Atelier Linguistique & « Big Data »

Le Traitement Automatique des Langues à l'ère du « Big Data » : le cas des publications scientifiques

Patrick Paroubek, Joseph Mariani,
Gil Francopoulo, Frédéric Vernier & Anna Koroleva

LIMSI-CNRS
Dépt. CHM - Groupe ILES
Rue John von Neumann, Campus Universitaire d'Orsay
Bât 508, 91405 Orsay cedex
pap@limsi.fr

30 nov. 2017 / Telecom ParisTech

Les données massives (« Big Data »)

- *10 To/jour*[1]*, 15 Po d'ici 2020, par le CNES sur PEPS=plateforme (libre & gratuit), données des satellites Sentinels,* `https://peps-mission.cnes.fr/fr`
- *200 Po, CERN Data Centre (*`https://home.cern/about/computing`*)*
- *Google ~20 Po/jour d'indexes (source : atelier TIM2017/DGA/juillet 2017)*

1. To=10^{12}=2^{40} octets, Po=10^{15}=2^{50} octets

Rediscovering 50 Years of Discoveries in Speech and Language Processing: A Survey.

Joseph Mariani[1]

Gil Francopoulo[2], Patrick Paroubek[1], Frédéric Vernier[1]

[1]LIMSI-CNRS, [2]Tagmatica

Context

- Extension to half a century of research in Speech and Language Processing (1965-2015)
  - Oriental-Cocosda 2017, Seoul
  - 20th Conference of the Oriental Chapter of the International Committee for the Coordination and Standardisation of Speech Databases and Assessment Techniques (Cocosda)

More information

- *Workshop on* ***Mining Scientific Publications*** (WOSP'2015)
  – Fort Knox, June 24-25, 2015
  – *D-Lib Magazine* **(**Nov./Dec. 2015, Vol. 21, N° 11/12)
- ***Computational Linguistics and Bibliometrics*** *(CLBib) Workshop*
  – *15th Int[al] Society of Scientometrics and Informetrics Conference* (ISSI)
  – Istanbul, June 29, 2015
- *BIRNDL: Joint Workshop on* ***Bibliometric-enhanced IR*** *(BIR) and* ***NLP and IR for Digital Libraries*** *(NLPIR4DL)*
  – ACM/IEEE Joint Conference on Digital Libraries'2016
  – Newark, June 23, 2016
  – *International Journal on Digital Libraries* Special issue (March 2017)
- *ACFAS:* ***Digital Libraries as Research Data***
  – Montreal, May 8-9, 2017
  – *Document Numérique* Special issue (December 2017)

Data

NLP4NLP Corpus

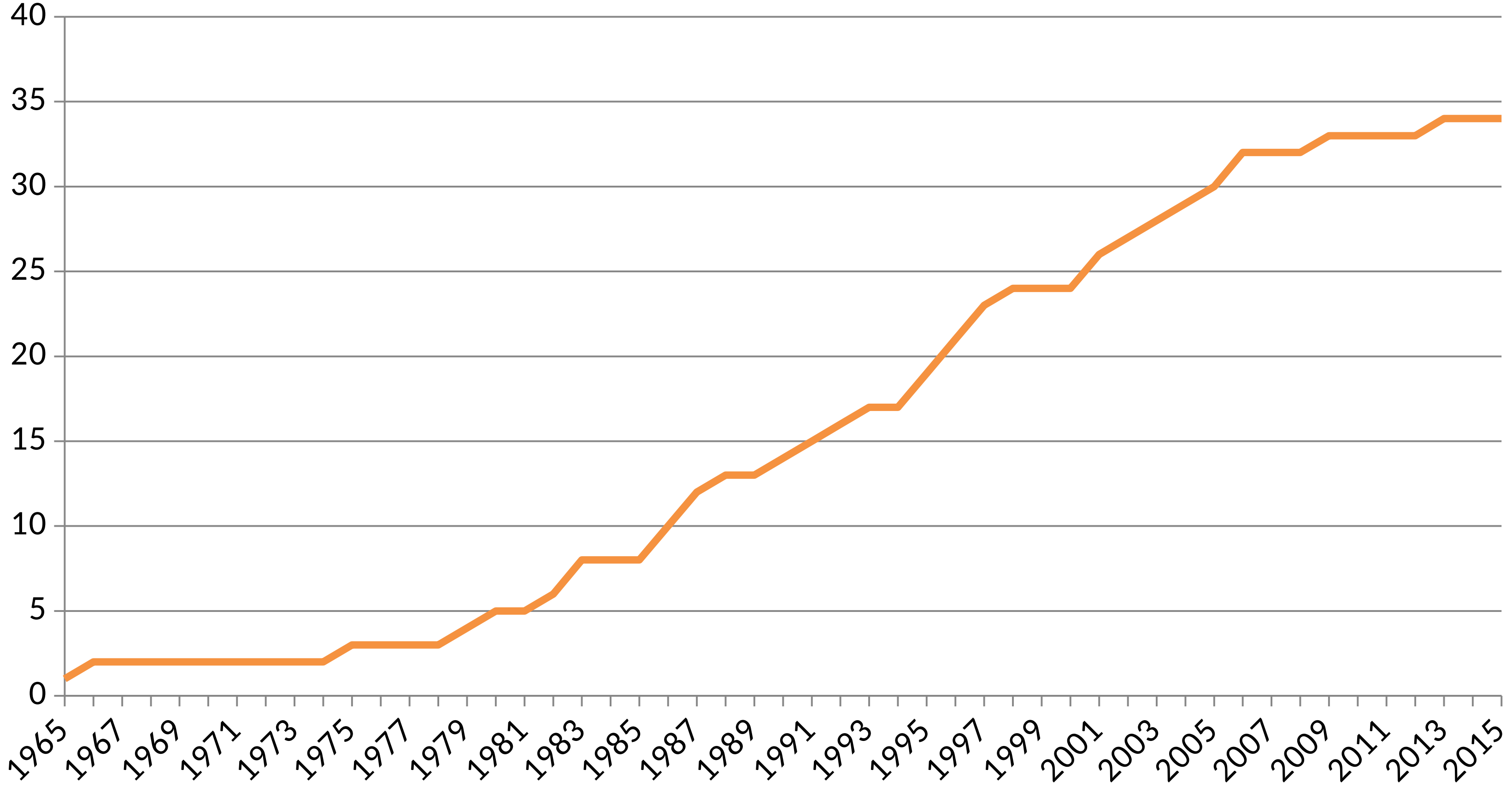
- Study of NLP domain (incl. written, spoken and signed language processing, and Information Retrieval) with NLP tools
- 34 publications over 50 years (1965-2015)
- Conferences (ACL, IEEE-ICASSP, ISCA-Interspeech, ELRA-LREC, etc.) and Journals (IEEE-TASLP, CL, SpeechCom, CSAL, LRE, etc.)
- 558 events
  - Conference venues
  - Journal Issues
- 65,003 documents
- 48,894 different authors
- 270 Mwords
- 324,422 bibliographical references
- Global Analysis and Comparative Analysis
  - Across 34 sources
  - Speech versus NLP

Data Processing

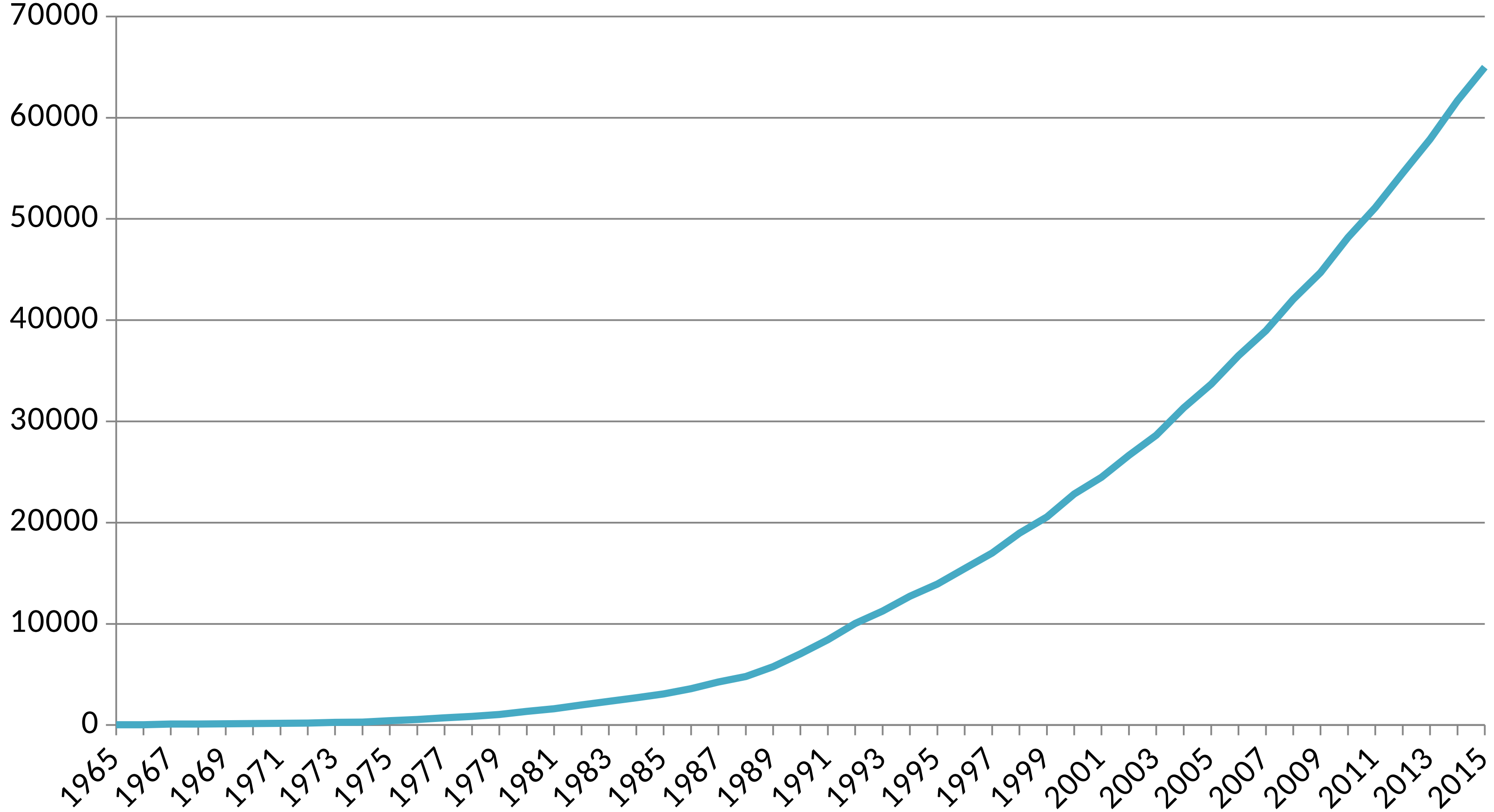
- Text as scanned images or textual format
  - OCR Software
- Existence of Metadata
- Automatic Extraction
  - Authors’ names
    - Affiliation, nationality, gender
  - Scientific Terms
  - Language Resources
  - Citation references
    - Authors, titles, sources
  - Funding agencies, etc.

