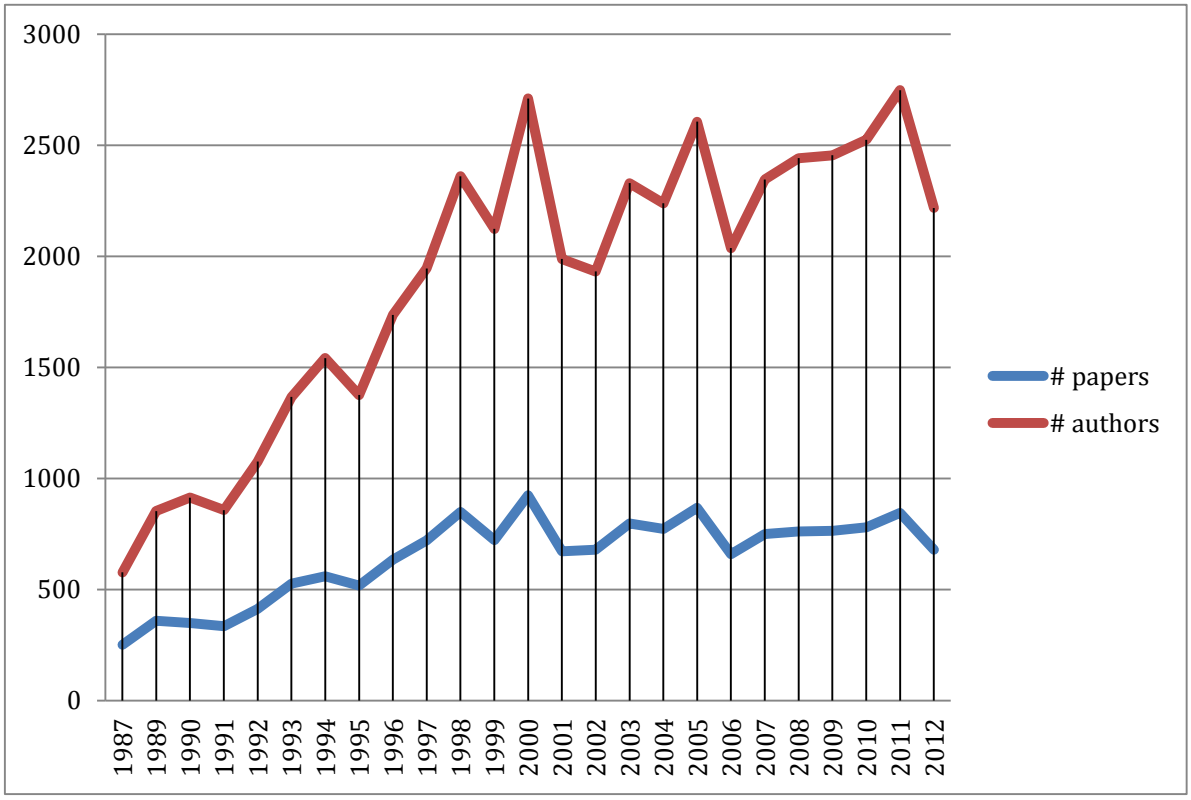


Figure 5. *Number of papers and authors over time*

2.3. The authors

2.3.1. Number of authors per conference

The number of authors also steadily rose up to 2,711 at Interspeech 2000. It then stayed very high at the level of 2000 and even reached 2,748 at Interspeech 2011. It went down at Interspeech 2012, also due to the higher rejection rate, which resulted in a lower number of papers (Fig. 5).

2.3.2. Number of authors per paper

Overall, most papers have 2 to 3 co-authors (Fig. 6). The largest number of co-authors for a paper is 21.

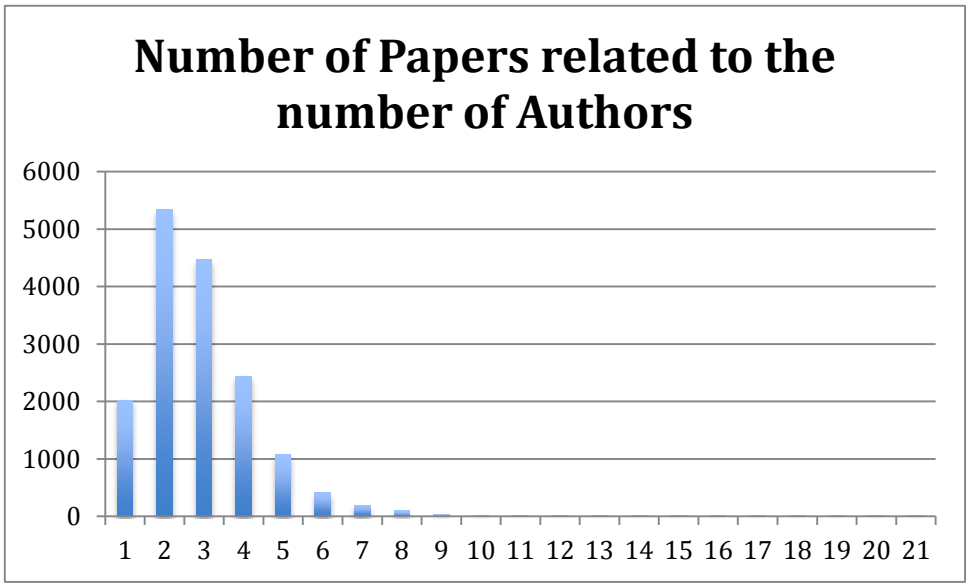


Figure 6. *Number of papers according to the number of authors*

2.3.3. Number of authors per paper over time

However, the average number of co-authors per paper increased over time, from 2.29 in 1987 up to 3.26 in 2012 (i.e. almost one more author on average) (Fig. 7).

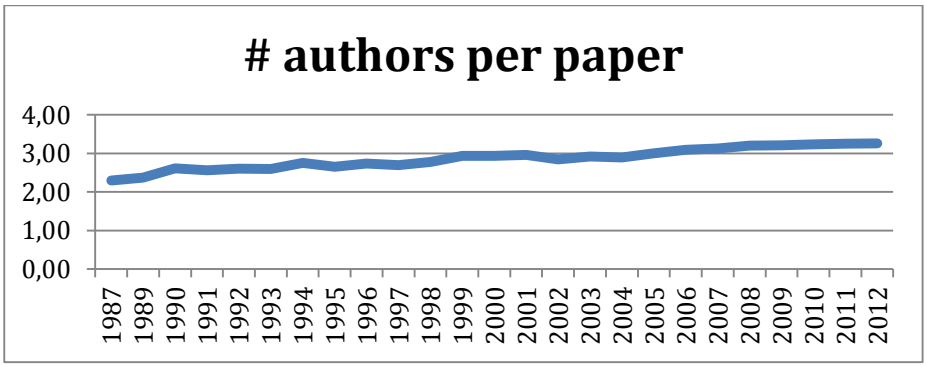


Figure 7. *Average number of authors per paper*

It is striking to notice that the number of papers with a single author was 30% in 1987 and went down to 5% in 2012, while the number of papers with 3 authors or more was 36% in 1987, and went up to 65% in 2012! This clearly demonstrate the change on the way research is conducted, going from individual research investigations to large projects conducted within teams or in collaboration within consortia, often in international programs.

2.3.4. Number of different authors

The study of the authors is difficult due to the various ways of writing their name (family name and given name, initials, middle initials, ordering, married name, etc.). It therefore necessitated a tedious cleaning process, which was made by hand. On an initial total of 16,445 authors' names, about 2,000 family names or given names were corrected, resulting in a list of 14,630 different authors. This clearly demonstrates the need for identifying uniquely each researcher.

2.3.5. Renewal of authors

We first studied the number of authors at each following conference (Table 3).

Year	Authors	Different authors	New authors	Completely new authors	% Similar authors	% New authors	% Completely new authors
1987	578	481	481	481	17%	100%	100%
1989	854	655	496	496	23%	76%	76%
1990	914	728	606	587	20%	83%	81%
1991	858	687	563	408	20%	82%	59%
1992	1,076	825	645	437	23%	78%	53%
1993	1,367	1,030	761	523	25%	74%	51%
1994	1,542	1,190	894	605	23%	75%	51%
1995	1,376	1,071	759	495	22%	71%	46%
1996	1,737	1,288	912	596	26%	71%	46%
1997	1,946	1,476	1,032	678	24%	70%	46%
1998	2,361	1,662	1,151	765	30%	69%	46%
1999	2,124	1,604	1,044	655	24%	65%	41%
2000	2,711	1,751	1,185	767	35%	68%	44%
2001	1,988	1,472	882	523	26%	60%	36%
2002	1,932	1,455	937	560	25%	64%	38%
2003	2,330	1,705	1,156	671	27%	68%	39%
2004	2,239	1,581	998	642	29%	63%	41%
2005	2,606	1,866	1,256	695	28%	67%	37%
2006	2,037	1,485	896	580	27%	60%	39%
2007	2,346	1,752	1,171	657	25%	67%	38%
2008	2,442	1,704	1,073	614	30%	63%	36%
2009	2,455	1,755	1,102	577	29%	63%	33%
2010	2,525	1,747	1,038	575	31%	59%	33%
2011	2,748	1,899	1,128	588	31%	59%	31%
2012	2,218	1,545	877	487	30%	57%	32%
Total	47,310	34,414	23,043	14,662	27%	67%	43%

Table 3. *Authors renewal Table*

The difference between the number of authors and the number of different authors reflects the number of authors whose name appear in several papers, what we may call the "authors variety", and the inverse "authors redundancy"). It appears that this redundancy slightly increased over time, showing a concentration of the papers authors (Fig. 8).

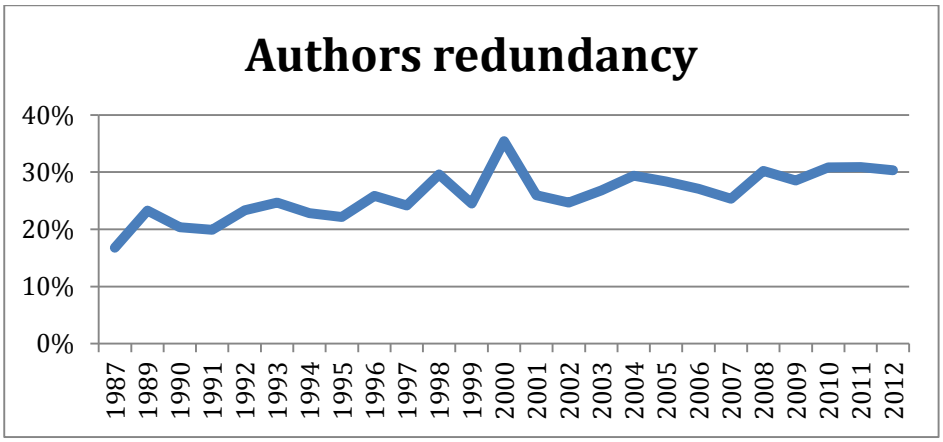


Figure 8. *Authors redundancy over time*

We then studied the authors' renewal. It clearly shows (Fig. 9) that the number of different authors from one conference to the next conference has been high and increased over time until Interspeech 2000, where there were about 1,200 new authors compared with 1999. It then stayed steadily important with a turn over of about 1,100 different authors each year. We also studied the turn over separately at Eurospeech and ICSLP conferences in order to check if it was different, but it seems to be comparable. The same appears for the number of totally new authors which increased every year up to Interspeech 2000, with 767 new authors that year, but then slightly decreased over time to 487 in 2012. This also appears in terms of percentages (Fig. 10) showing that the percentage of different authors from one year to the next decreased from 75% in 1989 to less than 60% in 2012, while the number of totally new authors decreased from 75% in 1989 to about 30% in 2012. This shows the stabilization of the research community over time, but may also reflect a lack of "new blood".

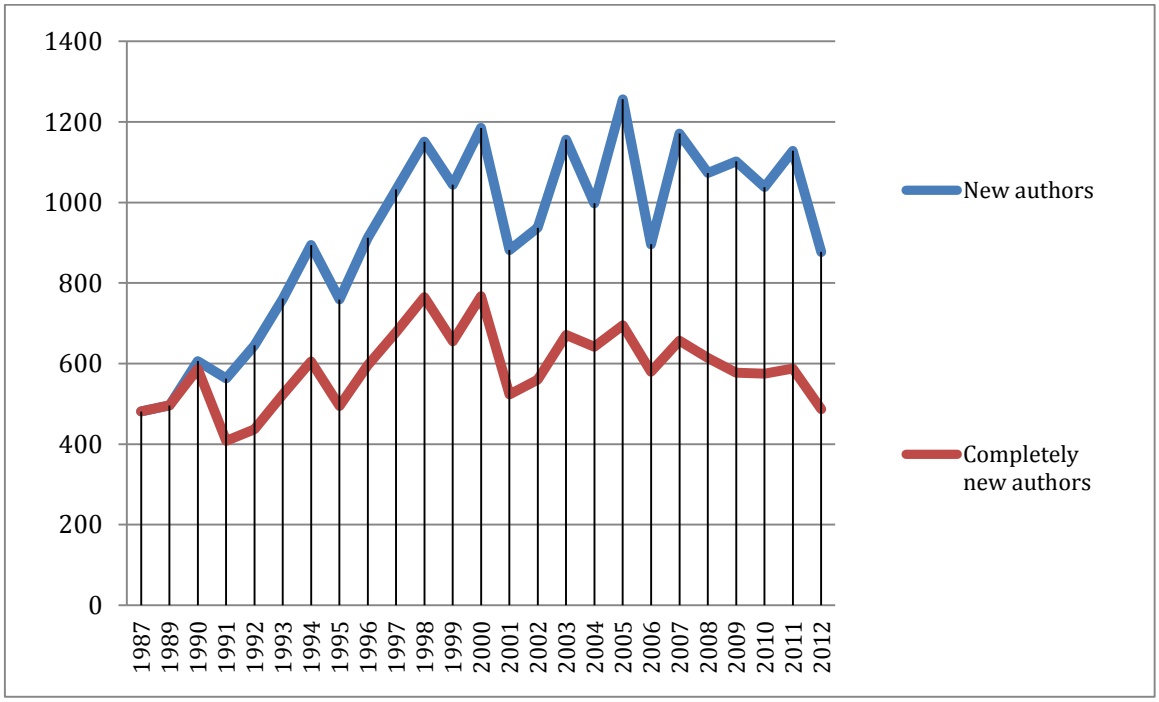


Figure 9. *Number of authors, new authors and completely new authors over time*

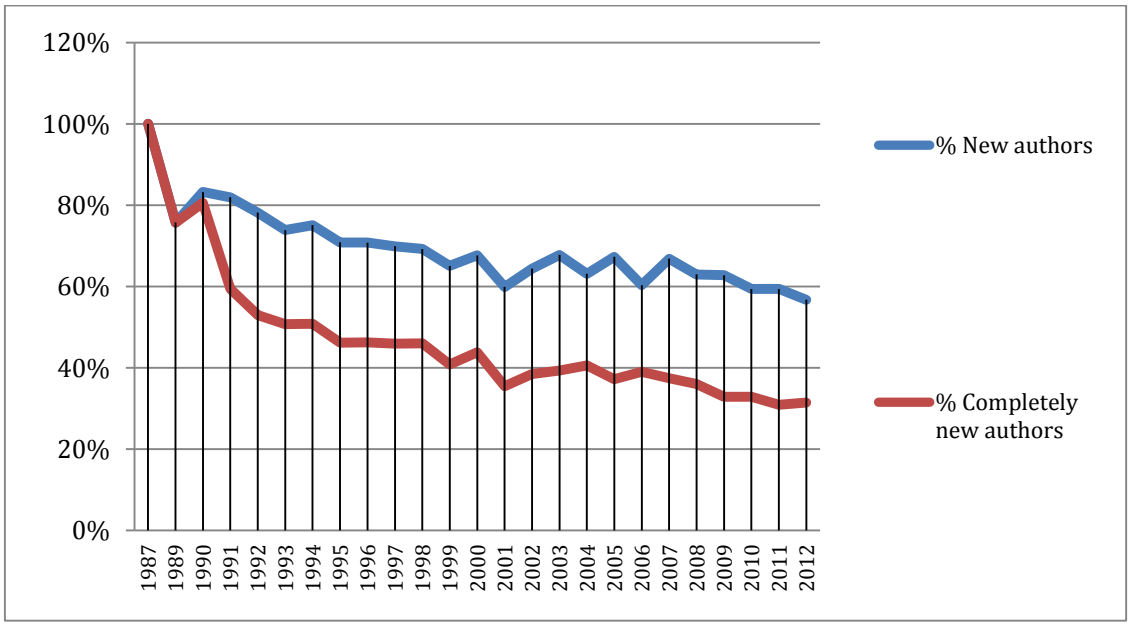


Figure 10. *Percentage of new authors and completely new authors over time.*

2.3.6. Gender of authors

The author gender study was performed with the help of a lexicon of 59,850 first names with gender information (54% male, 44% female, 2% epicene). Variations due to different cultural habits for naming people [18] (single versus multiple given names, family versus clan names, inclusion of honorific particles, ordering of the components etc.), changes in editorial practices and sharing of the same name by large groups of individuals, all contribute to make identifying the person referred to by a name a difficult problem, so much that initiatives exist to provide world-wide unique identifiers for researchers [19]. In this preliminary study we have used a crude normalization of proper names in ASCII, separating them into two components: given name and family name, allowing for compound forms in both parts. Note that for some of them, we only had an initial for the first name, which made gender guessing impossible, unless the same person also appears with his/her first name in full somewhere else. Although the result of the automatic processing was hand-checked by an expert of the domain for the most frequent names, the results presented here need to be considered with caution allowing for an error margin.

The analysis over the 25 conferences shows that 50% of the authors are male, while only 17% of the authors are female, with 1% of epicene gender but 31% are of unknown gender (Fig. 11 and 12). If we consider that the authors of unknown gender have the same gender distribution than the ones which are categorized, the ratio of male authors would be 74%, while the female authors are 26% (Fig. 13).

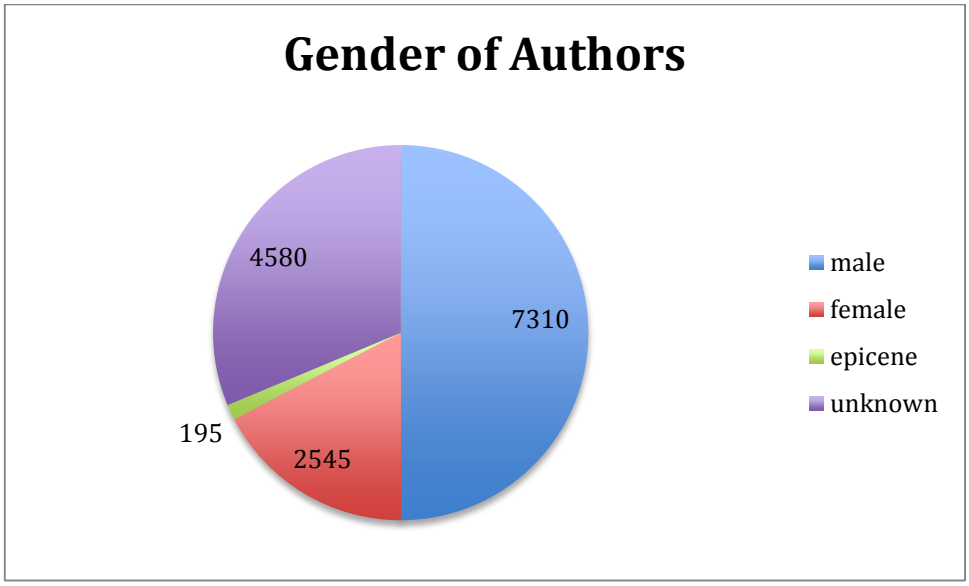


Figure 11 *Gender of the 14 630 authors over all*

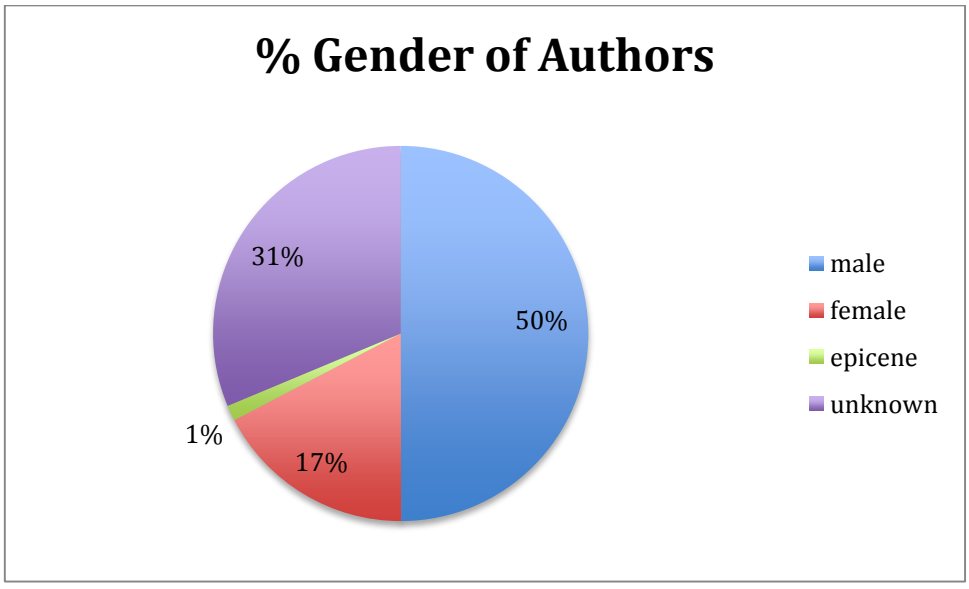


Figure 12. *Percentages of gender of the 14 630 authors over all*

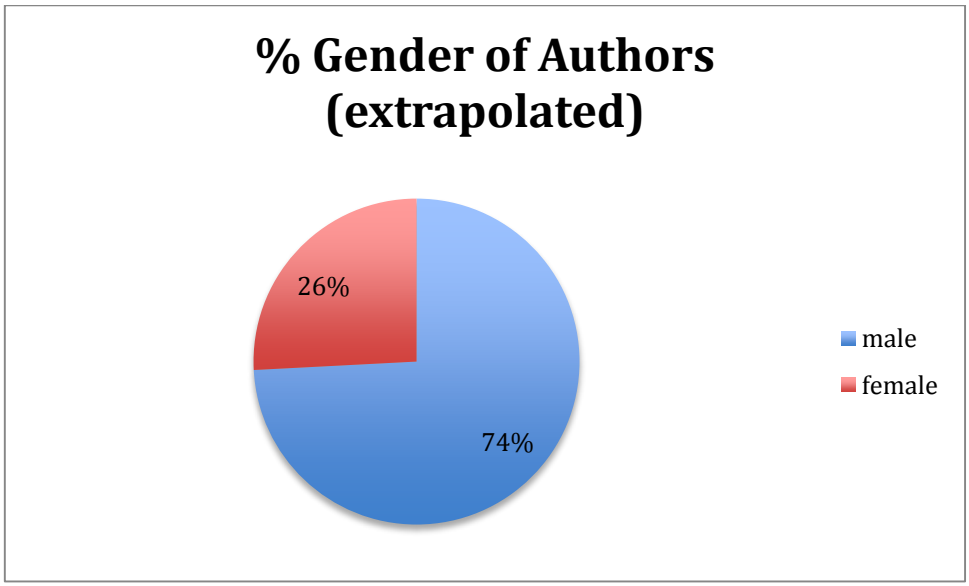


Figure 13. *Percentages of genders over all under the assumption that the distribution on unknown gender is similar*

If we now consider the contribution by gender over the 16,206 papers (Fig. 14), we find even a slight increase in the male contribution (78% against 22%).

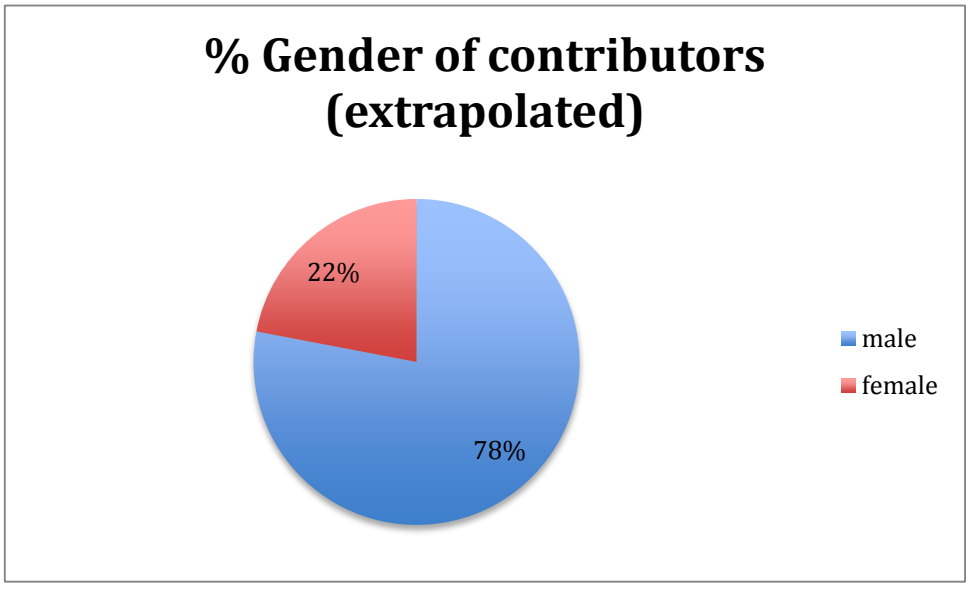


Figure 14. *Gender of the authors' contributions over all*

The analysis of the authors' gender over time (Fig. 15) however shows a slight decrease of male authors (from 83% in 1987 to 75% in 2012) and an increase of the female authors, from 17% to 25% (+50% relative).

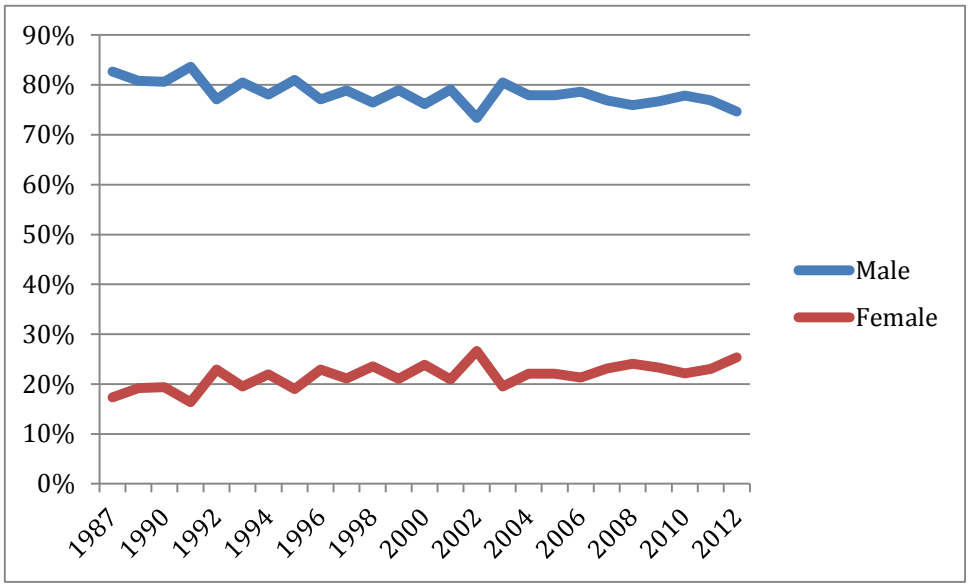


Figure 15. *Gender of the authors' contributions over time.*

2.3.7 Nationality of authors

We studied the nationality of the papers authors. When a paper is signed by several authors of the same country, it is counted as a single paper for the country. When it is signed by several authors of different countries, it is counted as one paper for each country. Papers have been published by authors of 73 countries (Fig. 16).

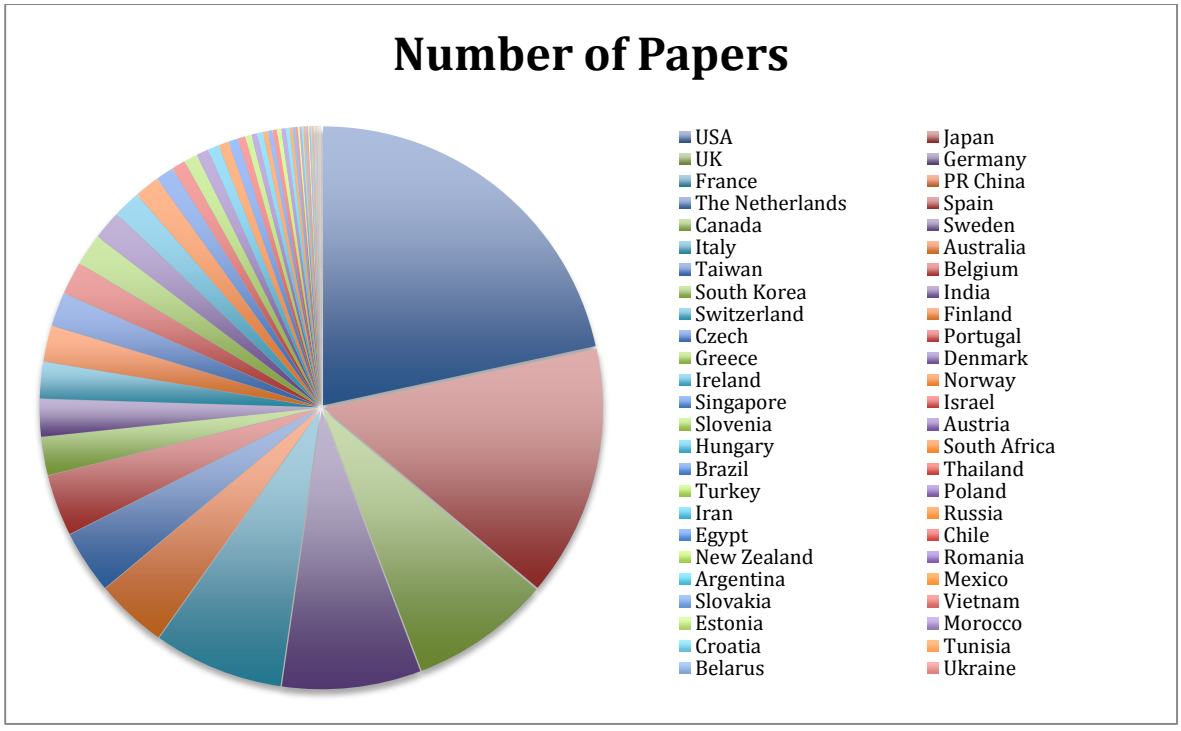


Figure 16. *Number of papers per country*

The 5 most publishing countries represent 60% of the authors: USA (22%), Japan (15%), Germany (8%), UK (8%), and France (8%). PR China comes next with 4% (Fig. 17).