Production analysis

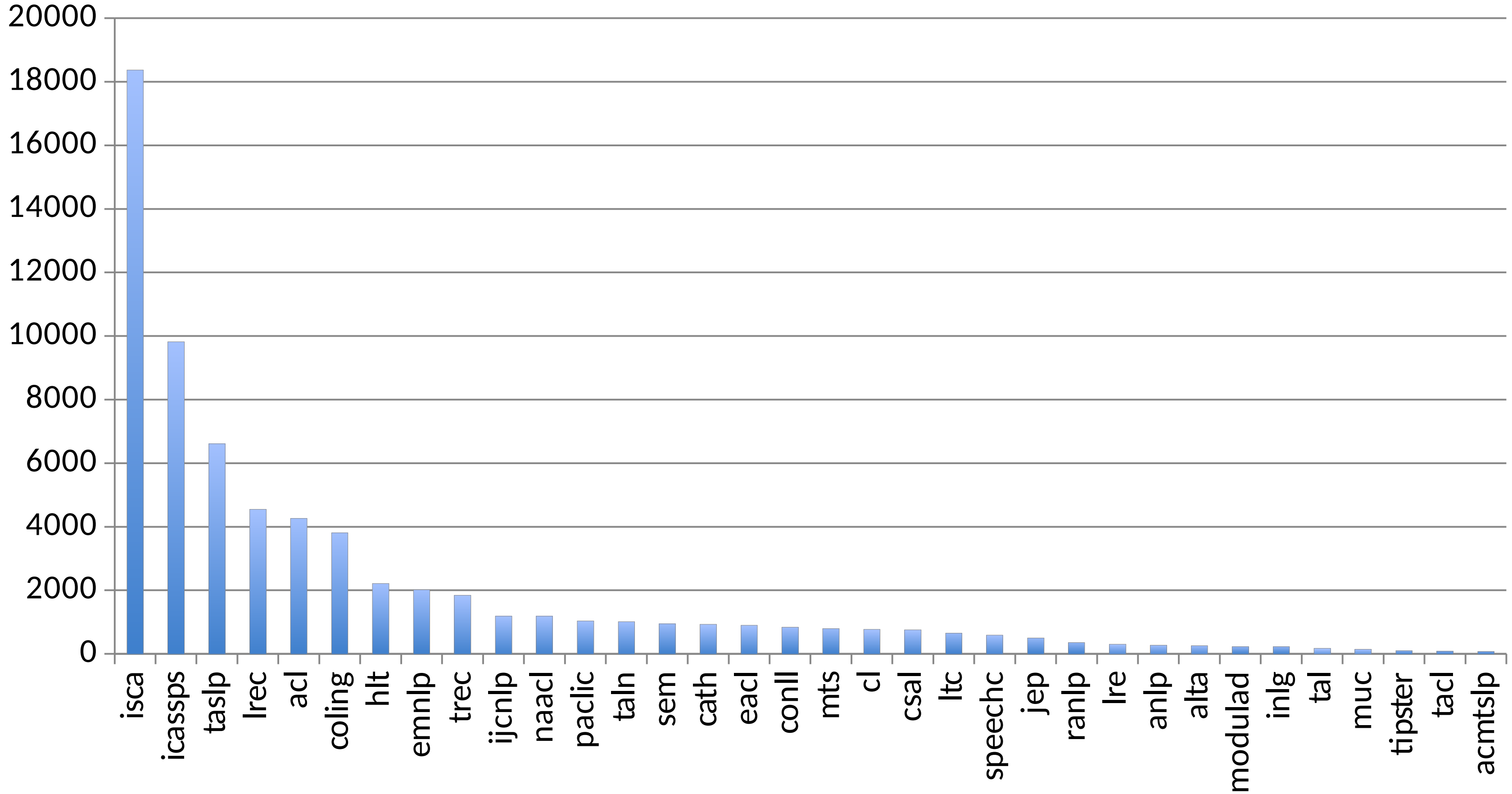
Cumulated number of sources



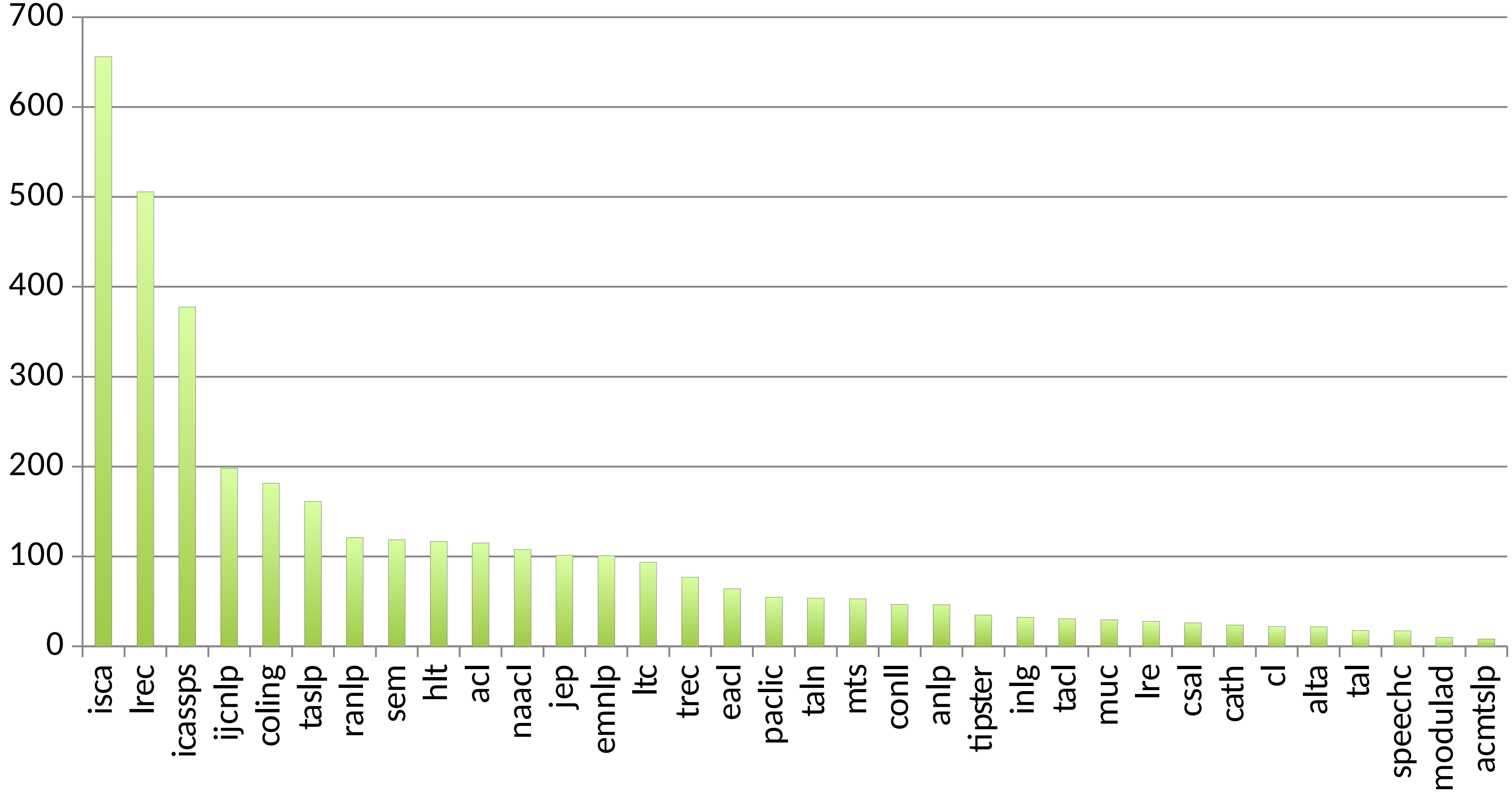
Cumulated number of papers



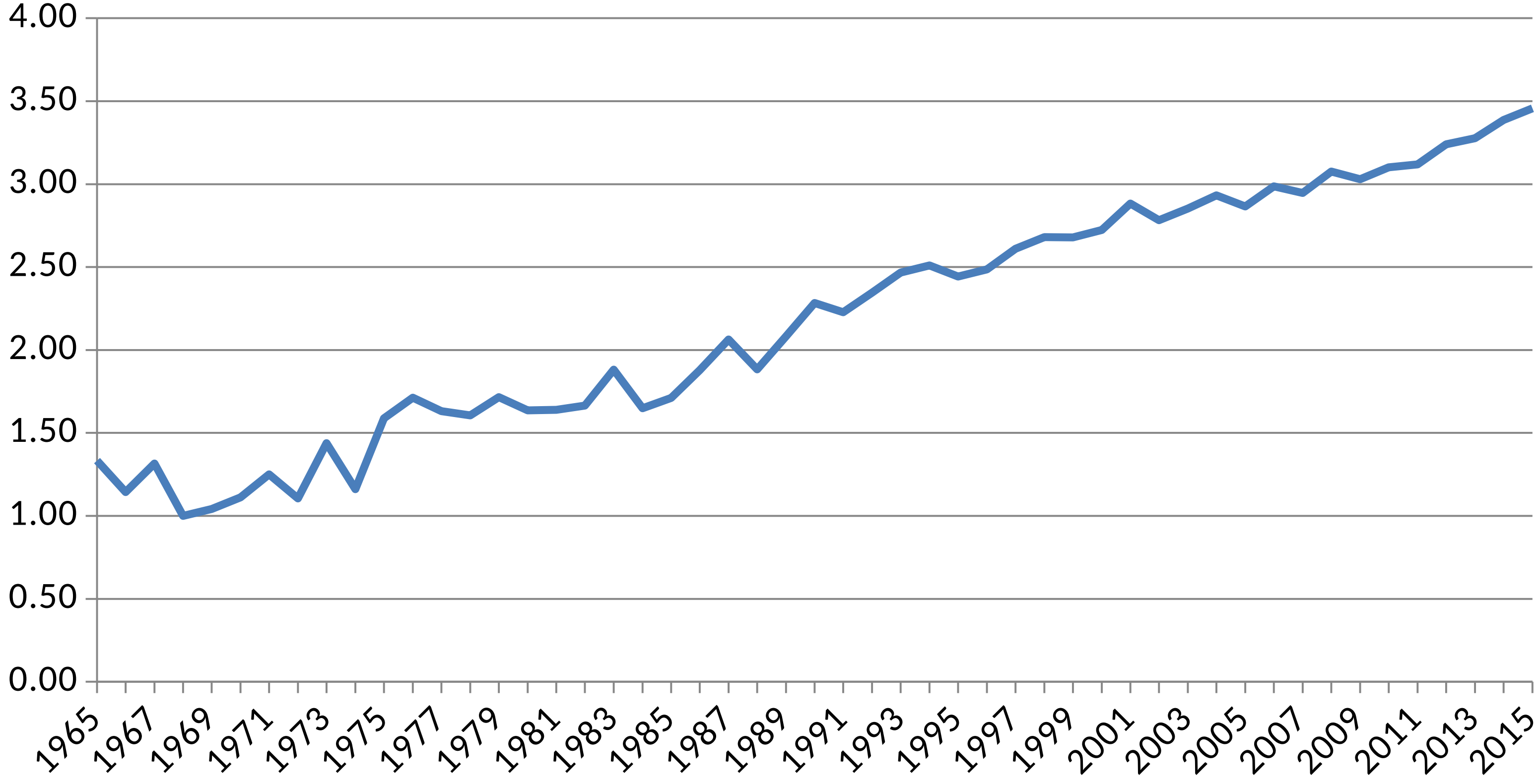
Number of papers in each source



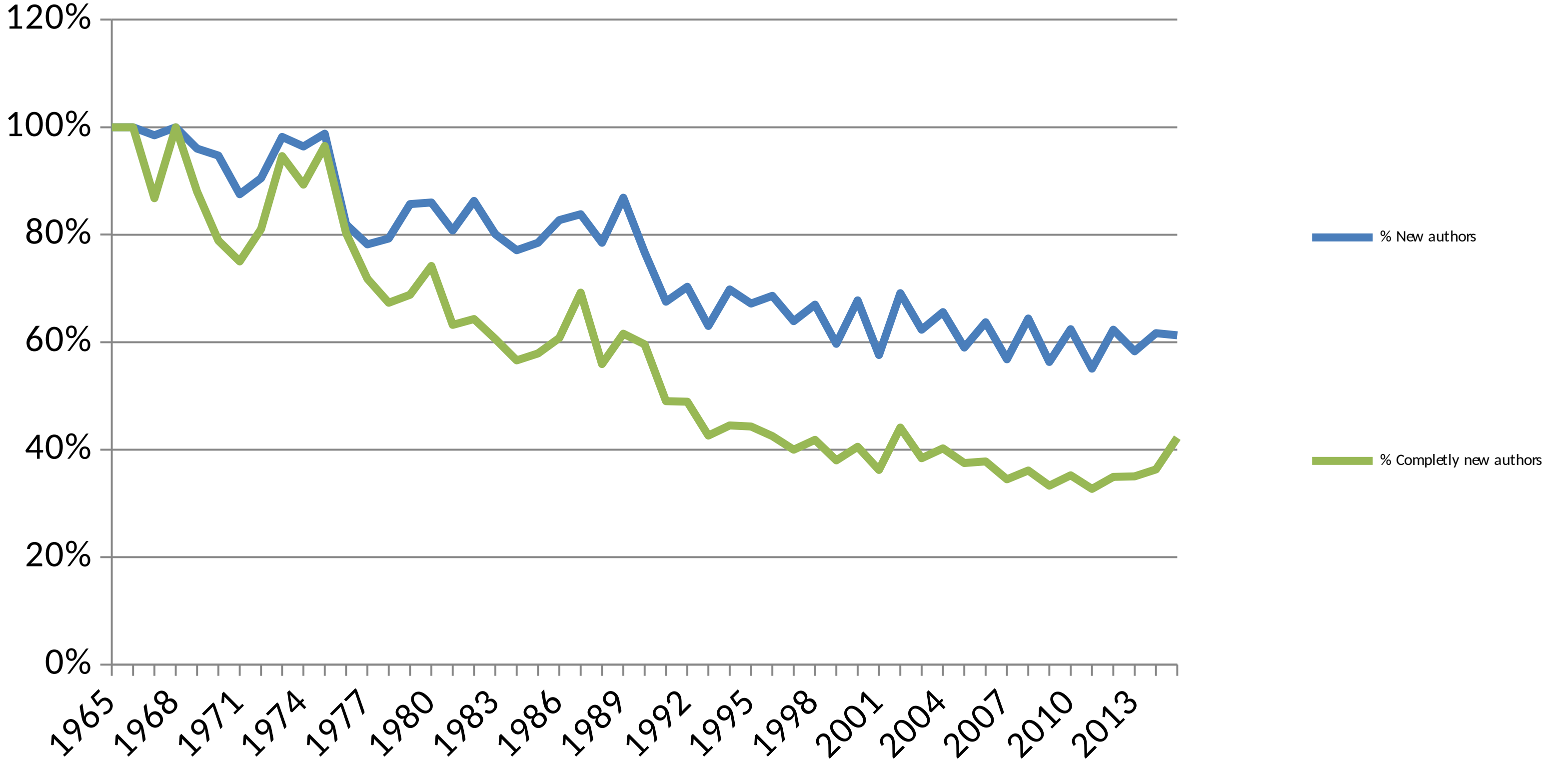
Number of papers at each event



Average number of authors per paper

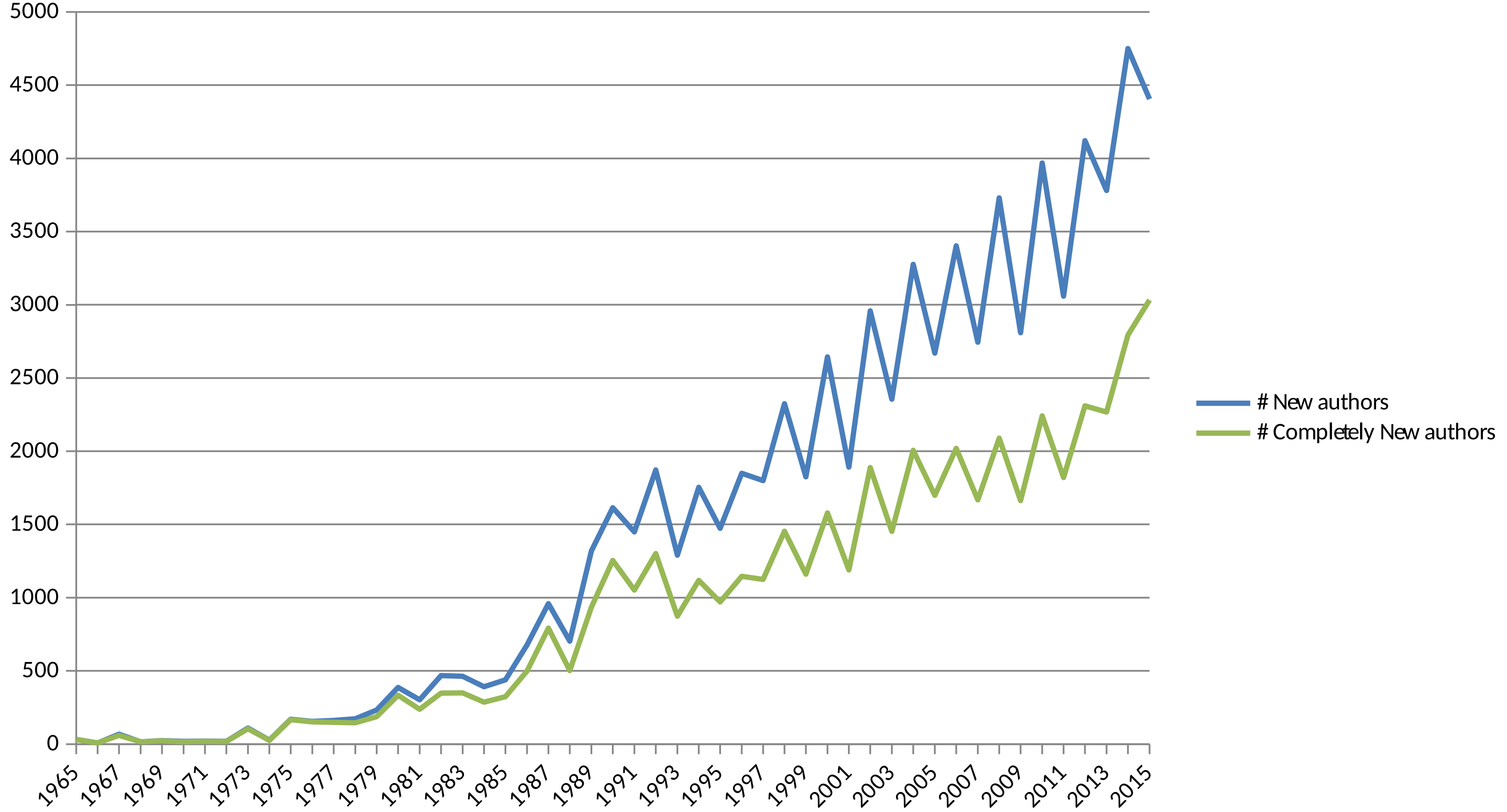


% New authors



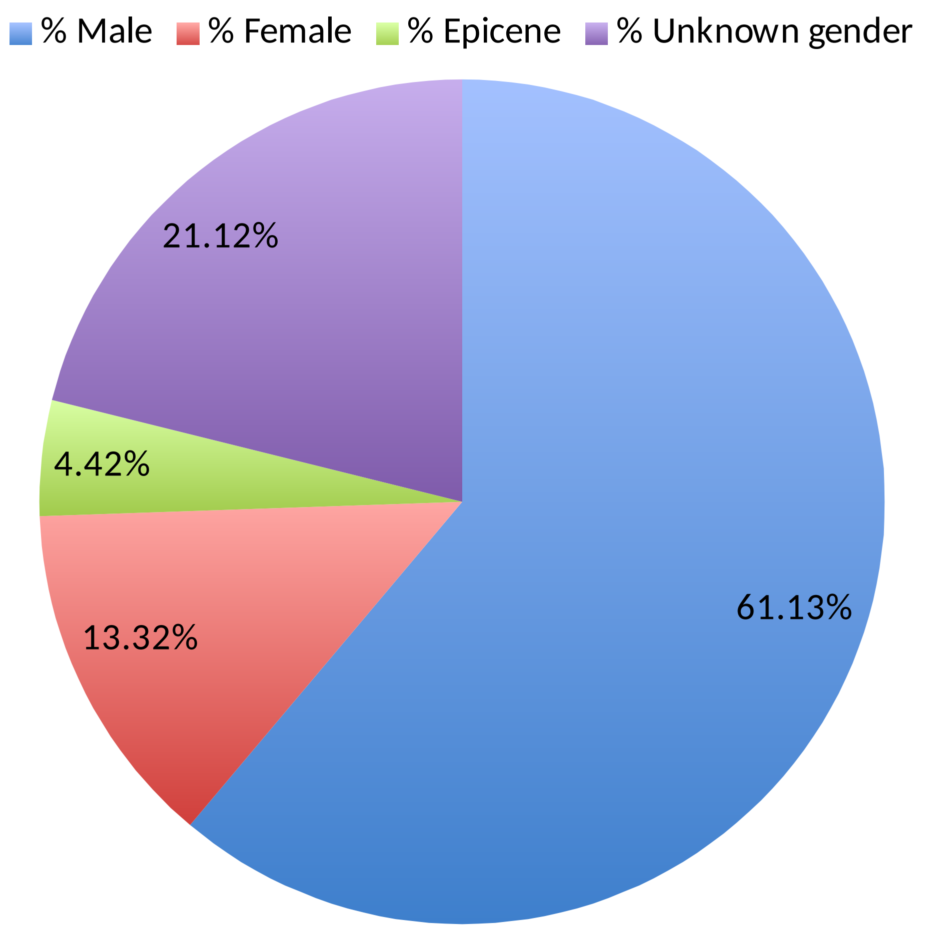
120%
100%
80%
60%
40%
20%
0%
1965
1968
1971
1974
1977
1980
1983
1986
1989
1992
1995
1998
2001
2004
2007
2010
2013
% New authors
% Completly new authors