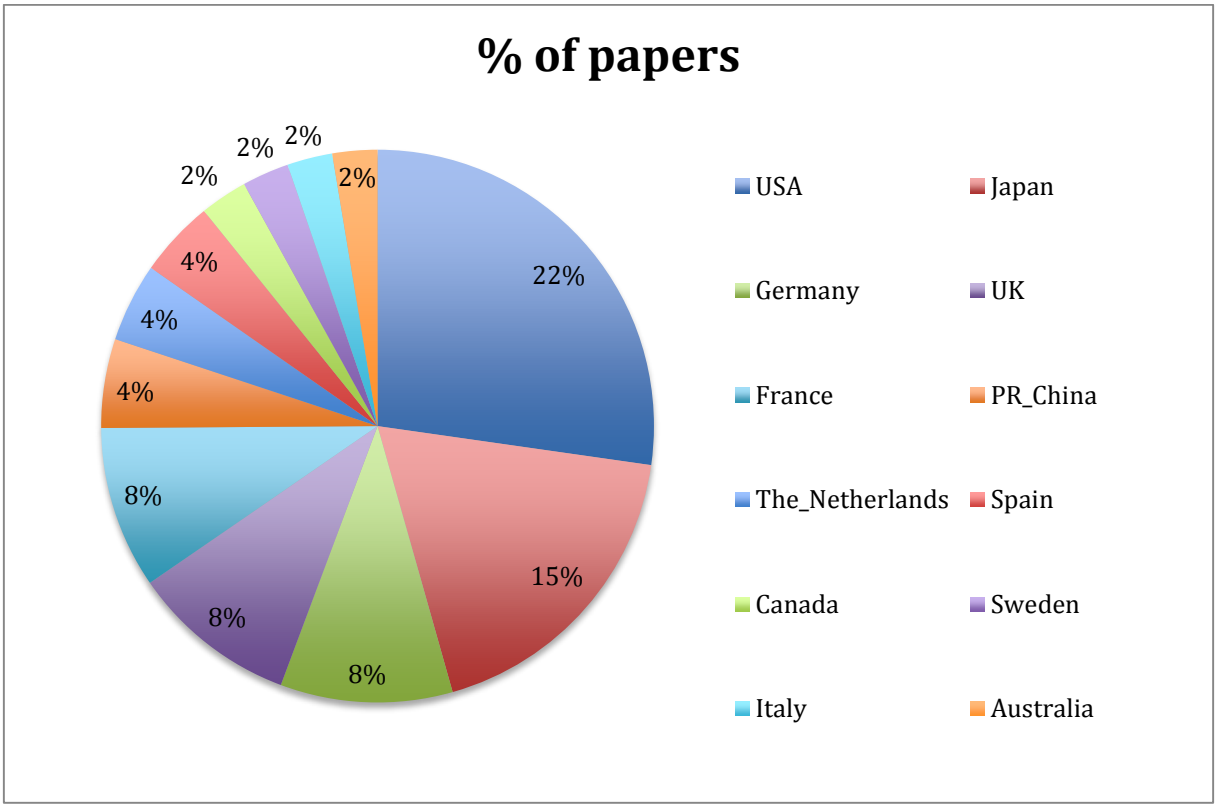


Figure 17. *Number of papers per country for the 12 most cited countries*

If we cluster the countries as we did for the ICASSP analysis (Cf 1.3), we see that Europe has the largest share (46% of the papers) (Fig. 18). The same appears if we consider continents (Fig. 19), but Asia (24%) gets then close to America (25%).

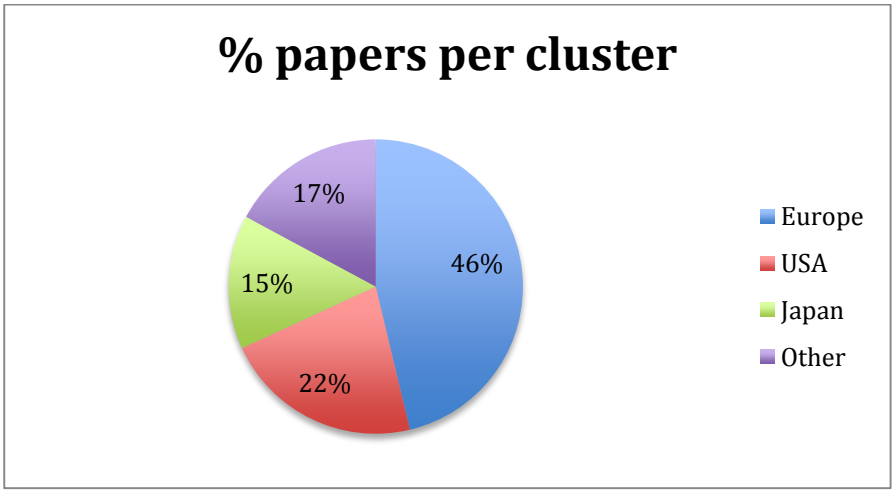


Figure 18. *Percentages of papers according to the same clustering of countries as in the ICASSP analysis.*

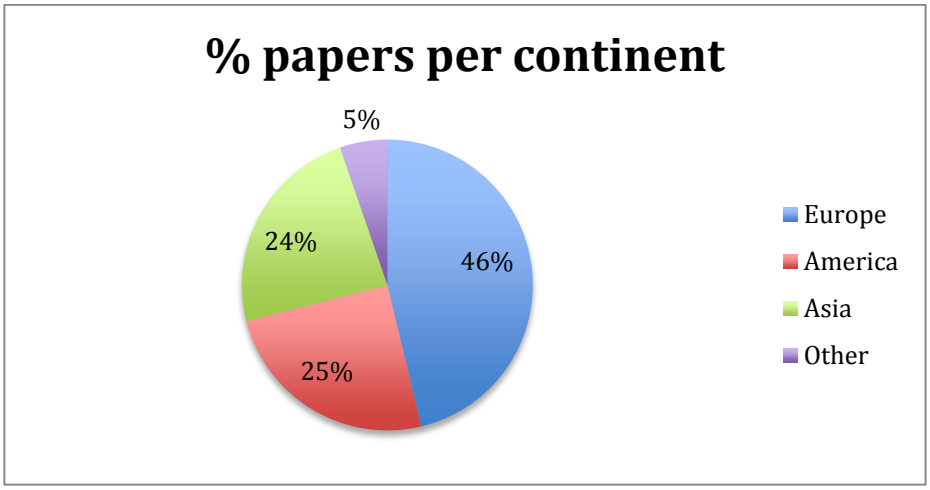


Figure 19. *Percentages of papers according to continents.*

If we now consider the evolution of the share of papers per country over time, for the 8 countries totaling more than 4% of the papers overall (Fig. 20), we see that the trend is that the share of USA slightly increased until 1996 and is steady since then (about 25%), while the share of Japan recently slightly decreased, starting in 2004. The share of Germany slightly increased and is now on a par with the one of Japan, while the share of PR China strongly increased and is now on a par with the ones of UK and France. The share of The Netherlands and Spain slightly decreased, except in particular when these countries were organizing the conference (Spain in 1995 and The Netherlands in 2007).

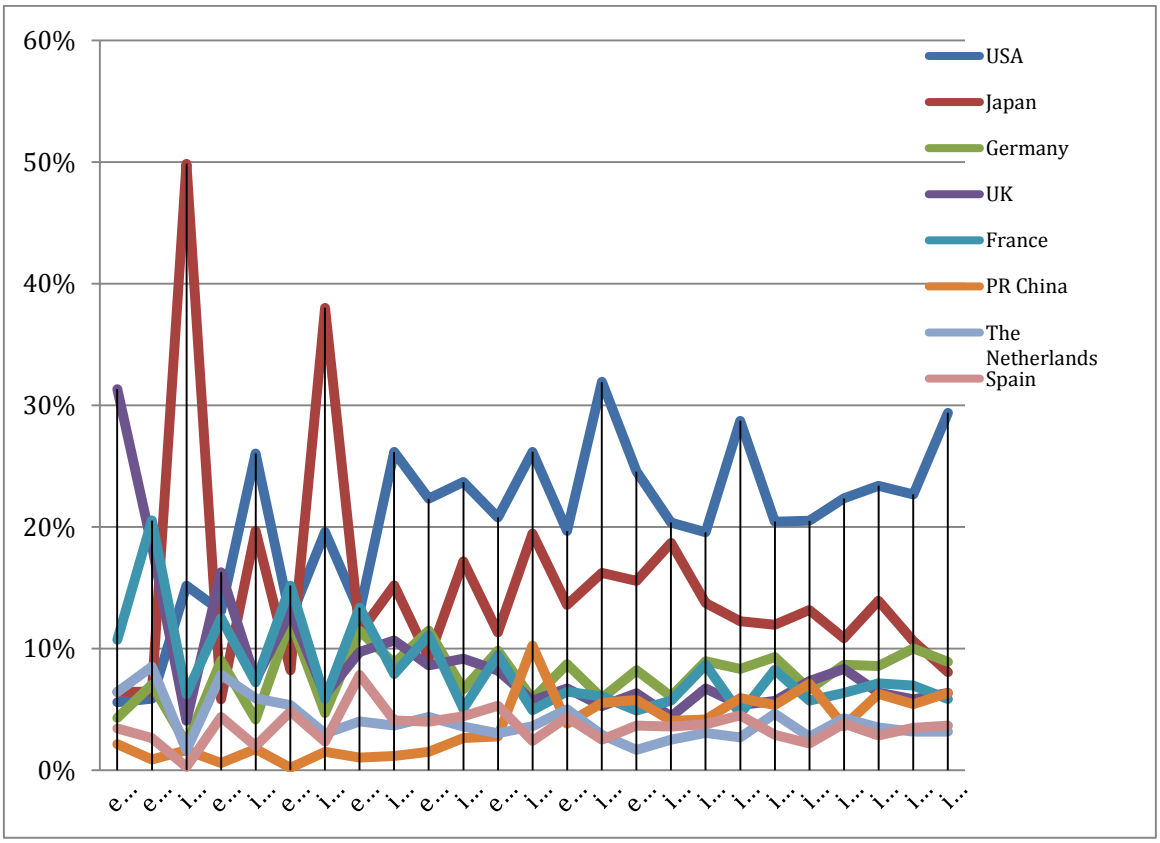


Figure 20. *Evolution of the share of papers per country over time for the 8 most cited countries.*

If we cluster the countries in the same way as we did for the survey on ICASSP conference series (Fig. 21), we see that the share of Europe is still the largest, but it decreased over time before stabilizing since 2000, with the strongest participation when the conference takes place on the European continent. As already mentioned, the share of Japan recently slightly decreased, while the share of the USA is steady and the share of the countries placed in the "Other" category strongly increased.

If we now cluster the countries into “Continents” (Fig. 22), we see that Asia is now on a par with USA, while the countries of the other continents are still lying behind, despite a slight increase over time.

We find the same phenomenon already mentioned in the ICASSP Survey: the participation of the representatives of various continents is related to the continent where the conference is taking place; hence the sawtooth appearance of the curves. Interestingly, it appears that the American participation is in phase with the Asian one, and in opposite phase with the European one until 2005, when this phenomenon gets weaker. A strong participation and good balance has been reached between Asia and Europe at Interspeech 2000 (Beijing, PR China) and 2004 (Jeju, Korea), and between Europe and the USA at Interspeech 2002 (Denver, USA) and 2006 (Pittsburgh, USA). The other continents were most represented when the conference took place in Australia (Sydney in1998 and Brisbane in 2008).