New authors

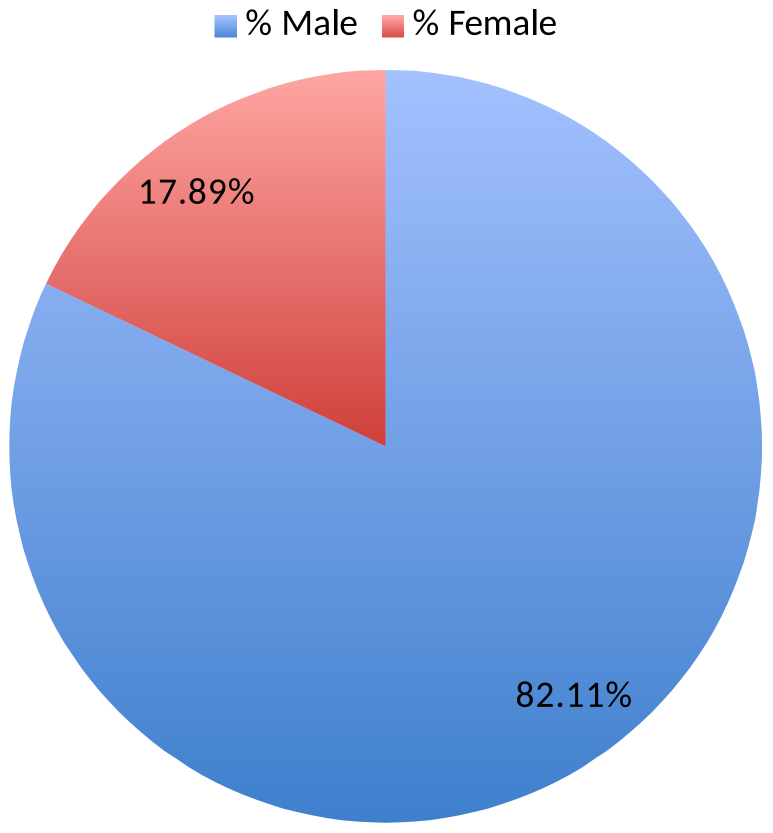


5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
1965
1967
1969
1971
1973
1975
1977
1979
1981
1983
1985
1987
1989
1991
1993
1995
1997
1999
2001
2003
2005
2007
2009
2011
2013
2015
New authors
Completely New authors

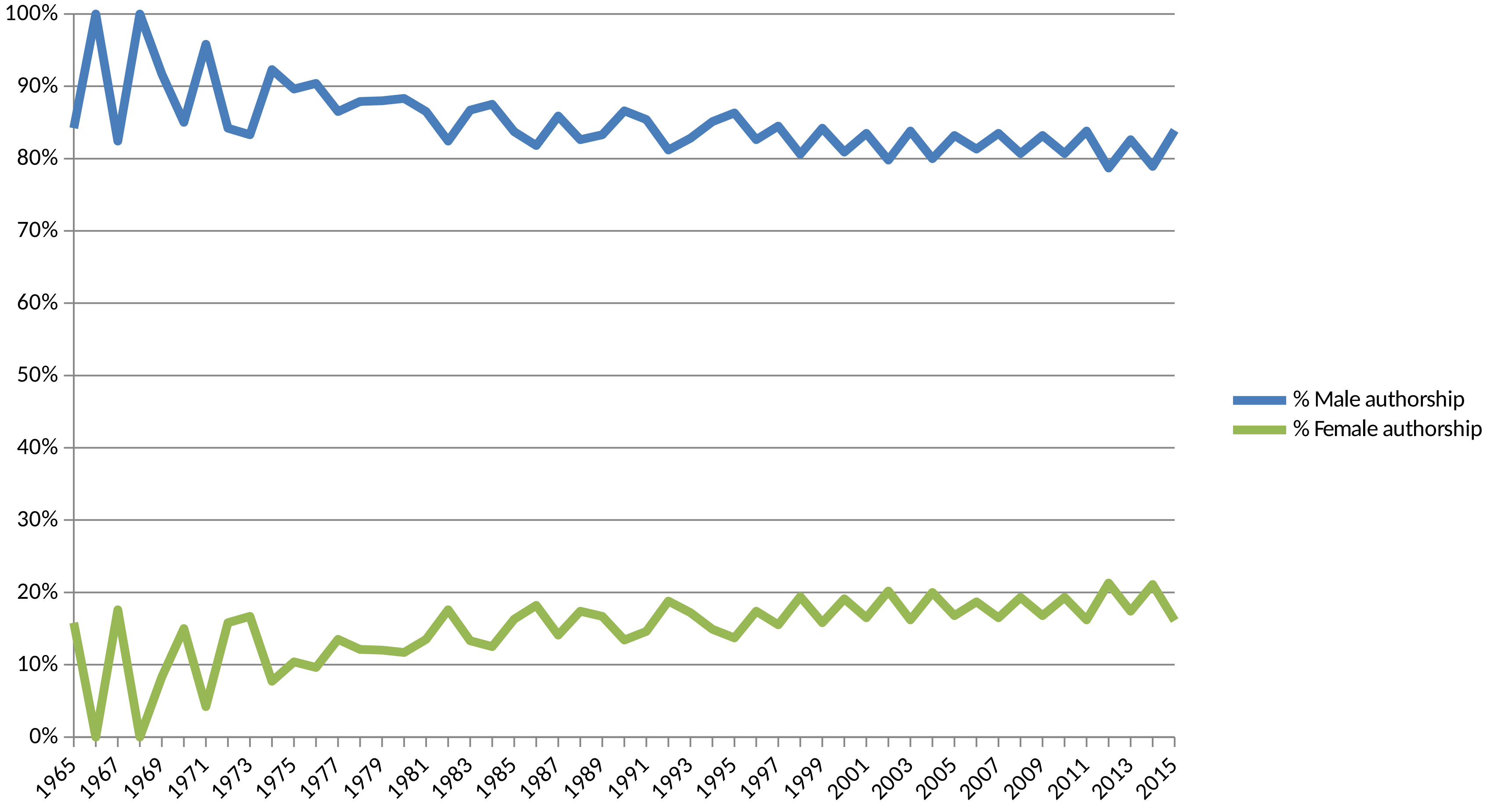
Authorship



Authorship (extrapolated)

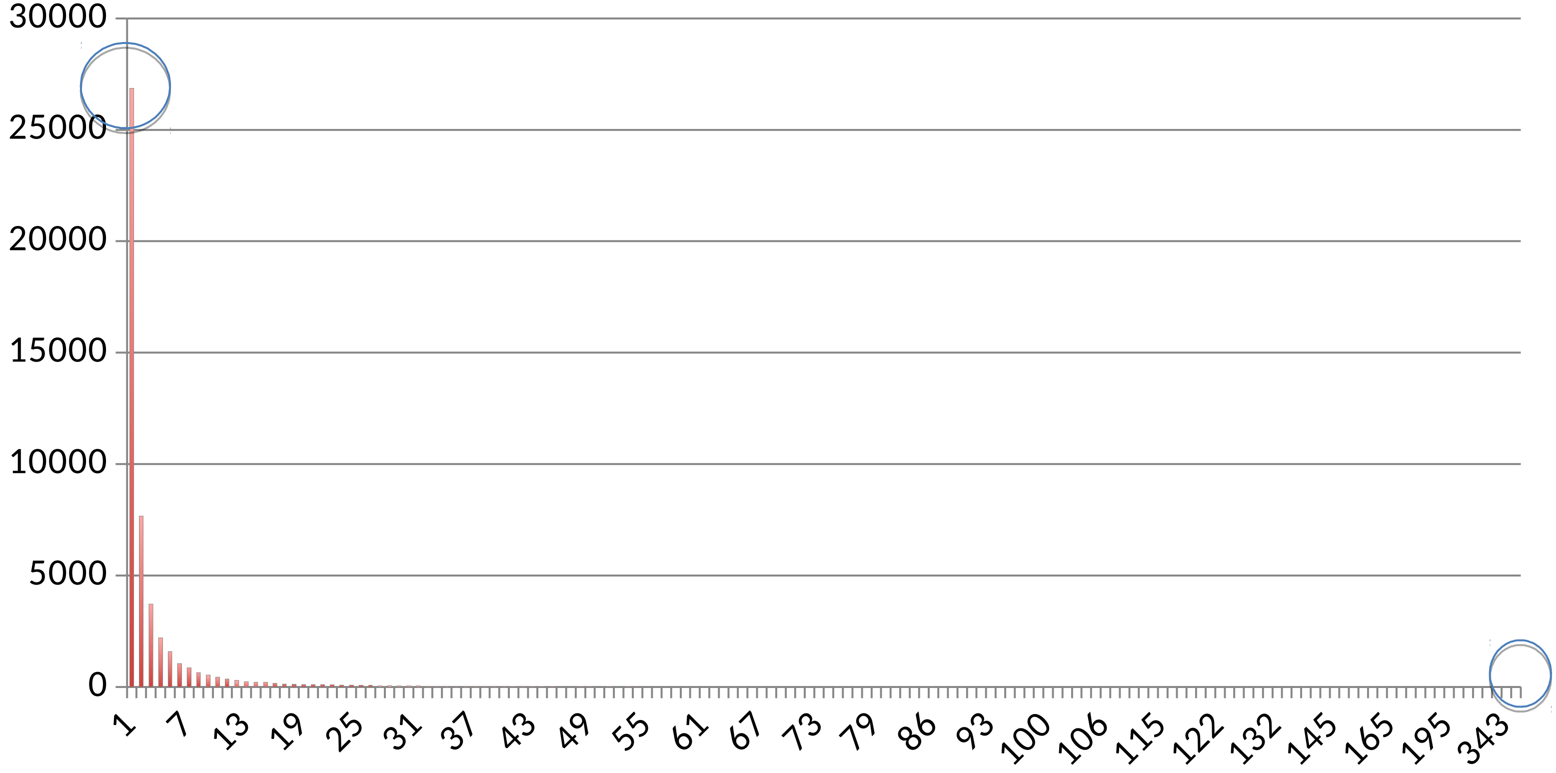


Male versus Female authors



Collaboration between authors

Number of papers per author



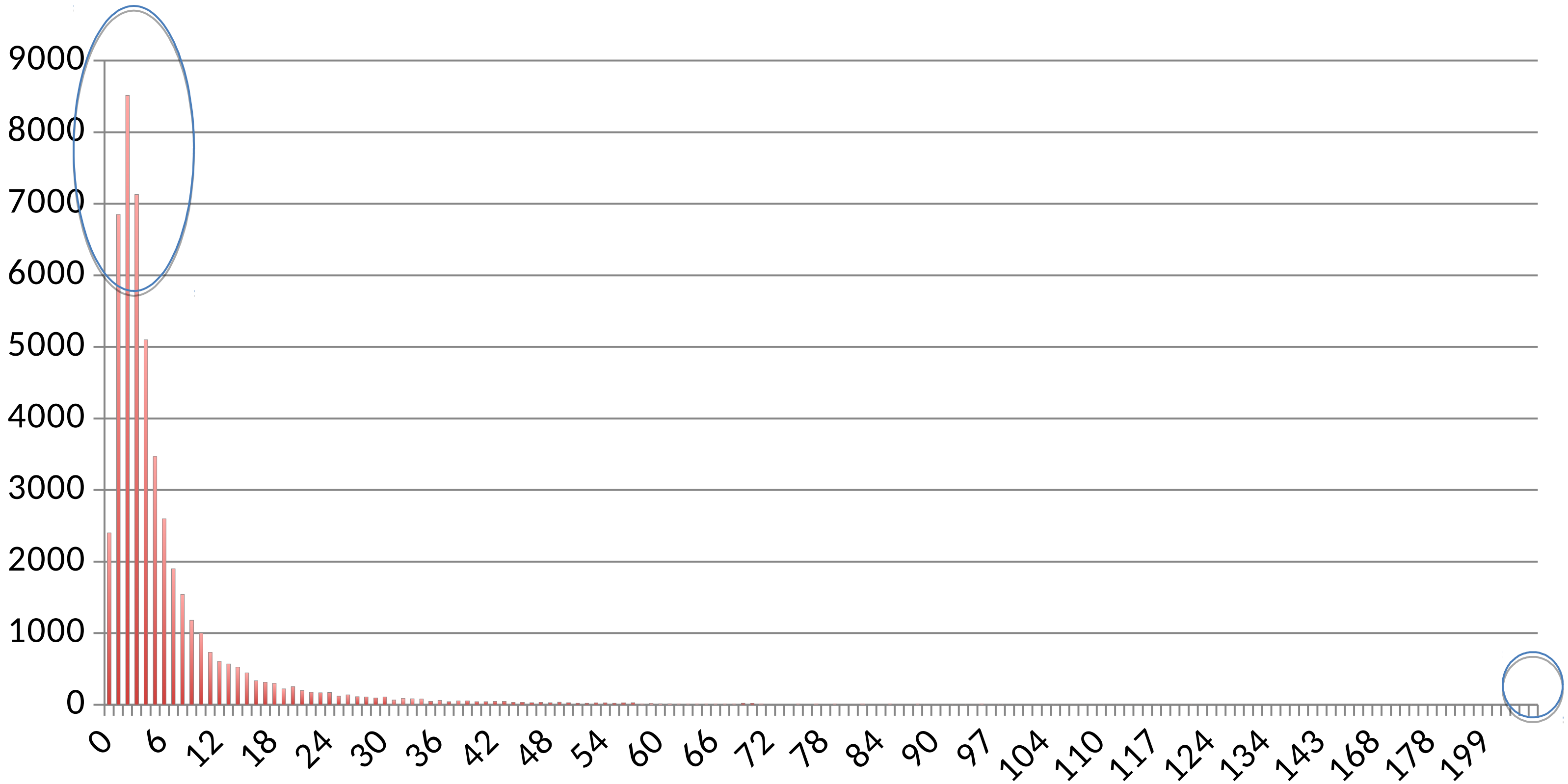
Number of papers per author

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

Number of papers as single author

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

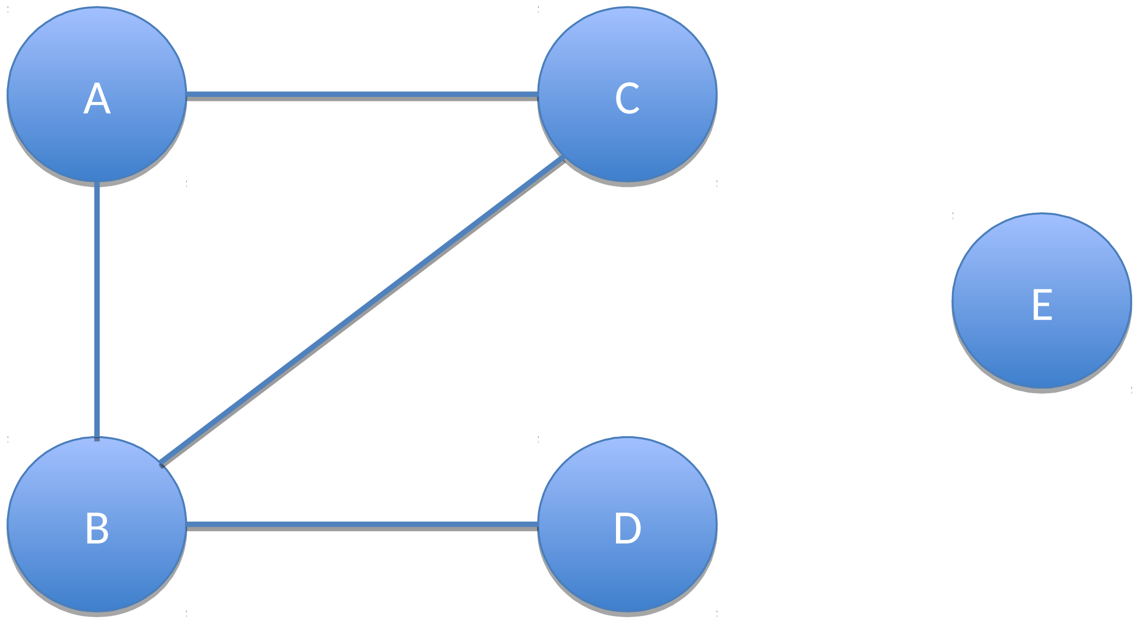
Number of co-authors



Number of co-authors

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

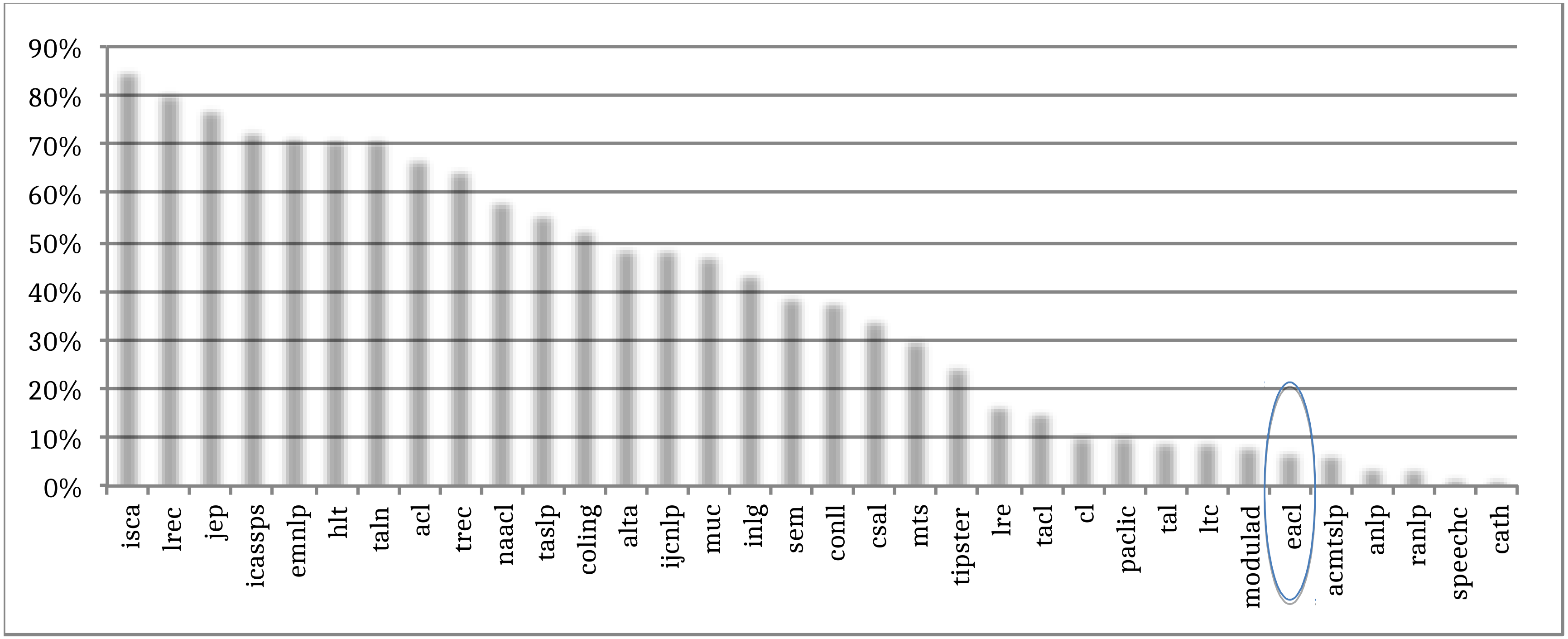
Collaboration Graph



Collaboration Graph: Connected Components

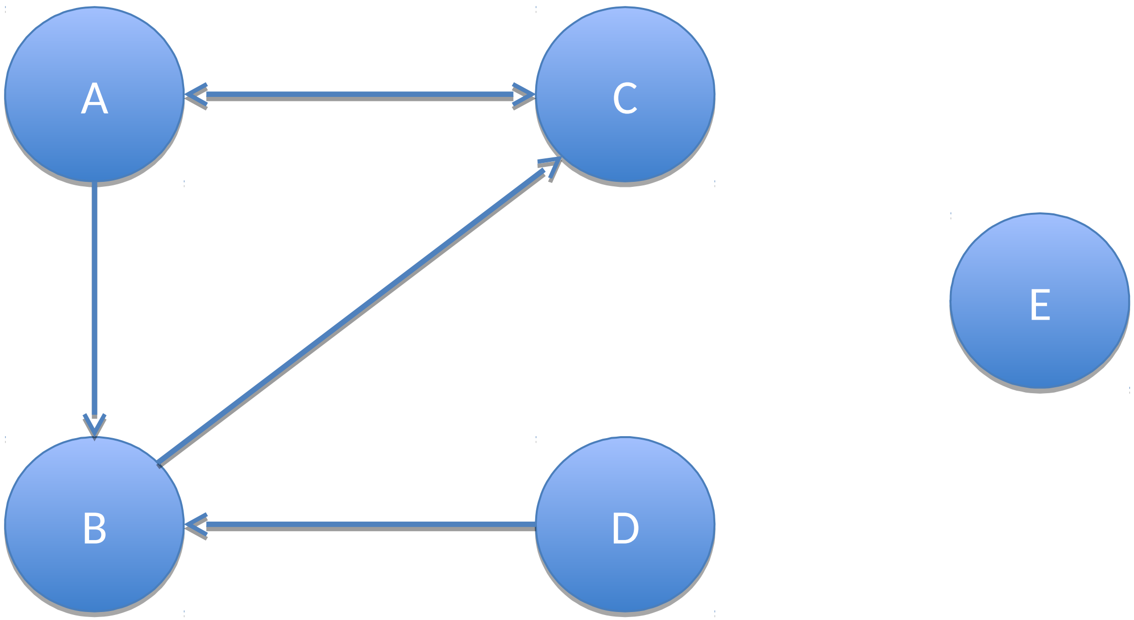
Connected Component Size	# of Connected Components	# of authors	% of Authors in the Connected Components	% of Connected Components
39744	1	39744	81%	0%
29	1	29	0%	0%
27	1	27	0%	0%
21	1	21	0%	0%
18	3	54	0%	0%
17	1	17	0%	0%
15	1	15	0%	0%
14	1	14	0%	0%
12	2	24	0%	0%
11	9	99	0%	0%
10	5	50	0%	0%
9	14	126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