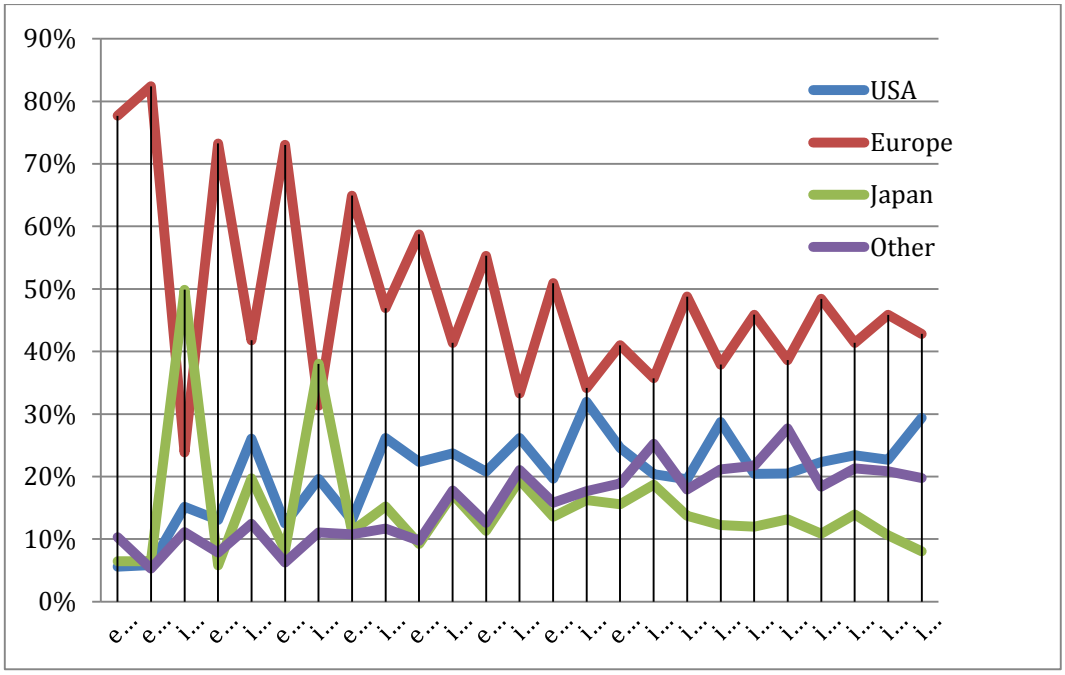


Figure 21. *Evolution of the share of papers per cluster.*

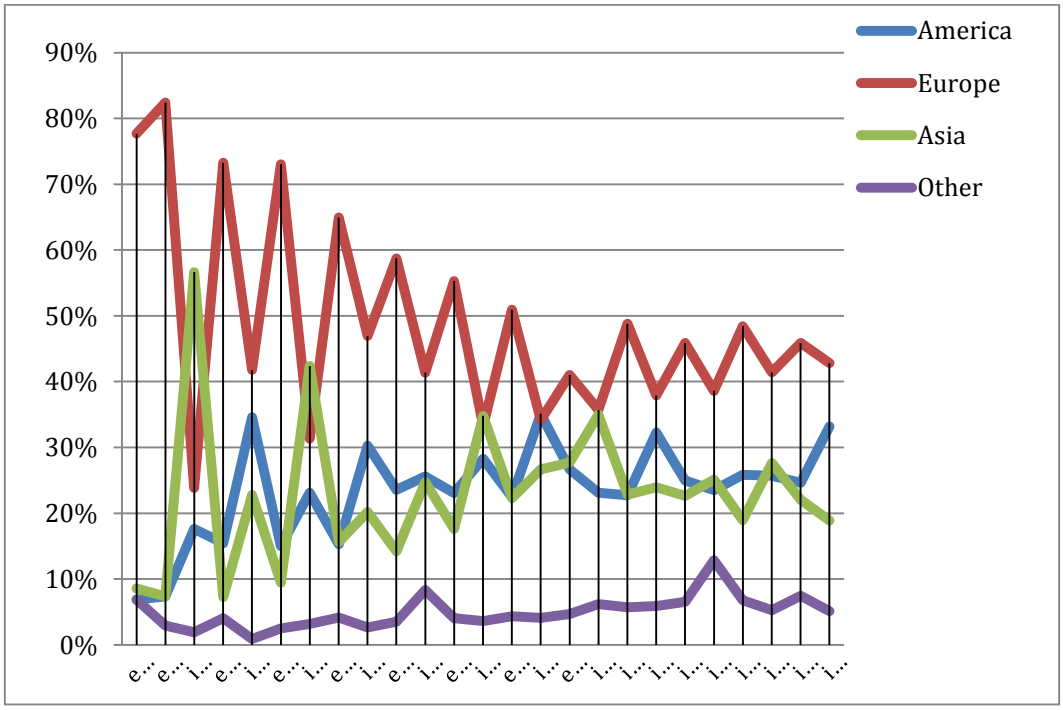


Figure 22. *Evolution of the share of papers per continent.*

2.3.8. Authors production

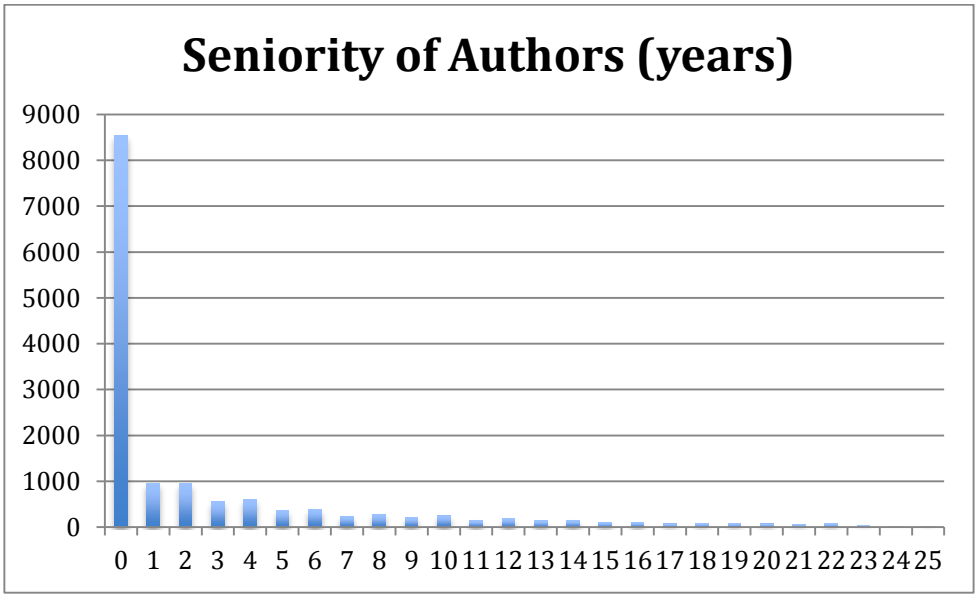


Figure 23. *Authors Seniority*

20 authors published over the time span of the 25 conferences, what we may call author seniority (Fig. 23):

William J. BARRY, Louis BOVES, Mike BROOKES, Martin P. COOKE, Maxine ESKENAZI, Yifan GONG, Phil D. GREEN, Hynek HERMANSKY, David HOUSE, Mark HUCKVALE, Tetsunori KOBAYASHI, Francisco LACERDA, Eduardo LLEIDA-SOLANO, Climent NADEU, José_Manuel PARDO, Josef V. PSUTKA, Hugo QUENE, Stephen RENALS, Isabel M. TRANCOSO and Bayya YEGNANARAYANA.

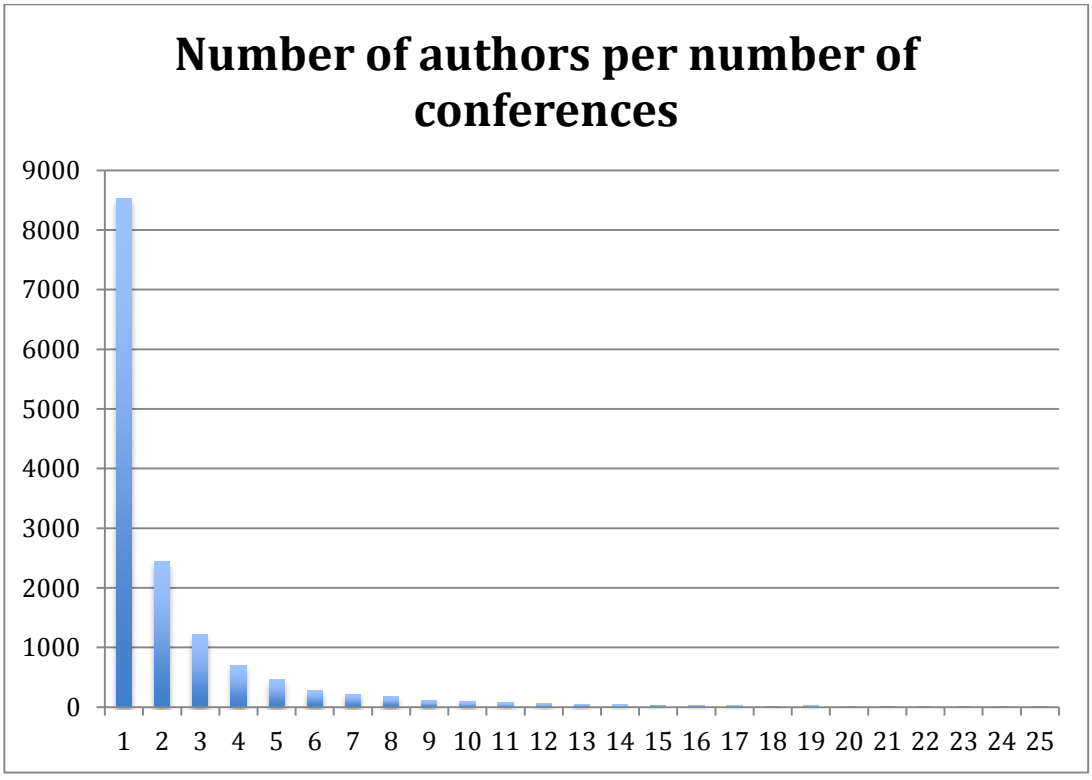


Figure 24. *Number of Authors per Number of Conferences*

A single author published at all 25 conferences (Louis BOVES), while 4 published at 24 conferences (Climent NADEU, Seiichi NAKAGAWA, Yoshinori SAGISAKA and Isabel TRANCOSO), 3 at 23 conferences (Javier HERNANDO, Chin-Hui LEE and Helmer STRIK),) and 9 at 22 conferences (Hervé BOURLARD, Nick CAMPBELL, Sadaoki FURUI, Jean-Paul HATON, Hynek HERMANSKY, Keikichi HIROSE, Tsuneo NITTA , Kiyohiro SHIKANO and Alex WAIBEL) (Fig. 24).

8,531 authors published at a single conference (58% of the 14,630 authors)

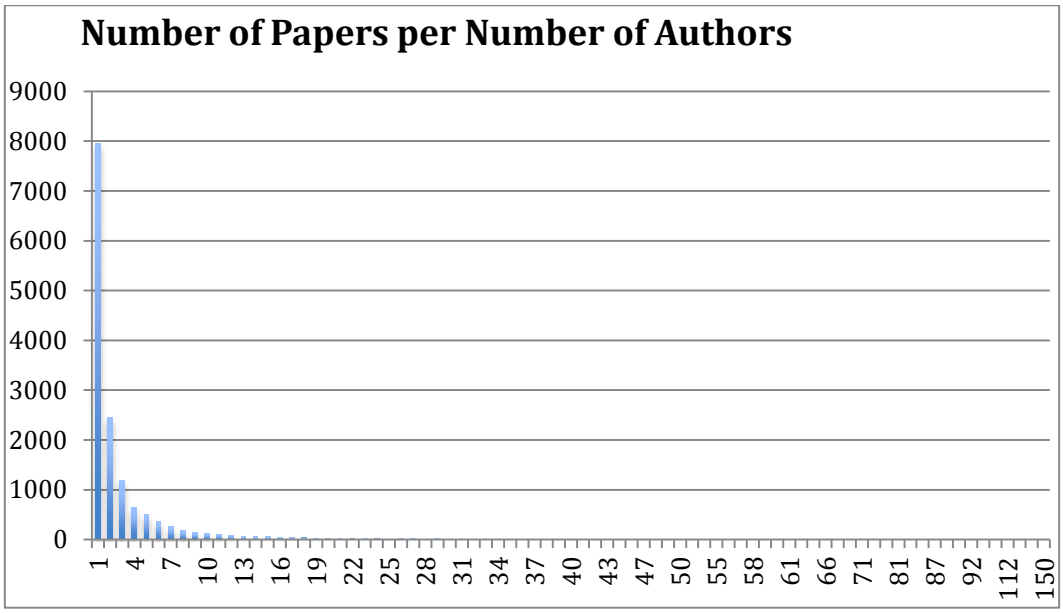


Figure 25. *Number of Papers per Number of Authors*

5 authors published more than 100 papers: Shrikanth NARAYANAN (150 papers), Keikichi HIROSE (138), John H.J. HANSEN (121), Hermann NEY (112), Kiyohiro SHIKANO (105).

300 authors published 20 papers or more, and about 1,000 published 10 papers or more, while 7,960 published only 1 paper (Fig. 25).

2.3.9. Factions or cliques

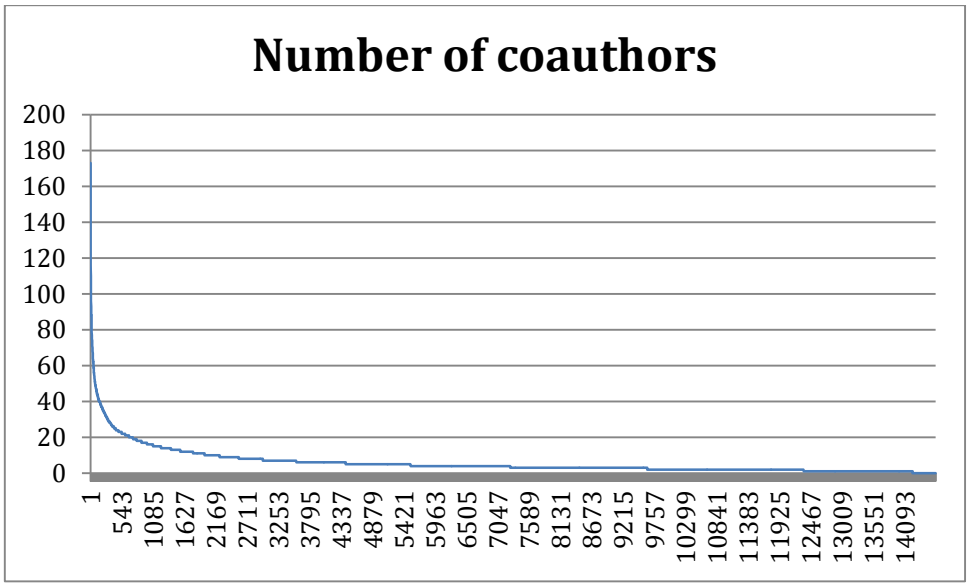


Figure 26. *Number of coauthors*

8 authors published with 100 or more different co-authors: Shrikanth NARAYANAN (170 co-authors), KiyohiroI SHIKANO (131), Keikichi HIROSE (118), Louis BOVES (116), Satoshi NAKAMURA (116), Hermann NEY (110), Alex WAIBEL (106), Frank SOONG (100), while 400 authors only published alone (Fig. 26).

We have only done a very preliminary study of the structure of the publishing communities inside ISCA. The study of the cliques, i.e. publishing groups of authors, extracted from the coauthor graph that links two author nodes when they have published a paper in common, results in 964 cliques. The largest one regroups 12,305 authors, which means that 84 % of the ISCA authors are somehow connected through a publication path, a good indicator of the cohesion of the community. The second largest clique contains only 24 authors who never published with any of the 12,305 previous ones. Only 2% of the authors have published alone.

2.4. Citations

We studied the citations only when the content of the paper was accessible in its digital form, i.e. for the conferences from 1996 to 2012 (17 years). It appeared that the papers were cited in many different ways, lacking homogeneity. We decided to consider the title of a paper as the most reliable identifier of a paper. However, it appeared that even the titles might be cited in different ways, especially when using acronyms. We therefore tried to compare titles with two different methods: either by simplifying the titles, of by comparing the titles using the Levenshtein distance, and taking a decision on the basis of the level of similarity But we faced a heavy computation time and did not complete this analysis yet. We therefore only considered for the time being the cited authors and the cited sources. In both cases, we had to extract the information from the xml document using a xpath query. We only focused our attention on the citations considered as valid by ParsCit. We extracted the cited authors from the "author" tag included in the ‘authors” tag, and the cited sources from two different tags: "booktitle" and "journal". The next step was to conduct a normalization process to handle the problem of different abbreviations for the same name of an author or of a source. Regarding the analysis of cited authors, we used the normalization table that we already used for normalizing the authors' names, with some specific additions adapted to these data. Concerning the cited sources, the table was realized manually in a tedious bootstrapping process, using first the names extracted from a random year (2007). Then, this table was extended with the most frequent variants across the 17 years. The cited sources' normalization table also contains the information of the category of the source: conference, workshop, journal, book or association. These normalization processes enabled us to make more realistic rankings.

2.4.1. Number of citations by papers overall and over time

Year	# papers with detected citations	# detected citations	Citations per paper	# detected cited authors	Cited authors per paper	Cited authors per citation
1996	400	3,284	8.21	7,112	17.78	2.17
1997	336	2,713	8.07	5,907	17.58	2.18
1998	668	5,097	7.63	11,353	17.00	2.23
1999	510	3,771	7.39	8,411	16.49	2.23
2000	659	5,084	7.71	11,233	17.05	2.21
2001	450	3,966	8.81	9,264	20.59	2.34
2002	445	3,404	7.65	7,143	16.05	2.10
2003	519	4,116	7.93	8,521	16.42	2.07
2004	635	5,764	9.08	13,512	21.28	2.34
2005	698	5,981	8.57	12,975	18.59	2.17
2006	632	5,674	8.98	13,459	21.30	2.37
2007	730	8,202	11.24	20,170	27.63	2.46
2008	703	7,293	10.37	18,442	26.23	2.53
2009	732	8,710	11.90	22,170	30.29	2.55
2010	744	9,047	12.16	24,106	32.40	2.66
2011	821	9,987	12.16	26,520	32.30	2.66
2012	667	8,387	12.57	22,688	34.01	2.71
Total	10,349	100,480		242,986		
Mean		5,911	9.71	14,293	23.52	2.46
Maximum			63		169	22

Table 4. *Number of bibliographical references and cited authors over time*

We detected at least one bibliographical reference in 10,349 of the 12,890 articles published from 1996 to 2012. We found 100,480 bibliographical reference in those 10,349 papers, i.e. 5,911 references in a conference on average. This means that there are about 10 references per paper on average, with a maximum of 63. The average number of bibliographical references per paper raised from 8.21 in 1996 to 12.57 in 2012 (Table 4).