Collaboration Graphs: % of authors in largest Connected Component / sources

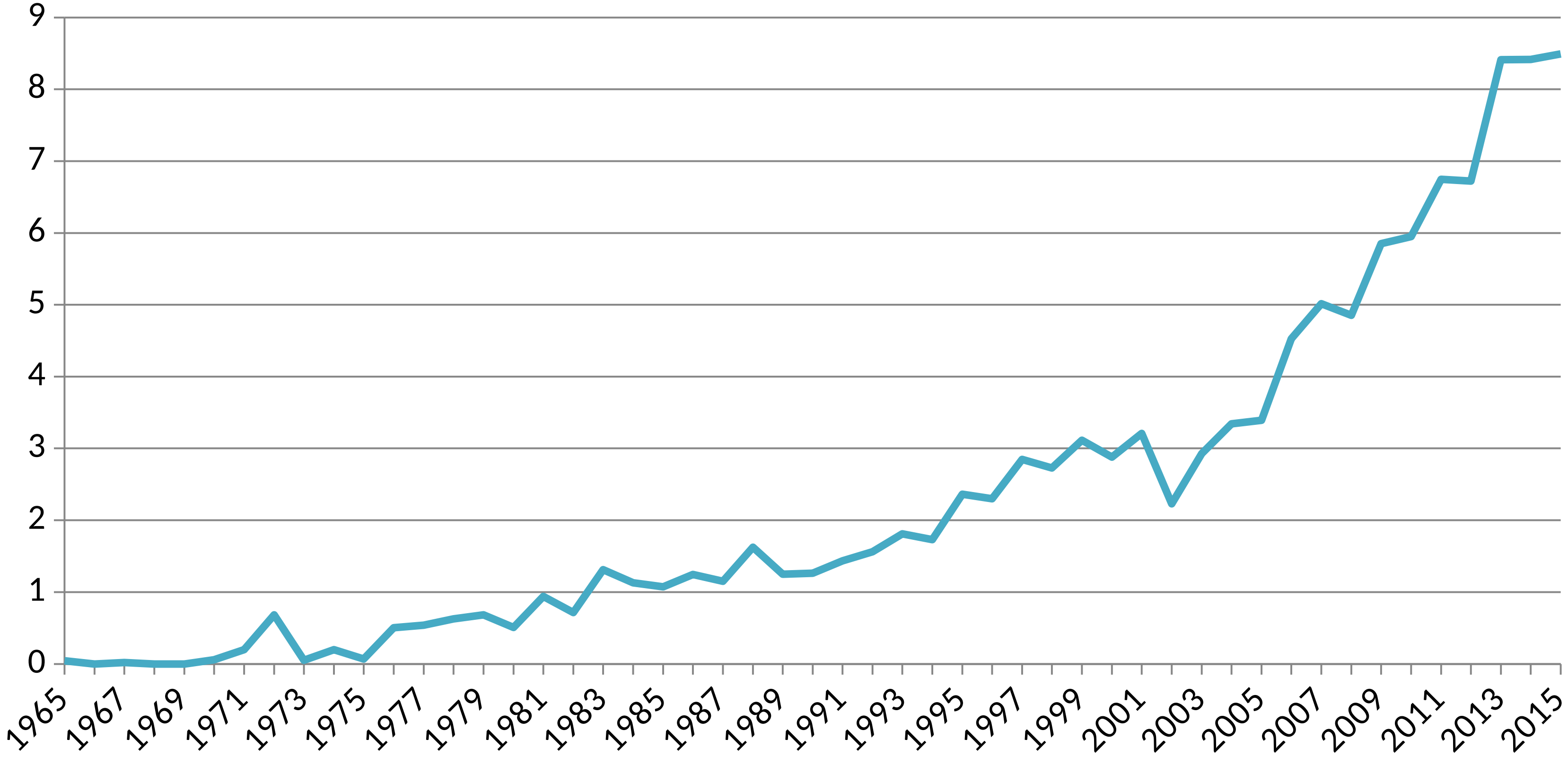


Citations of authors and papers

Citation Graph (authors)



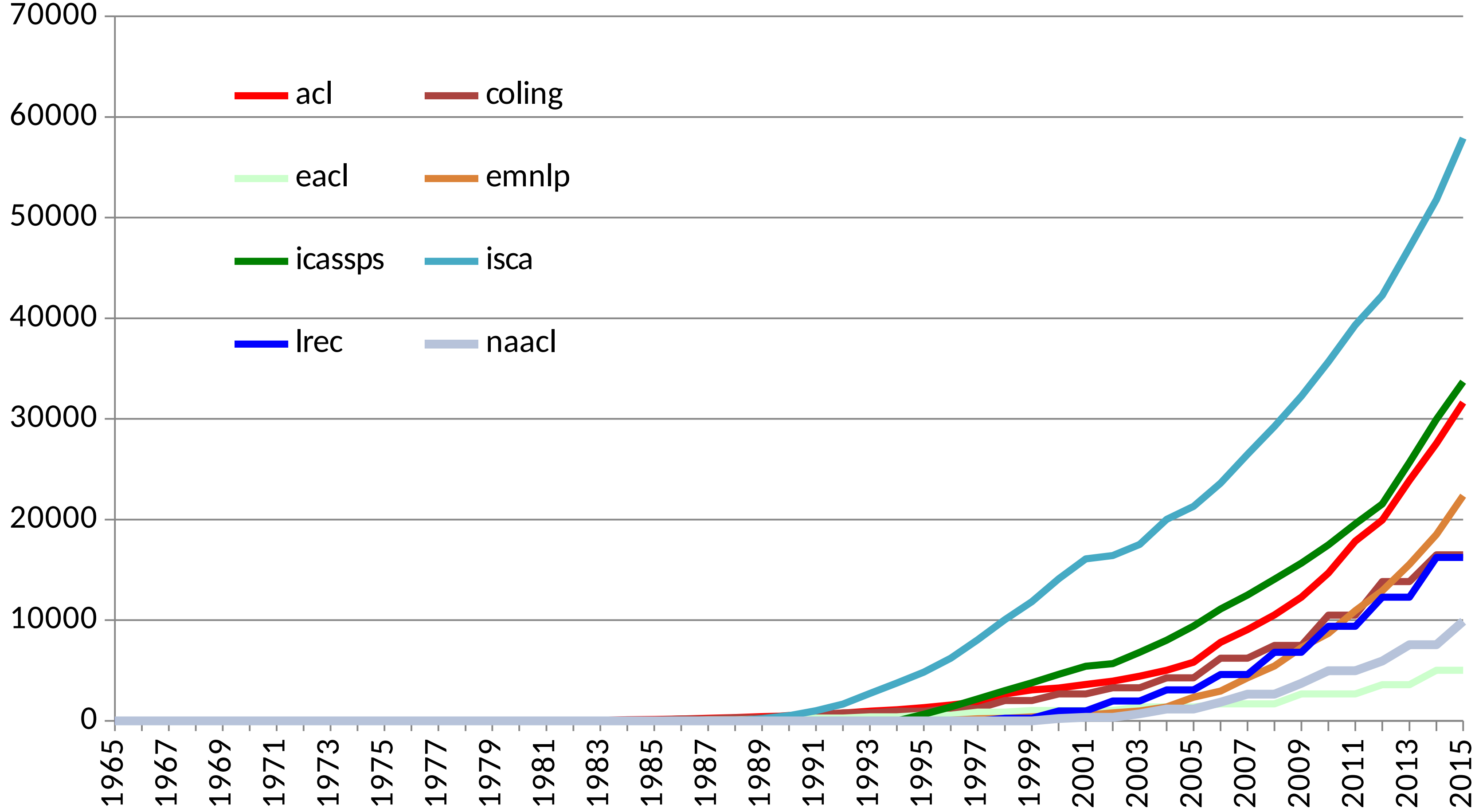
Average Number of References per Paper over the years



Annual versus Cumulative

- Annual: Number of references contained in papers ***on*** a given year
- Cumulative: Number of references contained in papers ***up to*** a given year

Cumulative Number of references in papers over the years (8 main conferences)



10 most Cited Authors

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

Citations

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

Evolution of research Topics

Definitions

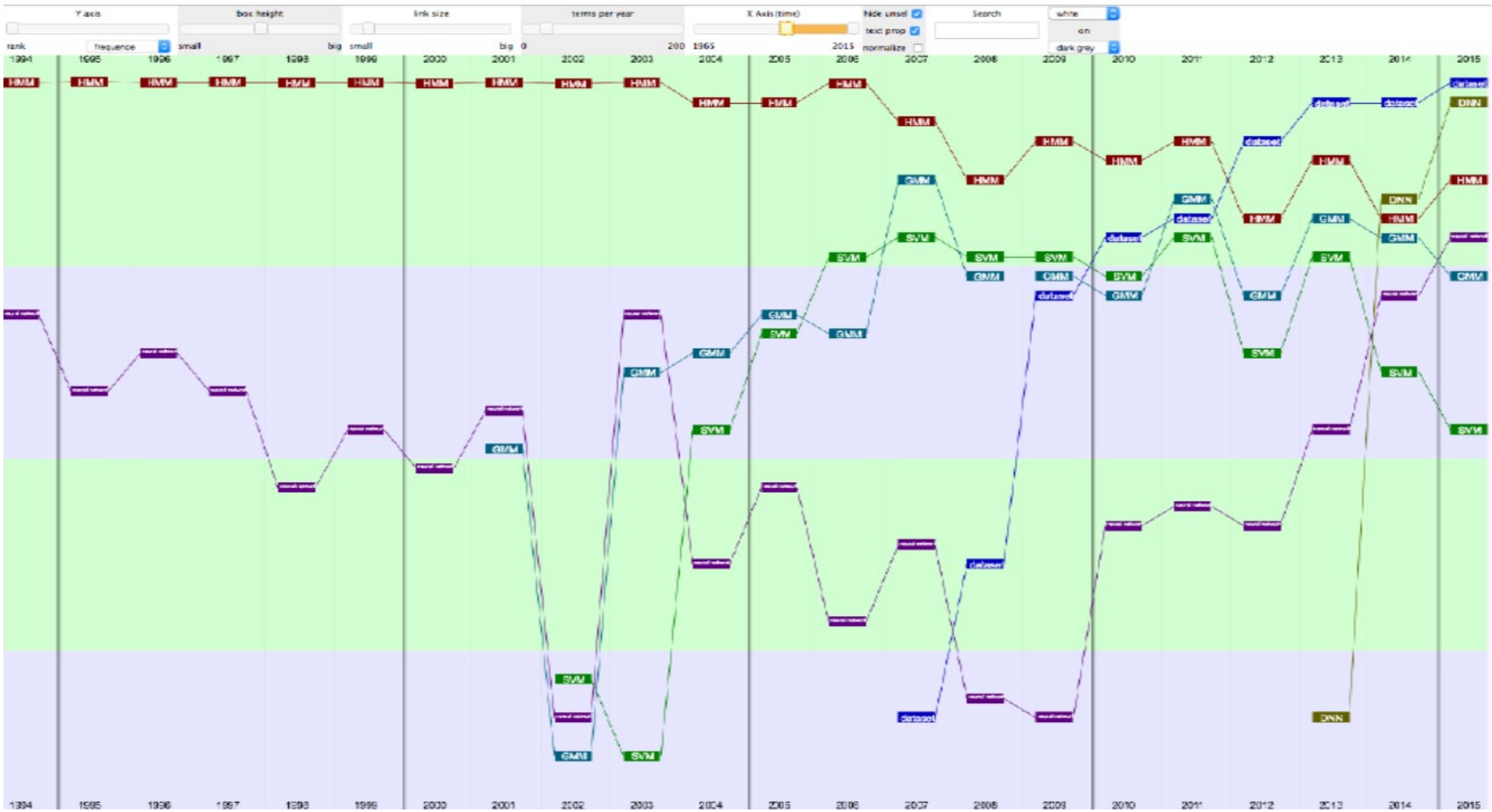
- Occurrence : mention of a word
- Frequency : # occurrences / # words
- Existence : mention of a word in a paper (0/1)
- Presence : # existences / # papers

- Technical Term corresponds to Research Topic
- Term: unigram, bigram, trigram
- Several variants for the Term

Most Frequent Topics

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

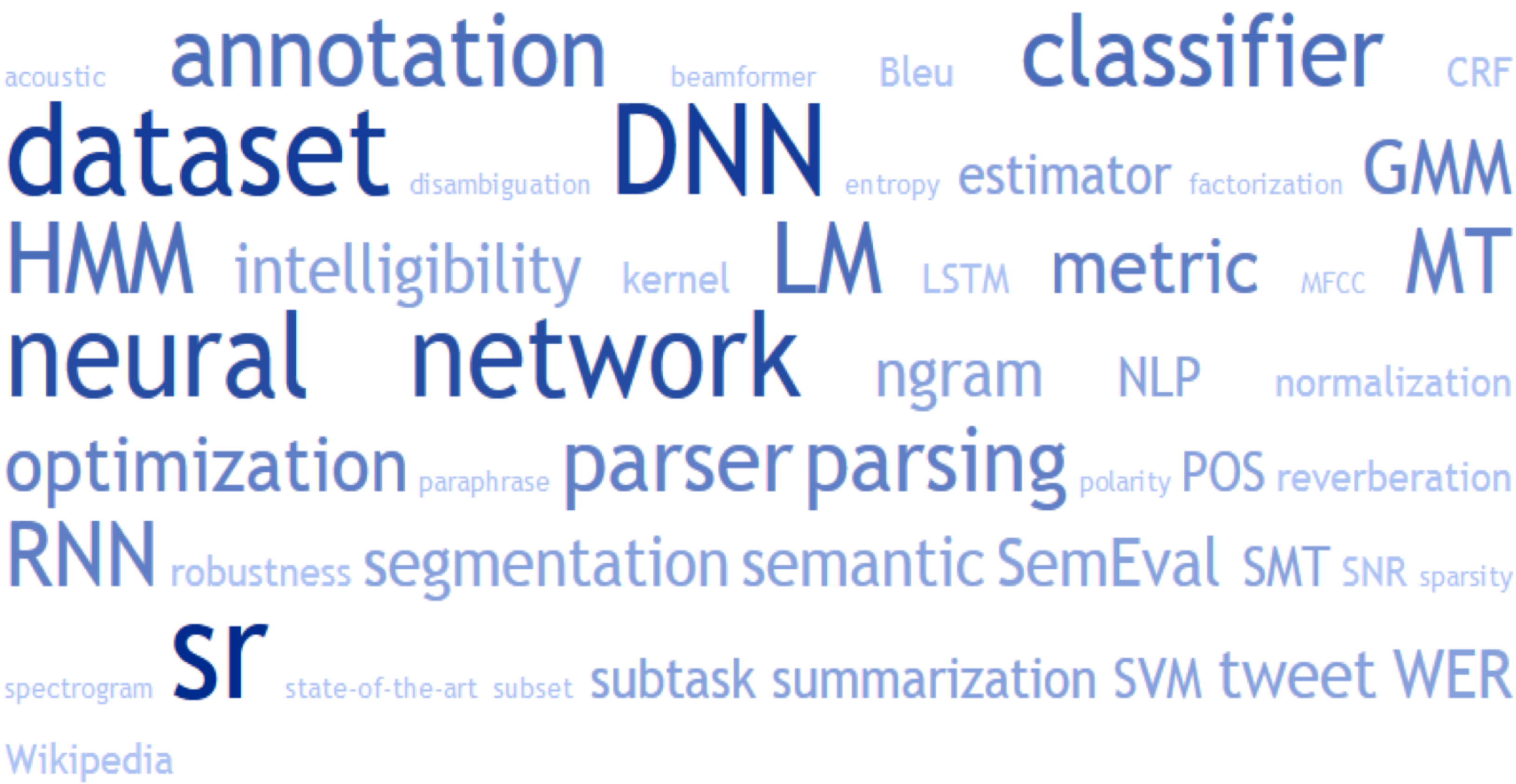
Topics Evolution over Time (Ranking 1994-2015)



TagCloud 1965

ACS automaton axiom calculus concatenation rule connective contraversial Dale derivation description of problem description of tree embedded equivalence of premise equlvalent example in linguistics Fortran Fortran IV system functioning of word GF IBSYS Ica inecke investisators kernel Kuno lexical lexicostatistics linguistical linguistical application LRS MANELSKL Modi MOP morpheme predicate verb Rom semantic sentence by phrase set of grammar set of operator set of string string against part string form structure grammar subset test vocabulary transformation on tree transformational VANDENBURGH variety of classification

TagCloud 2015



Tracking of Innovation

Introduction of the 10 most present terms in 2015

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

Manual checking

- *"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)."*
  Laurence Urdang, Computer and the Humanities, 1966

- *"Barring Arthur Clarke's reliance (in "2001, Space Odyssey") 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?"*
  Bonnie Lynn Webber, Conference of the ACL, 1980

Manual checking

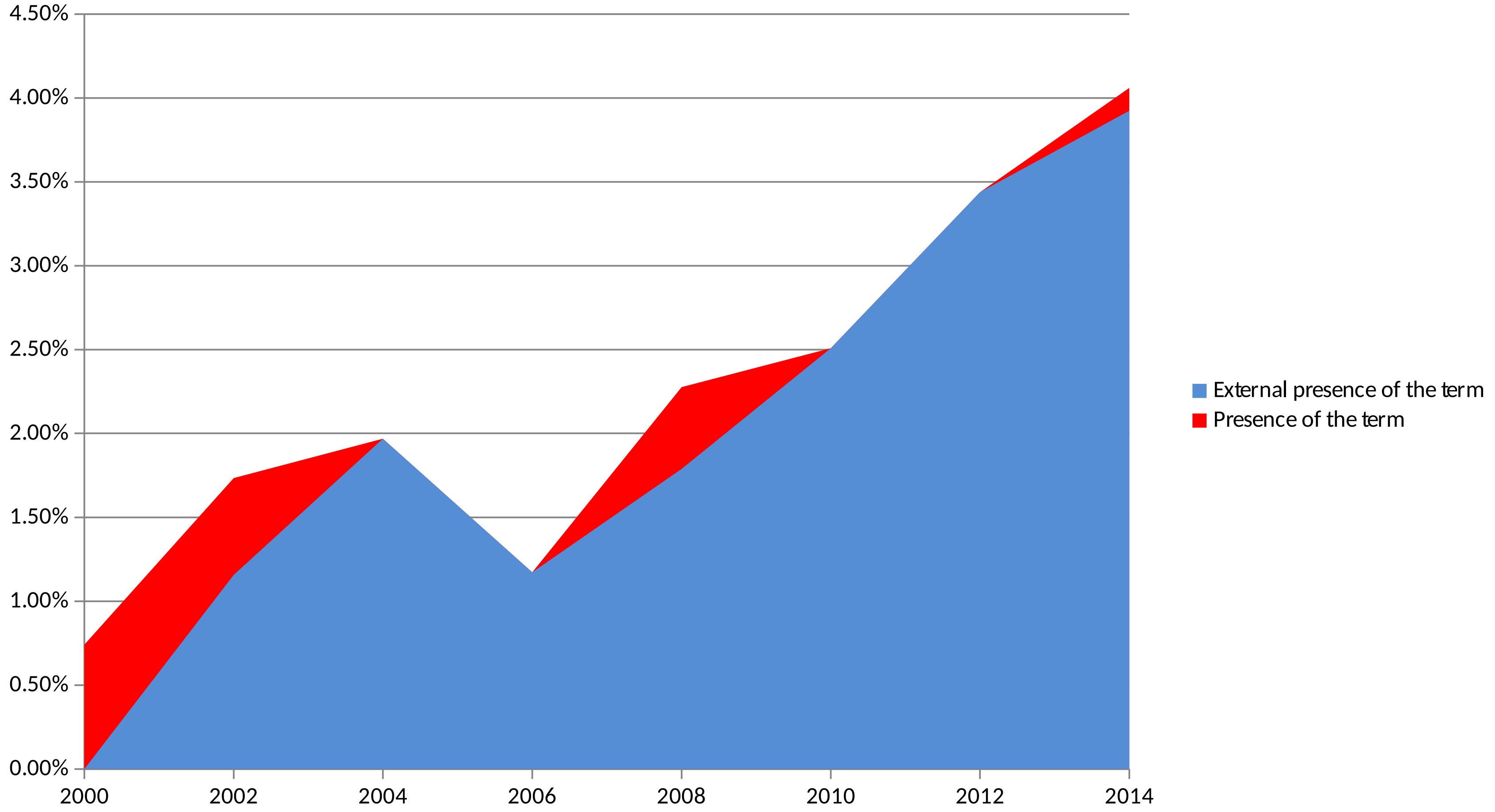
- First mention of HMM: Z.M. Shalyapina, *Problems of formal representation of text structure from the point of view of automatic translation*, Coling 1980

standpoint, the false implication is accounted for by the possibility, suggested by grouping the sentence units into the above two fragments, of interpreting and/or transforming these independently of each other, thus obtaining