2.4.2. Number of citations by authors overall and over time

We detected at least one author cited in 10,332 of the 12,890 articles published from 1996 to 2012, and 98,877 of the 100,480 bibliographical references cite at least one author. We found 242,986 authors citations in the bibliographical references, i.e. 14,293 cited authors in a conference on average. This means that there are about 24 cited authors per paper on average (with a maximum of 169), and 2.46 cited authors per article on average (with a maximum of 22). The average number of cited authors in a bibliographical reference raised from 2.17 in 1996 to 2.71 in 2012, following with some delay the general trend of the increase of the number of co-authors of a paper.

2.4.3. Most cited authors

After a tedious cleaning process which fortunately benefited from the existence of the ISCA Archive conference series cleaned authors' list, the 242,986 cited authors correspond to 50,653 "different" authors.

The 10 most cited authors get 700 citations or more. Those are: Phil Woodland (1066), Steve Young (993), H. Ney (977), Doug Reynolds (954), Chin-Hui Lee (89), Andreas Stolcke (814), Alan Black (796), Keiichi Tokuda (792), Hynek Hermansky (707) and Shrikanth Narayanan (700). Most of those 10 most cited authors get a high ranking over the considered 17 years, apart from Doug Reynolds who appears in the 50 most cited authors only since 2000 and became the most cited author in 2011. We see that the list of cited authors is very sparse (Fig. 27).

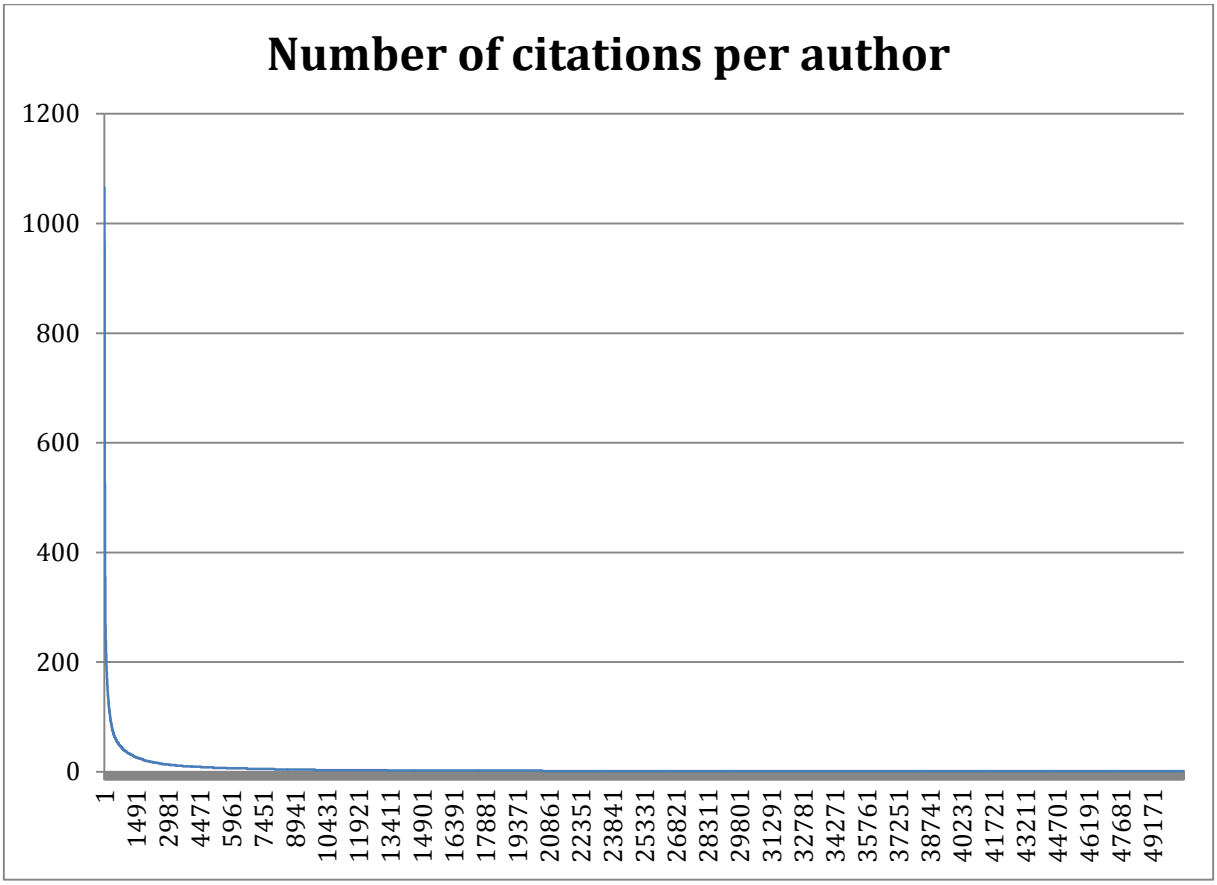


Figure 27. *Number of citations per author*

2.4.4. Most cited sources

Here also, the conferences, journals and books are cited in many different ways, and a tedious cleaning process had to be conducted in order to identify the sources. Only the sources cited more than once were considered in this cleaning process.

The 20 most cited conferences or workshops are given in Fig. 28. IEEE ICASSP is followed by ICSLP, Eurospeech, Interspeech, the Automatic Speech Recognition and Understanding (ASRU) workshop, the International Congress of Phonetic Sciences (ICPHS) and the Language Resources and Evaluation Conference (LREC).

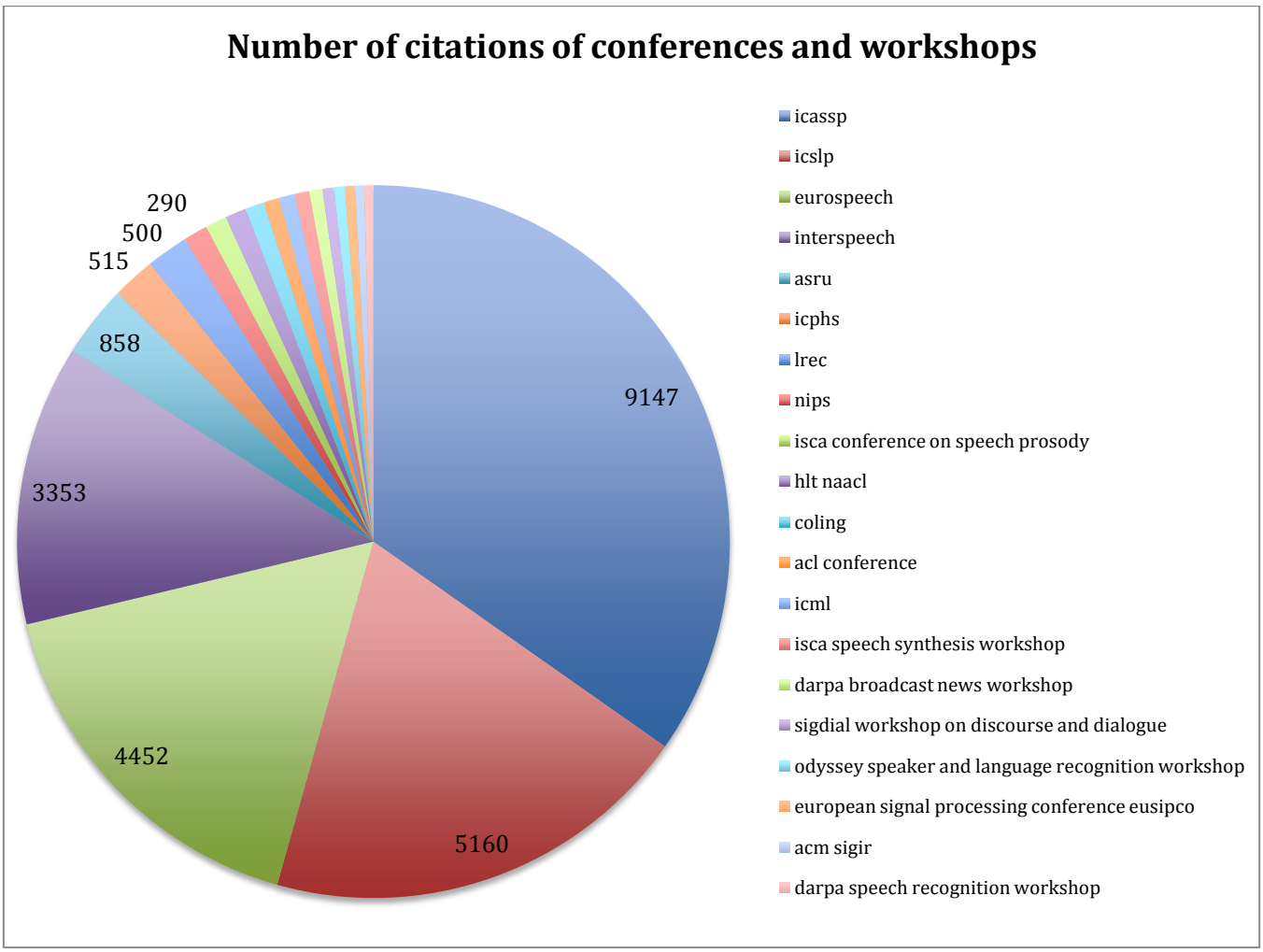


Figure 28. *20 most cited conferences or workshops (1996-2012)*

The 20 most cited journals or books are given in Fig. 29. The Journal of the Acoustical Society of America (JASA) ranks first, given that it also serves as the Proceedings of the ASA Conference series. It is followed by the Speech Communication Journal, the various avatars of the initially called IEEE Transactions on Acoustic, Speech and Signal Processing, and Computer Speech and Language.

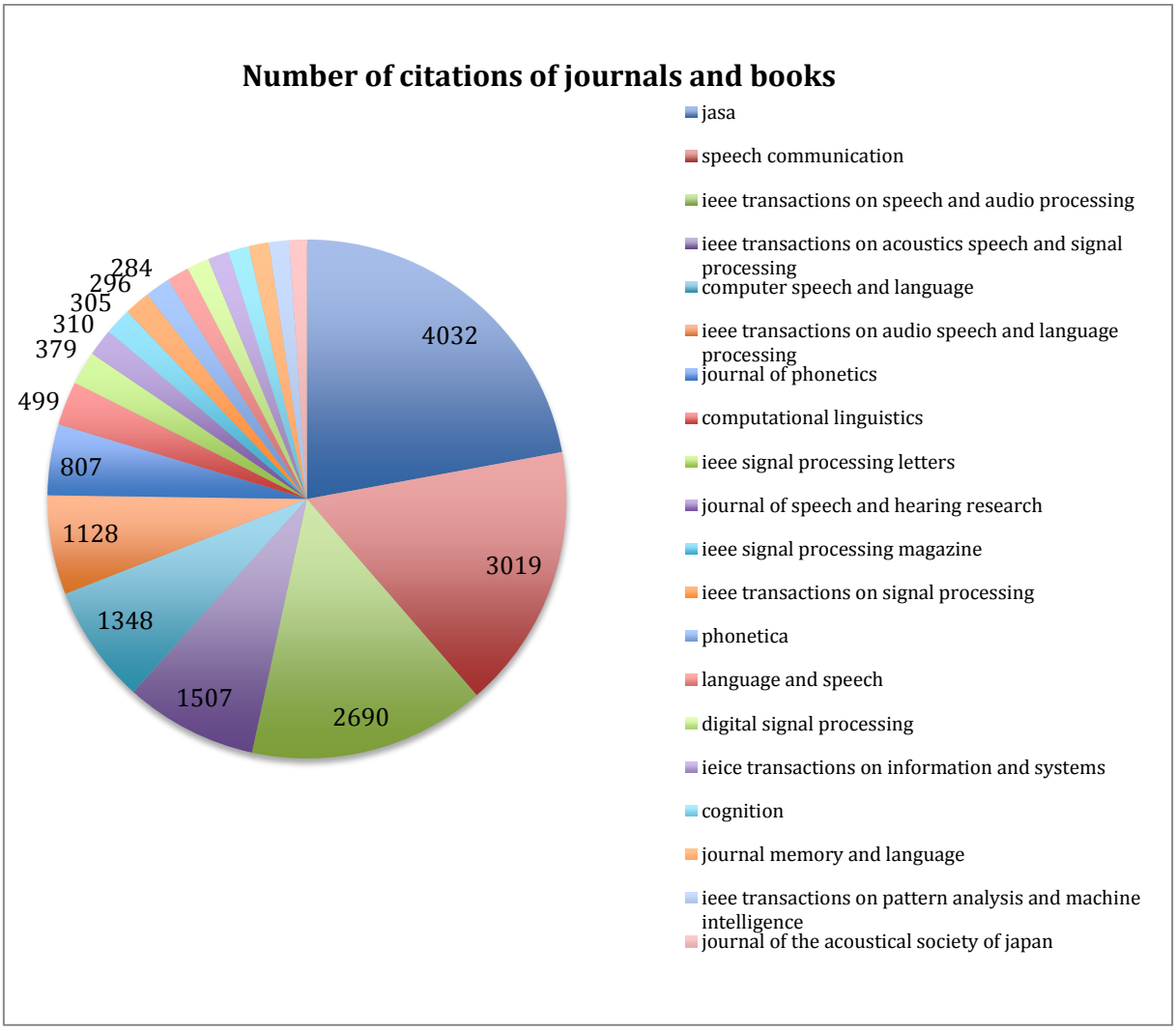


Figure 29. *20 most cited journals or books (1996-2012)*

If we now merge conferences which are related such as ECST, Eurospeech, ICSLP and Interspeech, or the various HLT conferences held under various umbrellas, and Journals which are related such as the IEEE Transactions on Acoustic, Speech and Signal Processing, which successively became Transactions on Speech and Audio Processing, and now Transactions on Speech, Audio and Language Processing, we find that the 6 more cited sources are in ranked order the ISCA conference series (12,965 citations), ICASSP (9,147), the IEEE Transactions series (5,325), the JASA (4,032), Speech Communication (3,019) and Computer Speech and Language (1,348) (Fig. 30).

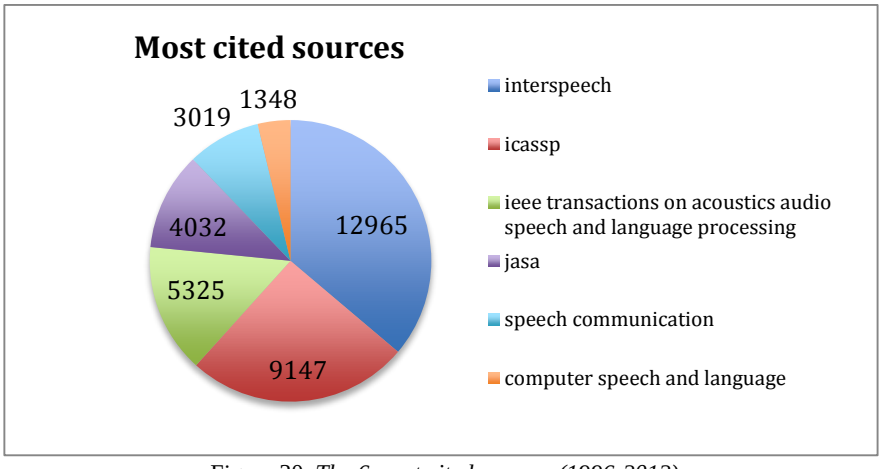


Figure 30. *The 6 most cited sources (1996-2012)*

We then considered those 6 most cited sources and studied their evolution over the 1996-2012 period (Fig. 31). It appears that their relative ranking is stable over time, and that all show a large increase, by a factor of 2 to 4, in that period.

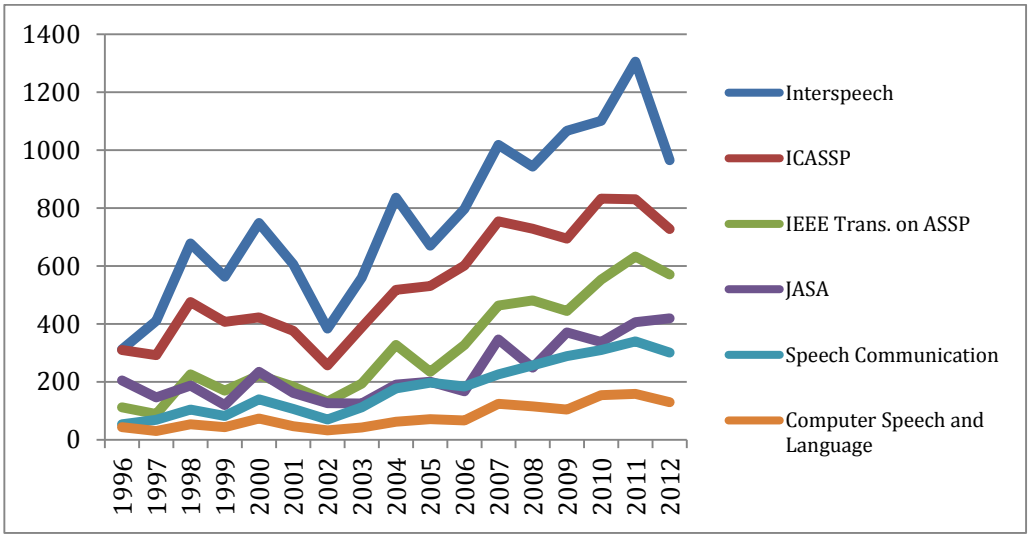


Figure 31. Evolution of the 6 most cited sources over time (1996-2012)

2.4.5. Most cited Funding Agencies

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

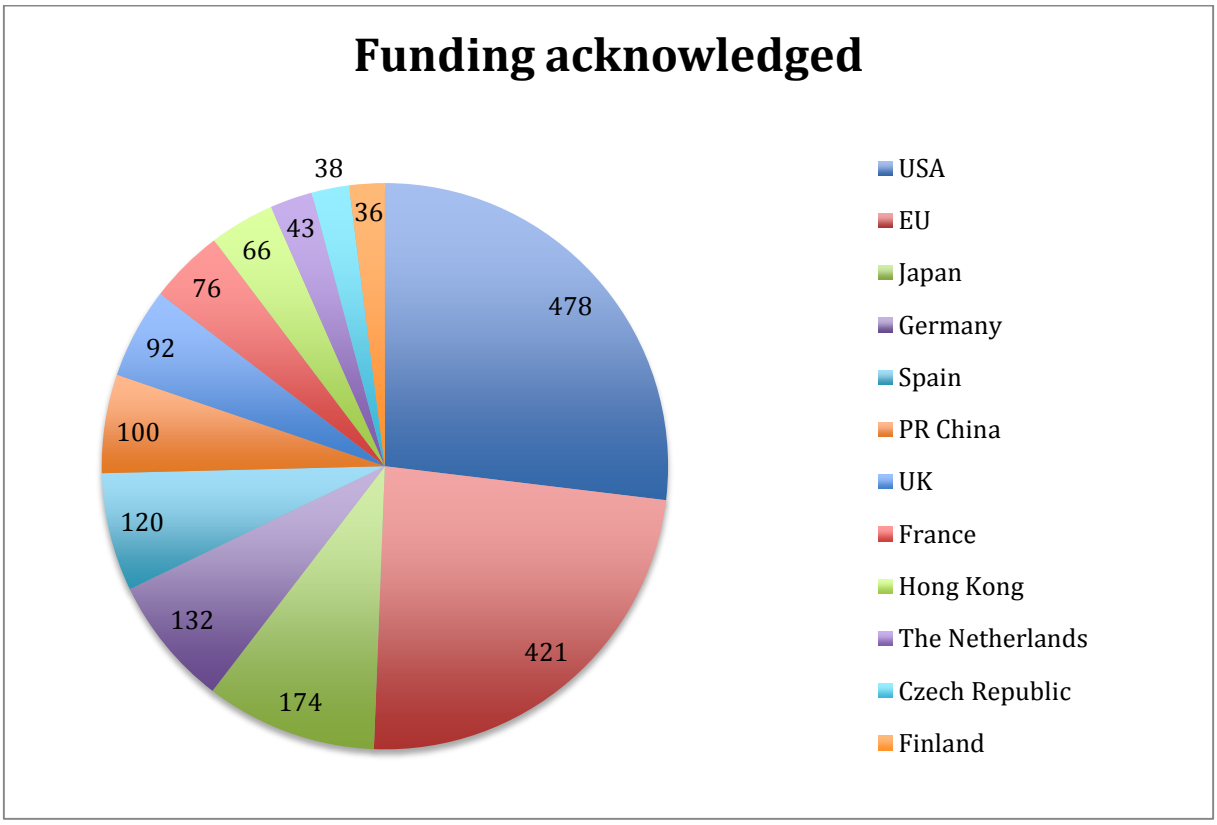


Figure 32. Share of the funding acknowledgement for the 12 most cited countries

If we consider the 12 most cited countries (Fig. 32), we see that the USA are ranked first, but that the European Union at the Community level is close. Japan comes third, followed by Germany, Spain, PR China, UK and France.

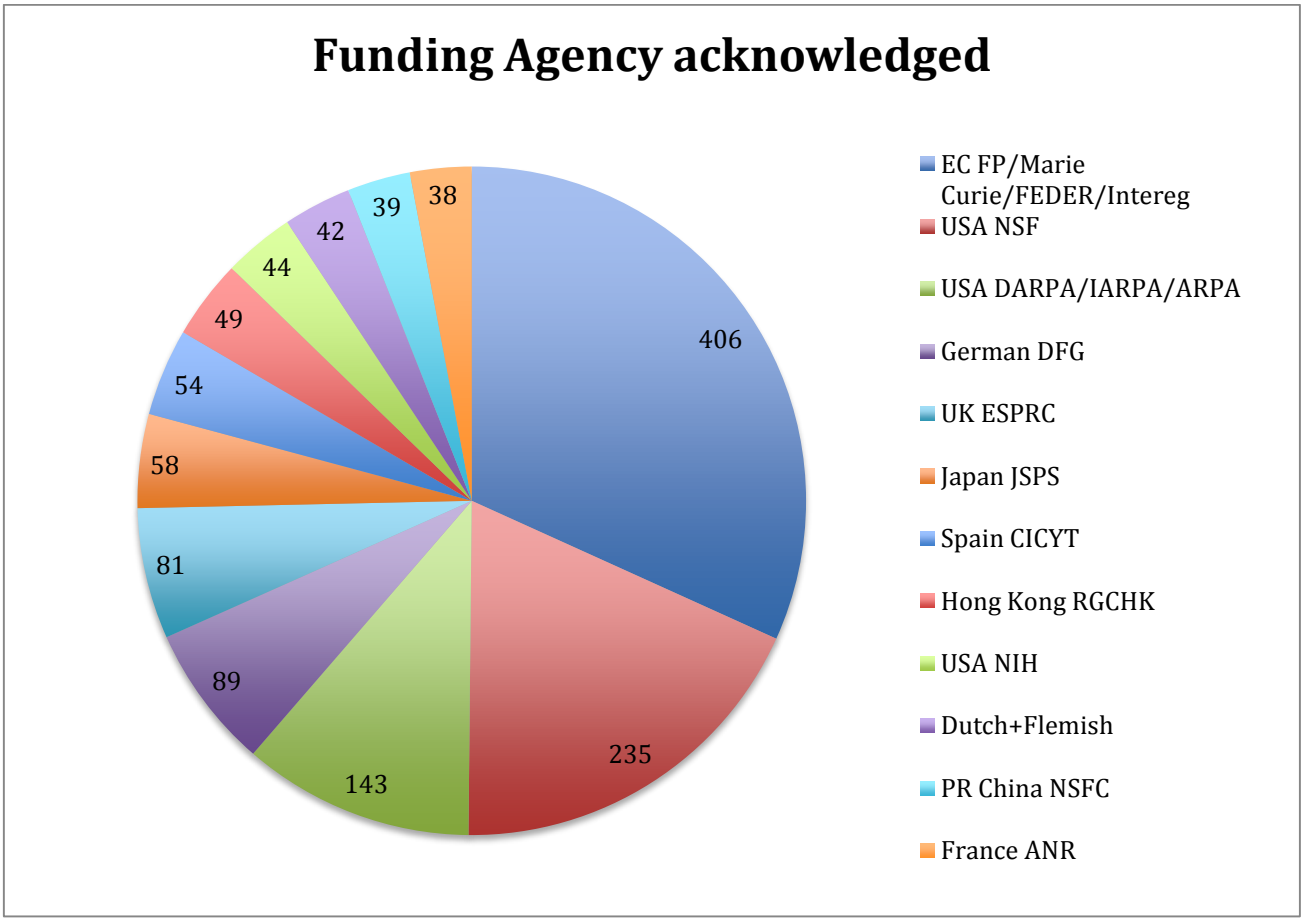


Figure 33. Share of the funding acknowledgement for the 12 most cited agencies

If we now consider the 12 most cited agencies (Fig. 33), we see that the European Commission comes first, with its various programs, followed by the US NSF and DoD DARPA, ARPA and IARPA Agencies. Interestingly, we see that the US NIH (National Institutes of Health) also appear in this chart, and we mention the joint efforts of The Netherlands and Belgium Flemish governments to support the Dutch language.

2.5. Topics

2.5.1. Term based topic analysis

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

Just as for the study of citations, our initial input is the textual content of the papers, which is only available from 1996 to 2012. Over these 17 years, the archives contain a grant total of 241,232,235 English words.

In order to identify the main topics, we used two approaches. First, in a top down approach, we started from the index of a book on Spoken Language Processing [20]. However, this index may not include all the terms related to speech communication, and may not include the most recent terms, which appeared after its year of publication (2009). We therefore then tried a bottom up approach.

As our aim is to study the terms of the Spoken Language Processing domain, we do not want to get noise from some frequent formula "ordinarily" used in the English language. For this purpose, as a first step, we processed a vast amount of "ordinary" English texts in order to compute a statistical language profile. More precisely, we applied a deep syntactic parser called TagParser (www.tagmatica.com) [14] and got the noun phrases. For each sentence, we kept only the noun phrases with a plain noun as a head, thus excluding the situations where a pronoun, a date or a number is the head. We also made a special dispatching for co-ordinations. We retained the various combinations of sequence of adjectives, prepositions and nouns

excluding initial determiners according to unigrams, bigrams and trigrams sequences, and we stored the result on the hard-disk. This process was applied on a corpus gathering the British National Corpus (aka BNC) (www.natcorp.ox.ac.uk) [15], the Open American National Corpus (aka OANC (www.americannationalcorpus.org)) [16], the Suzanne corpus release-5 (www.grsampson.net/Resources.html), the English EuroParl archives [17] (years 1999 until 2009) (www.statmt.org/europarl), plus a small collection of newspapers in the domain of sports, politics and economy. The total of words was 200M words. It should be noted that, in selecting this corpus, we took care to avoid any text dealing with Spoken Language Processing.

This statistical language profile being recorded on disk, we were ready to process the second step which was to parse the ISCA Archive and, with the same filter, to compute the difference. In other words, we made the hypothesis that when a sequence of words is INSIDE the ISCA archive and NOT INSIDE the "ordinary" profile, we consider that this term is specific to Spoken Language Processing.

The twenty most frequent terms in Spoken Language Processing were computed over the period of 17 years, with the following strategy. First, the most frequent terms were computed in a raw manner, and secondly the synonyms sets (aka synsets) for all most 50 frequent terms of each year (which are frequently the same from one year to another) were manually declared in the lexicon of TagParser. Around the term synset, we gathered the variation in upper/lower case, singular/plural number, US/UK difference, abbreviation/expanded form and absence/presence of a semantically neutral adjective, like "artificial" in "artificial neural network". Thirdly, the most frequent terms were recomputed with the amended lexicon. This processing took 4 hours on a mid-range workstation (a Dell Precision workstation based on a single Xeon E3-1270V2 with 32 Gb of RAM) and gave the results that follow.

2.5.2. Most frequent terms (based on content analysis)

The 20 most frequent terms (lemmas) over time (1996-2012) are the following (Table 5):

Term synset	# Occurrences	Frequency
HMM: HMM(s), Hidden Markov Model(s)	19688	0.79
SR: SR(s), ASR(s), Automatic Speech recognition(s), Speech Recognition(s)	18290	0.74
LM: LM(s), Language Model(s)	16985	0.68
GMM: GMM(s), Gaussian Mixture Model(s)	11226	0.45
Recognizer: recogniser(s), recognizer(s)	10324	0.42
Segmentation: segmentation(s)	8907	0.36
Modeling: modeling(s), modelling(s)	8838	0.36
Classifier: classifier(s)	8211	0.33
WER: WER(s), Word Error Rate(s)	8055	0.32
MFCC: MFCC(s), Mel Frequency Cepstral Coefficient(s)	7562	0.30
Formant: formant(s)	7281	0.29
Normalization: normalization(s), normalisation(s)	6726	0.27
Dialog: dialog(s)	6107	0.25
SNR: SNR(s), Signal Noise Ratio(s)	5650	0.23
SVM: SVM(s), Support Vector Machine(s)	5372	0.22
Lexicon: lexicon(s), lexica	4825	0.19
Prosody: prosody	4776	0.19
Neural Network: NN(s), ANN(s), Neural Network(s), Artificial Neural Network(s), NeuralNet(s)	4603	0.19
Lattice: lattice(s)	4415	0.18
Covariance: covariance(s)	4208	0.17

Table 5. *20 most popular terms overall*

2.5.3. Change in Topics

We studied the ranking among the 50 most popular terms (mixing unigrams, bigrams and trigrams) representing several topics of interest. The terms are followed by their ranking in 1996 and 2012 (R1996/R2012).

Keywords remaining popular (Fig. 34): We studied in this category the following keywords, which stayed in the 5 top over 15 years: HMM (1/2), SR (2/3), LM (4/4).

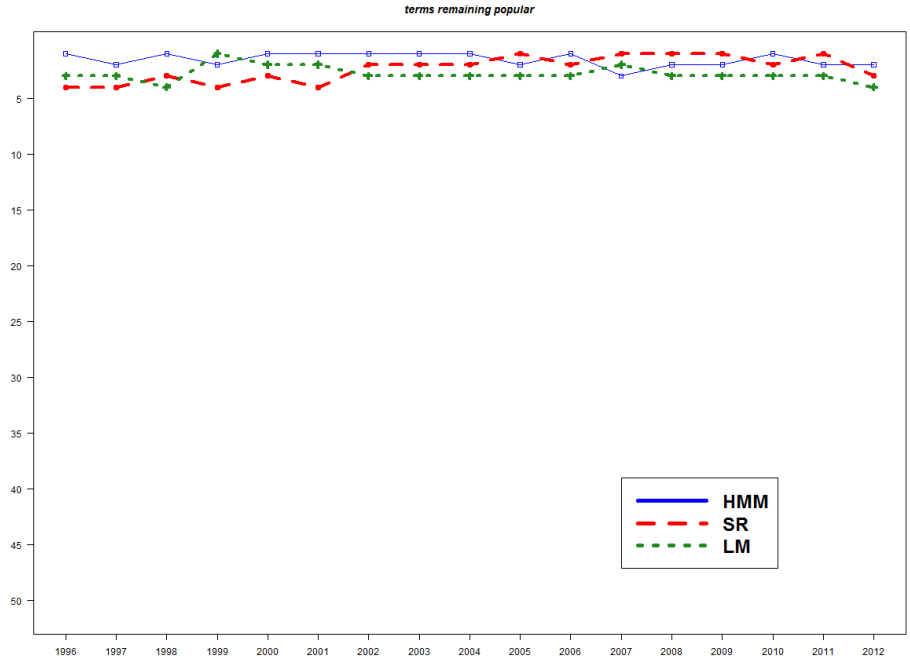


Figure 34. *Terms remaining popular*

Keywords becoming popular (Fig. 35): We studied in this category the following keywords, which became more and more popular over time: GMM (Gaussian Mixture Model) (less than 50/1), Classifier (less than 50/5), WER (Word Error rate) (31/7), SVM (Support Vector Machine) (less than 50/11), Normalization (21/6), Ngram (less than 50/27).

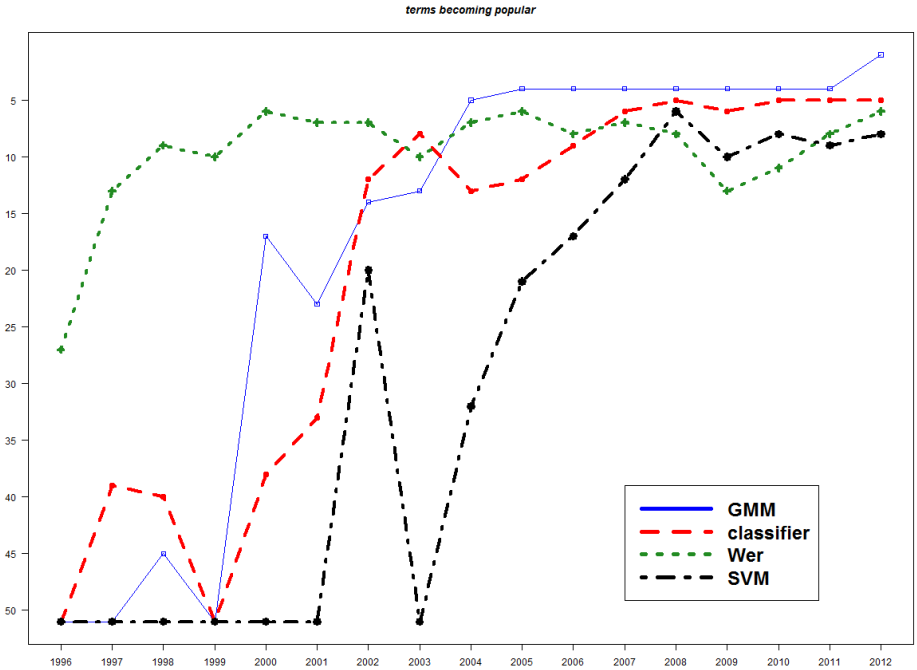


Figure 35. *Terms becoming popular*

Keywords losing popularity (Fig. 36): We studied in this category Codebook (22/less than 50) and Perplexity (11/Less than50).

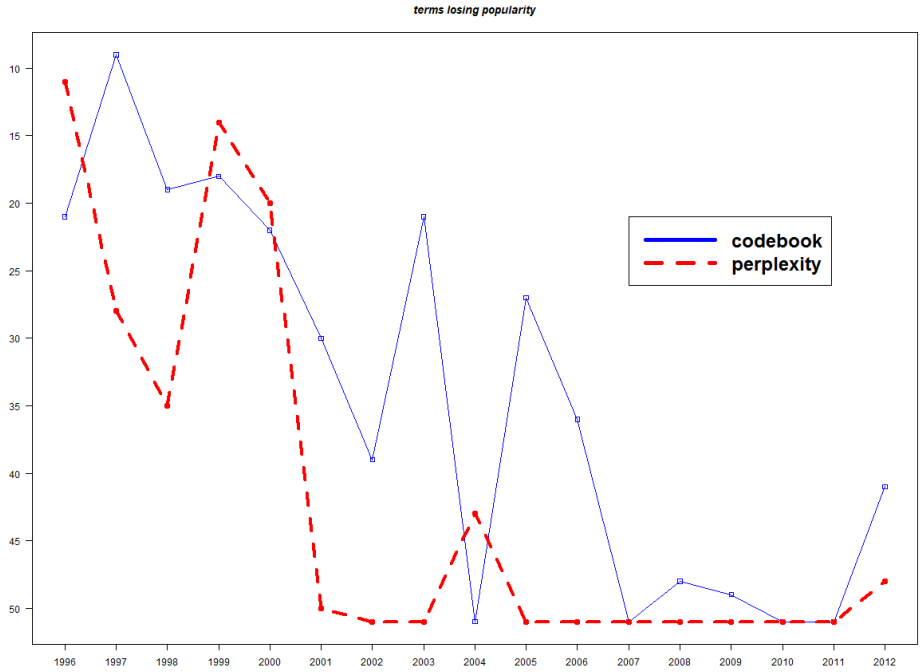


Figure 36. *Terms losing popularity*

We studied especially the disappearing of the terms “bigram” and “trigram” replaced by “ngram” (Fig. 37): Bigram (8/less than 50), Trigram (28/less than 50), Ngram (less than 50/27)

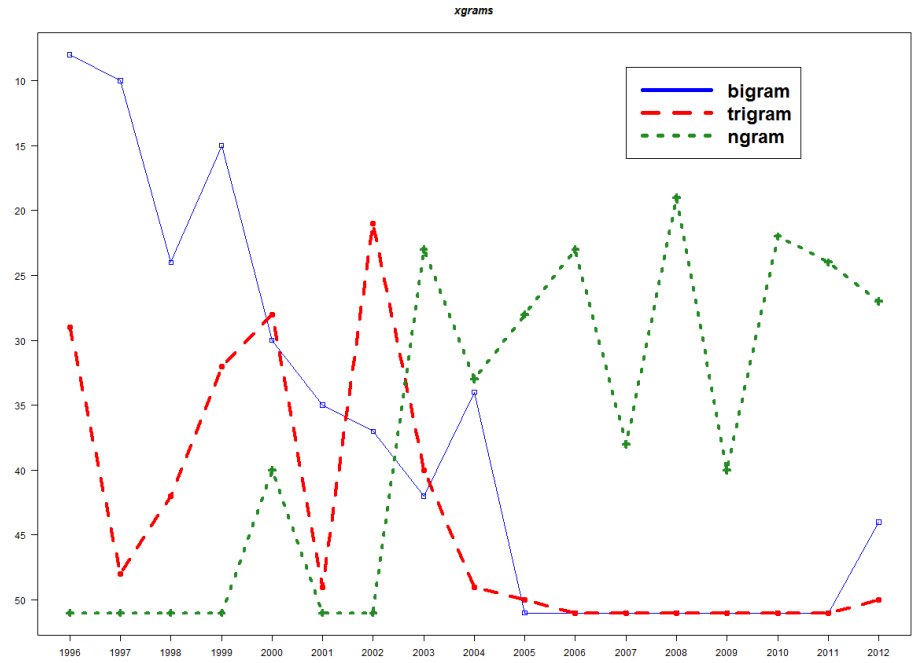


Figure 37. *Comparison of bigram, trigram and ngram over time*

Keywords fluctuating (Fig. 38):

We studied in this category terms that stayed popular over time but possibly with a different meaning, or terms which show a large fluctuation. Here, the terms are followed by examples of extreme ranking changes over time: Segmentation (5/13/10), Normalization (21/8/18/6), Formant (6/23/17), Dialog (33/5/23/5/34), Prosody (23/10/39/12/37). The term Neural Networks (10/37/17), which was popular by the end of the 90s, lost its popularity in the early 2000s and recently regained popularity.

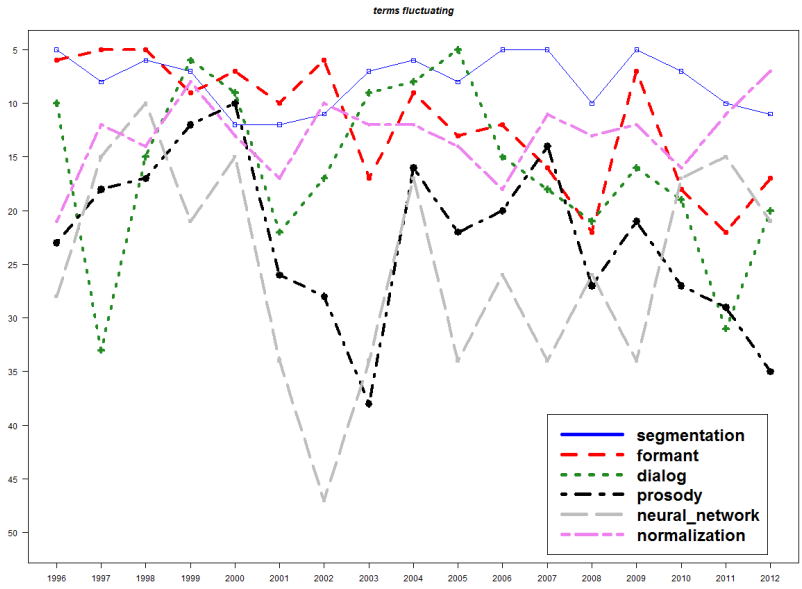


Figure 38 *Terms fluctuating*

The http://vernier.frederic.free.fr/Infovis/rankVis/ site provides an interactive "Top term visualization of ISCA conference between 1996 and 2012 » application allowing to explore the landscape of the 50 most popular terms over 1996-2012.

2.5.4. Specific study on the "17-year friends" of the "becoming popular" terms

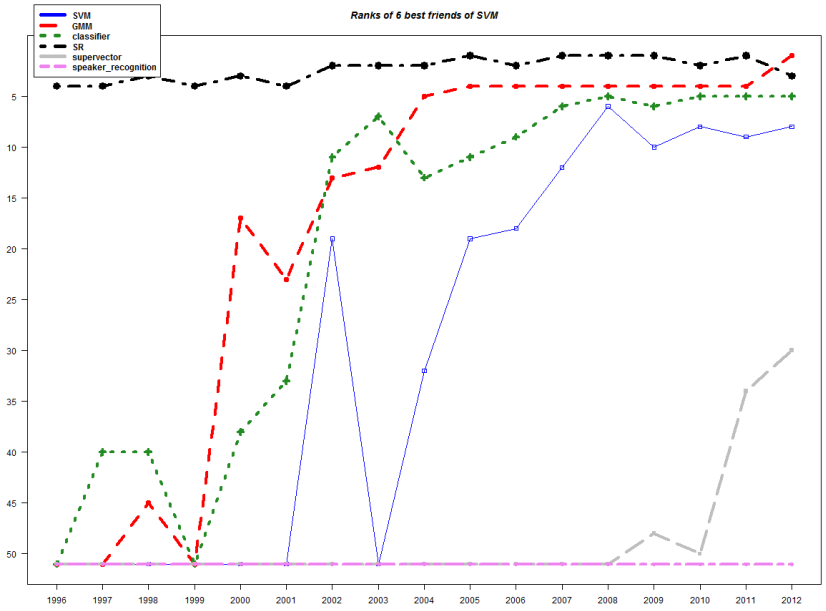
A selection of terms has been studied with respect to both their time-behavior and semantically closeness. The aim is to detect trends and related properties between the terms of the domain.

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

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

The case of "Support Vector Machines" (SVM):

Such a computation gives the following diagram, with "SVM" as a focus and with a display limit of 6 terms to ease the reading:

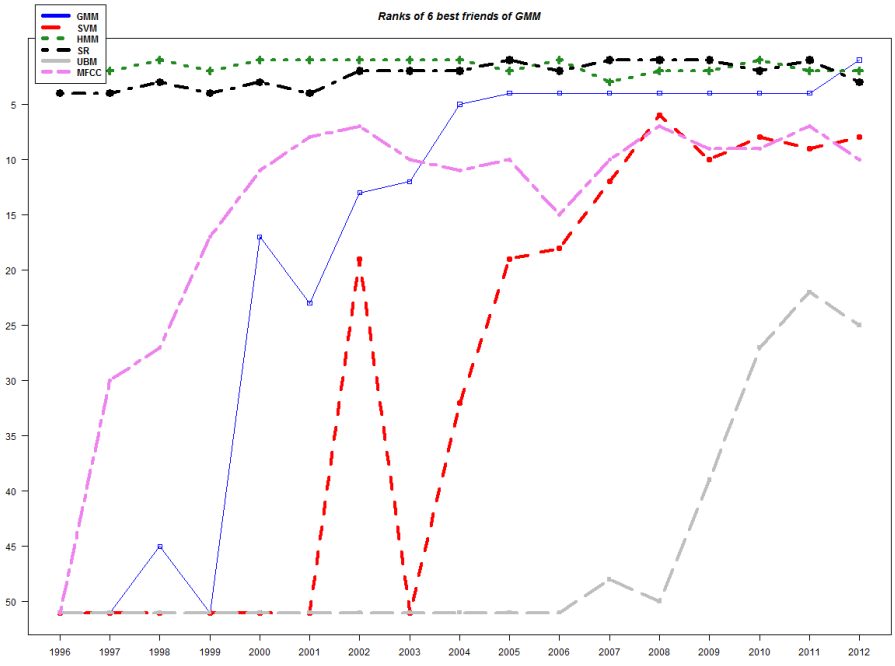


We can make the three following comments:

- The term "*supervector*" (in light gray) seems to be a weak signal. With respect to "SVM", "supervector" is both a friend and has a similar curve. In other words, "supervector" appears in the same abstracts and, like "SVM", it becomes more and more popular, although more recently.
- The terms "GMM" and "classifier" have more or less the same level of popularity and have also the same curve shape than "SVM".
- The term "SR" is very popular from 1997 to 2012 but it is also a friend of SVM. In contrast, the terms "HMM" and "LM" do not appear on this diagram but they are globally constantly popular on the whole period, as shown on the figure "Terms remaining popular" in the previous section. This means that "SR" is deeply related to "SVM" and shared by a lot of other terms. On the contrary, "HMM" and "LM" are shared by a lot of papers but not deeply related to "SVM".

The case of "Gaussian Mixture Models" (GMM):

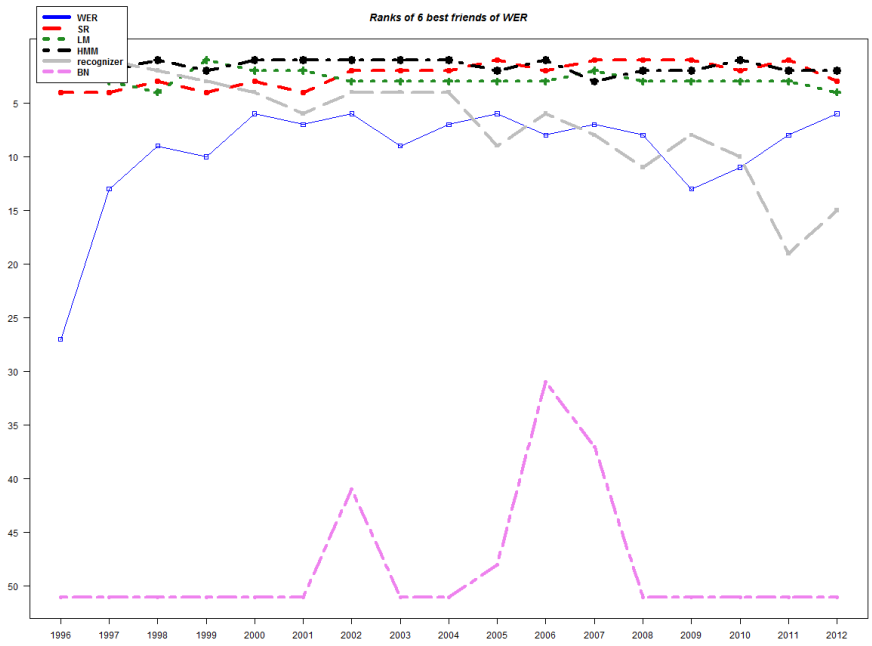
With "GMM" as the focus, the diagram is as follows:



The terms "HMM" and "SR" are continuously popular and "friends" of "GMM". The diagram shows a weak signal in the recent raising of the "*Universal Background Model*" (UBM), while the presence of the "*Mel-frequency cepstral coefficients*" (MFCC) is normal .

The case of "Word Error Rate" (WER):

With WER as the focus, the diagram is as follows:



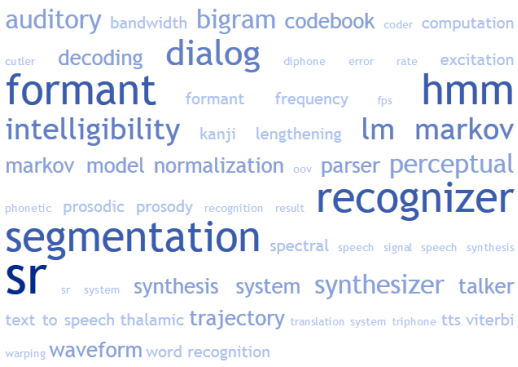
This diagram exhibits the increasing importance of evaluation and Word Error Rates (aka WER) over time and picks related to "BN" (*Broadcast News*), the Broadcast News campaigns organized by NIST for DARPA in 1996-1999.

2.5.5. *Tag Clouds for frequent terms*

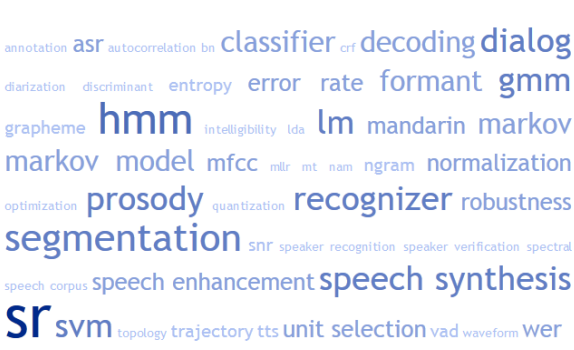
The aims of this current section is to have a global estimation of the main terms of a specific year and to have an idea of the stability of the terms over the years. The line-based diagrams presented in the previous section allow for a fine grain presentation but they do not permit a global view. For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

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

Results:

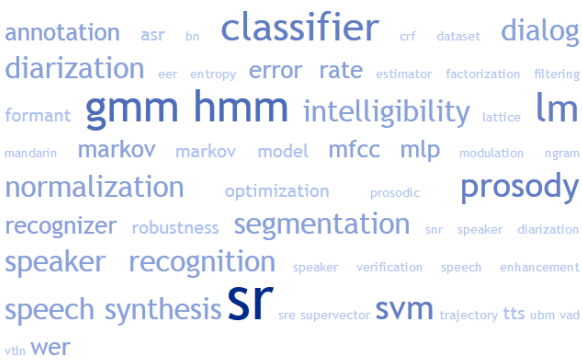


Tag Cloud based on the 1996 abstract

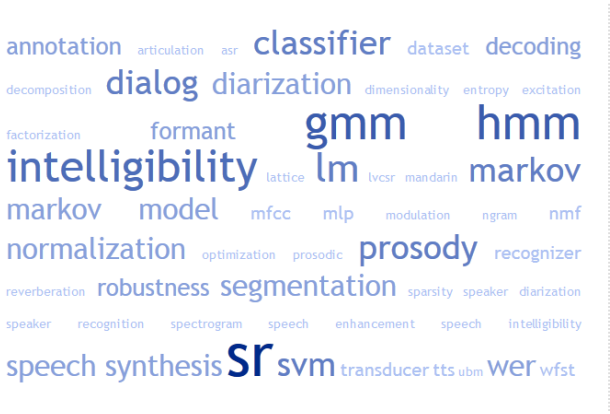


Ten years latter, Tag Cloud based on the 2006 abstracts

Tag Cloud based on the 2010 abstracts



Tag Cloud based on the 2011 abstracts



Tag Cloud based on the 2012 abstracts

Globally, it appears that no big change has occurred in the most frequent terms mentioned across the years, and the "pictures" look similar. Progress has therefore be steadily constant, without any big "conceptual rupture" in the period.

We see in the period between 1996 and 2006 the already mentioned disappearing of "Bigram" and "Codebook", and the stronger presence of "classifier", "Prosody" and "GMM".

We can notice that the terms "formant" and "recognizer", which were rather popular in the first years (i.e. 1996) are less popular in the recent years (i.e. 2010, 2011 and 2012). The three clouds of the recent years are rather similar, which means that the terminology of the domain over this period is quite stable.

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

The objectives were twofold. First, we wanted to study whether or not it is possible to facilitate the automatic clustering of papers into a limited number of sessions based solely on the parsing of the content. Secondly, we wanted to exhibit possible hidden links between apparently unrelated papers.

Our process relies on the same terminological extraction as the previous sections. Let's recall that this extraction computes the terms of the domain from the difference between a statistical profile of "ordinary" English templates (recorded on a disk) and the syntactic patterns of the papers of the conference. Once the terms are collected, the TF-IDF of each term is computed. Without entering into mathematical details, let's say that the TF-IDF value reflects how important a term is to represent a document within a corpus (see http://en.wikipedia.org/wiki/Tf-idf for details) [21]. A consequence of this computation is that the popular terms over the whole conference (like HMM, for instance) do not have a high TF-IDF value: only specific terms have a high value.

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

Then, we considered these salient terms as the representation of the paper and from these terms, we automatically clustered the papers using a hierarchical clustering algorithm (UPGMA) using the cosine similarity between papers. Once each cluster is built, the terms of the clusters are ranked according to their TF-IDF in order to get a list of terms that are representative of the cluster (see "clusterTerms" in the following table). The clustering process gives the following result.

CLUSTER#1 Size=5 clusterTerms=cue trading, pronunciation map, MCEPs, knowledge integration, AF value		
	An Information-Extraction Approach to Speech Analysis and Processing	paperTerms=knowledge integration, stage by stage, Safra, bank of detector, ASAT
	Articulatory Feature based Multilingual MLPs for Low-Resource Speech Recognition	paperTerms=AF detector, AF, language phone, ASAT, system combination system
	Modeling Cue Trading in Human Word Recognition	paperTerms=cue trading, AF value, GOF, AFS, AF
	Modeling a Noisy-channel for Voice Conversion Using Articulatory Features	paperTerms=MCEPs, AFS, amount of speaker, information in AFS, mapper
	Real-time Visualization of English Pronunciation on an IPA Chart Based on Articulatory Feature Extraction	paperTerms=pronunciation map, a_J, MLNs, value of AF, AFS

CLUSTER#2 Size=4 clusterTerms=NNLM, hNNLM, RNNLM, word prediction accuracy, treebank corpus		
	Conversion of Recurrent Neural Network Language Models to Weighted Finite State Transducers for Automatic Speech Recognition	paperTerms=RNNLM, treebank corpus, discretization, pruning criterion, one of ngram
	Improving WFST-based G2P Conversion with Alignment Constraints and RNNLM N-best Rescoring	paperTerms=alignment lattice, tom arc, foreach, RNNLM, bptt
	Large Scale Hierarchical Neural Network Language Models	paperTerms=NNLM, hNNLM, vocabulary word, RNNLM, unpruned ngram LM
	Towards Recurrent Neural Networks Language Models with Linguistic and Contextual Features	paperTerms=word prediction accuracy, RNNLM, lemma, php, perplexity of RNNLM
CLUSTER#3 Size=4 clusterTerms=DTNN, WFST DNN, tensor layer, NNR, DNN		
	A Initial Attempt on Task-Specific Adaptation for Deep Neural Network-based Large Vocabulary Continuous Speech Recognition	paperTerms=NNR, sti, DNN, LBP, CD DNN
	Boosting Classification Based Speech Separation Using Temporal Dynamics	paperTerms=Struct, DNN, previous IBM, multi-layer perceptrons, neigh[1]
	Integrating Deep Neural Networks into Structured Classification Approach based on Weighted Finite-State Transducers	paperTerms=WFST DNN, DNN, DNN model, MTE, arc sequence
	Large Vocabulary Speech Recognition Using Deep Tensor Neural Networks	paperTerms=DTNN, tensor layer, DNN, tensor, DP layer
CLUSTER#4 Size=3 clusterTerms=Cllr, Ecrps, Enlpd, WPPCA, calibration of score		
	Age Estimation from Telephone Speech using i-vectors	paperTerms=WPPCA, SVR, age estimation, GMM WPPCA, PCA SVR
	Calibration of probabilistic age recognition	paperTerms=Ecrps, Enlpd, SVR, Cllr, distribution over age
	The Role of Score Calibration in Speaker Recognition	paperTerms=Cllr, calibration of score, ranking of system, miscalibration, calibration for system
CLUSTER#5 Size=3 clusterTerms=SSANOVA, ND, frequency word, neighborhood density, trajectory between vowel		
	Convolutive Non-Negative Sparse Coding and New Features for Speech Overlap Handling in Speaker Diarization	paperTerms=CNSC, base activation, OIP, LLK, overlap detection
	Heterogeneous Convolutive Non-Negative Sparse Coding	paperTerms=CNSC, convolution range, separation task, uniformed, CNSC algorithm
	Speaker Diarization of Overlapping Speech based on Silence Distribution in Meeting Recordings	paperTerms=sli, SPI, ovi, overlap labelling, overlap detection

[1] *This is due to a bug in the management of hyphenation.*

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

The three first clusters put together several papers dealing respectively with "*Articulatory Features*" (AF), "*Recurrent Neural Networks Language Models*" (RNNLM) and "*Deep Neural Networks*" (DNN), even if the terms do not appear

The two last ones show the ability to establish a sort of chain made of shared terms, which transitively, term by term, links some papers together. For instance, in the fourth cluster, the first two papers share the topic of "*age recognition*", while the two last ones share the topic of "*calibration*", which could be of interest for the first one. Similarly, in the fifth cluster, the first and third papers share the topic of "*speech overlap*", while the two first papers share the topic of "*sparse coding*", which could be of interest for the second one.

3. Perspectives

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

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

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

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

It will allow to check whether the domains of interest depend on the author's gender, and to study the changes in the topics of interest for authors or factions.

4. Conclusions

In this analysis exercise, we faced a great difficulty in the use of the available data. Most of the oldest information could not even be used, because it is not available in a machine-readable format easily convertible into text. This resulted in the fact that whereas we've been able to use all the 25 years conference series metadata, we've only been able to use the paper content since 1996. Whereas metadata is freely available online, the content of the papers is only accessible to ISCA members, contrary to the situation at the ACL where all the 50 years ACL conference data is freely available online. ISCA may consider moving to the same policy in the time of Open, Shared and Linked Data.

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

The research in Spoken Language Processing has achieved major advances over this period through constant and steadily scientific efforts, that gained efficiency thanks to the availability of a necessary infrastructure made up of publicly funded programs, largely available language resources, regularly organized evaluation campaigns initiated in the USA by the 80s. It also very importantly benefited of a scientific social network bridging the community which increased its momentum 25 years ago with the creation of the European Speech Communication Association (ESCA) and benefited from the ECST, Eurospeech, ICSLP and Interspeech conference series to share ideas and make progress.

This preliminary analysis allowed us to extract salient or hidden information and trends which, we hope, provide a better understanding of the past 25 years of research in Spoken Language Processing worldwide. We hope it will also serve as a precious experience for building up the next 25 years.

5. Acknowledgements

The authors wish to thank the ACL colleagues, Ken Church, Sanjeev Khudanpur, Amjbad Abu Jbara, Dragomir Radev and Simone Teufel, who helped them in the starting phase, Isabel Trancoso, who gave her ISCA Archive analysis on the use of assessment and corpora, Wolfgang Hess, who produced and provided the 14 GBytes ISCA Archive, and Frédéric Vernier who kindly provided his visualization software. And all the organizers, reviewers and authors over the 25 years conference without whom this analysis could not be conducted!

6. Apologies

This survey has been conducted on textual data, which covers a 25 years long period and has therefore been encoded in different formats, which even made it too difficult to process the content of the articles anterior to 1996. The analysis uses tools, which automatically detect the various parts of scientific papers (Title, authors, affiliations, abstracts, references, acknowledgements, etc.) and may make errors. Therefore, the results should be considered as containing an error margin, and the authors wish to apologize for any errors that the reader may detect and that they will be glad to take into account in future releases of the present survey.

7. References

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

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

[3] 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

[4] J. Mariani (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

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

[6] http://swish-e.org/

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

[8] http://www-igm.univ-mlv.fr/~unitex/

[9] http://flex.sourceforge.net/

[10] David Auber, Daniel Archambault, Romain Bourqui, Antoine Lambert, Morgan Mathiaut, Patrick Mary, Maylis Delest, Jonathat Dubois, Guy Mélançon, Research Report, p31, January 2012, http://hal.archives-ouvertes.fr/hal-00659880

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

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

[13] Isaac G. Councill, C. Lee Giles, Min-Yen Kan, ParsCit: An open-source CRF reference string parsing package. In Proceedings of the Language Resources and Evaluation Conference (LREC 08), Marrakesh, Morocco, May, 2008.

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

[15] The British National Corpus, 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/

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

[17] Philipp Koehn, Europarl: A Parallel Corpus for Statistical Machine Translation,, MT Summit 2005.

[18] Yu Fu, Feiyu Xu and Hans Uszkoreit, 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.

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

[20] J. Mariani ed. (2009), Spoken Language Processing, ISTE-Wiley, 2009, ISBN 978-1-84821-031-8

[21] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, *Introduction to Information Retrieval*, Cambridge University Press. 2008., ISBN: 0521865719