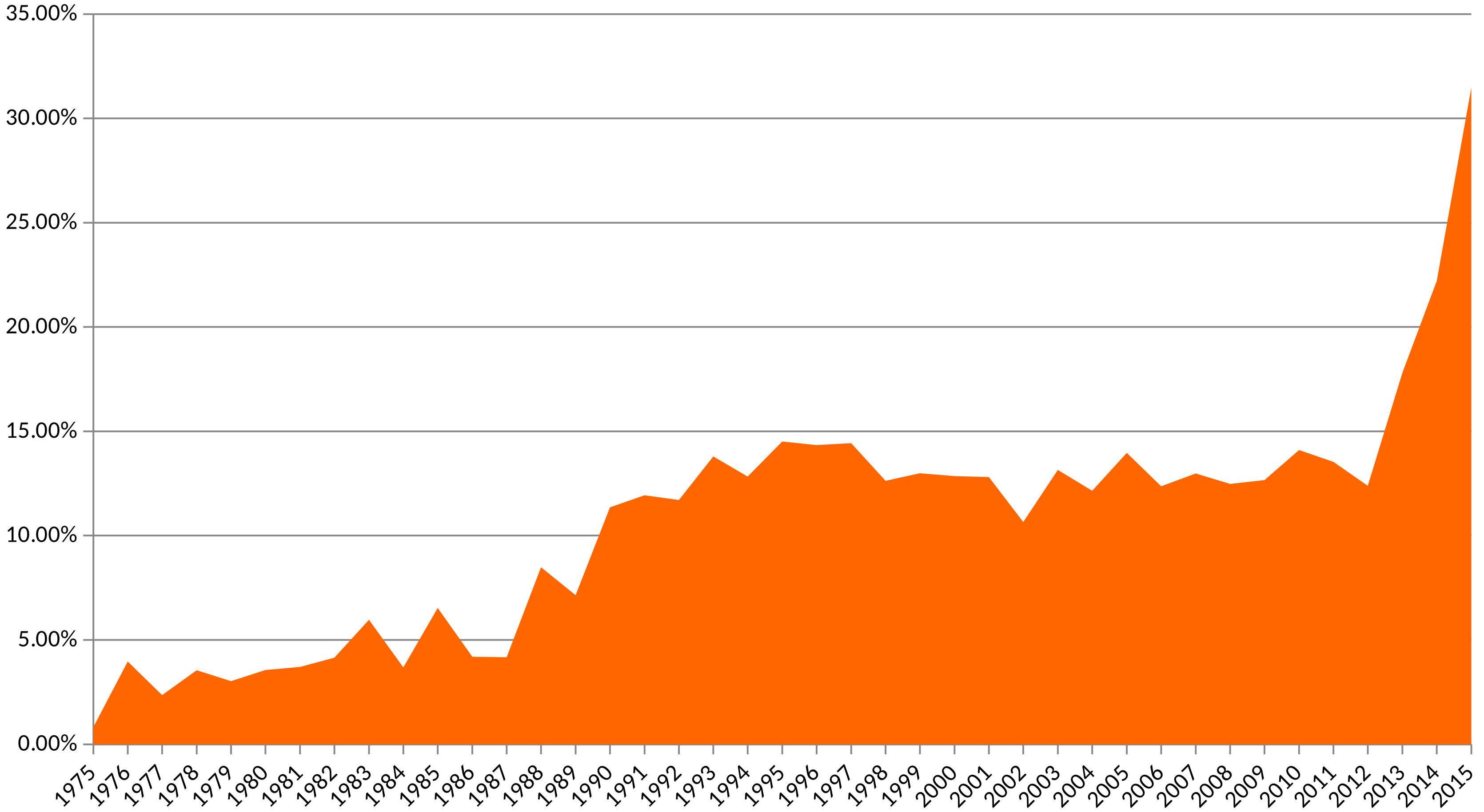
Любые из используемых нами вещей домашнего обихода изнашиваются, даже если ими пользоваться долгое время ("All of the things we use daily wear out, even if used for a long time").

No matter which one of the two explanations be taken as true (the second one seeming more plausible, while the first one suggesting simpler check-ups in processing texts) it is clear that the translation problem is to achieve in Russian the same syntactic grouping as in the

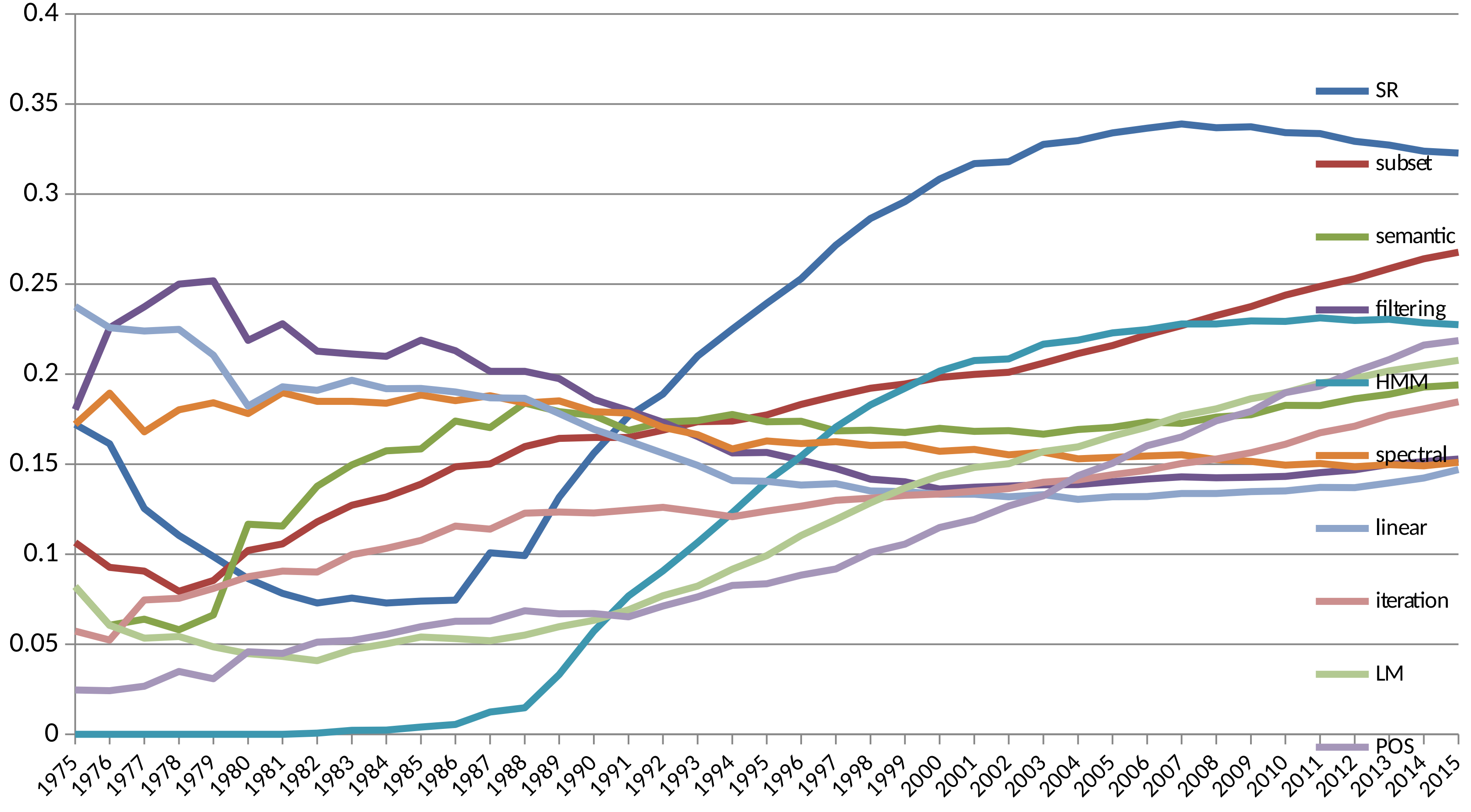
Innovation: Presence of the term over the years (e.g. “cross validation”)



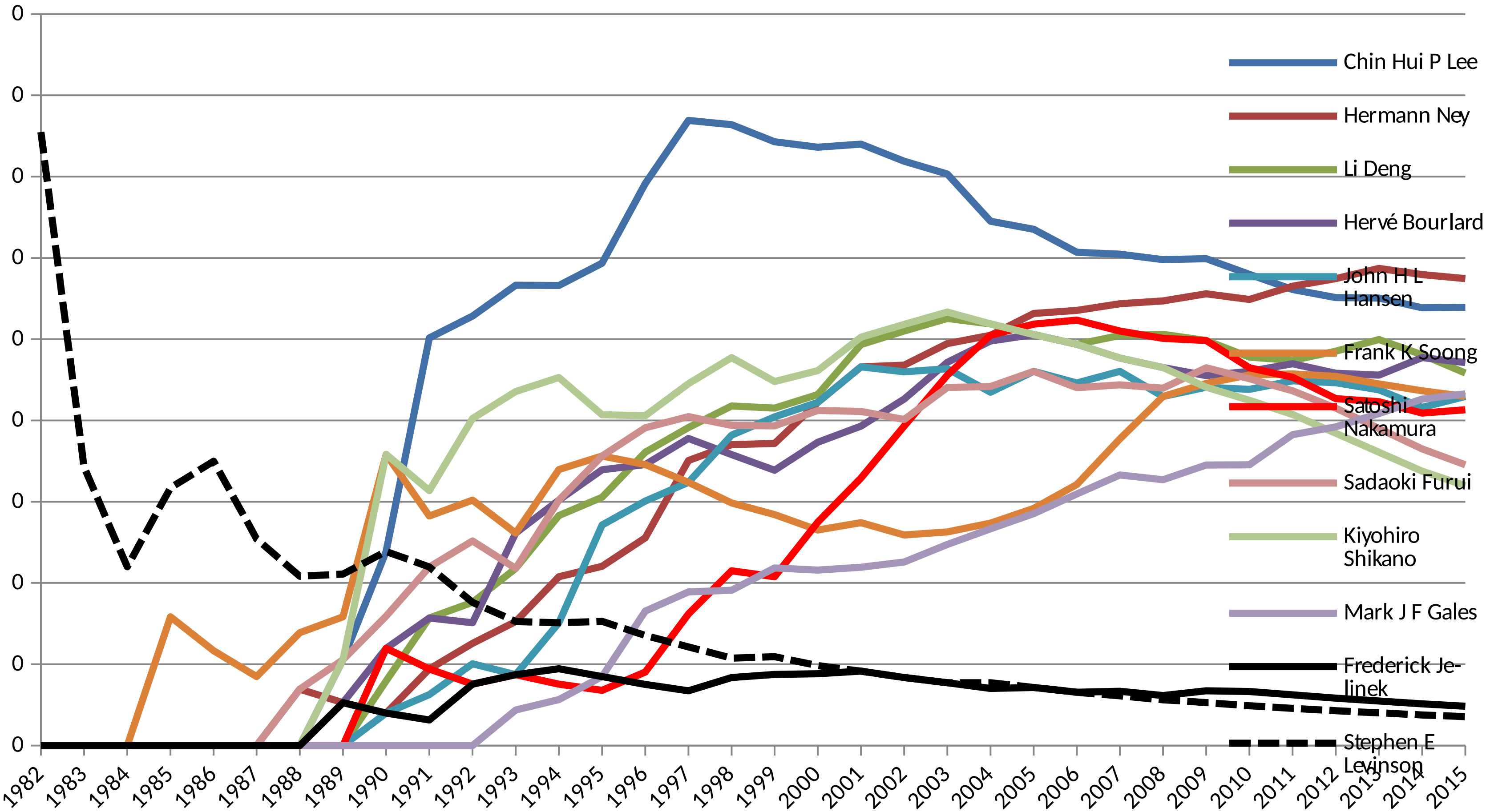
Innovation: Presence of the term over the years (e.g. "Neural Networks")



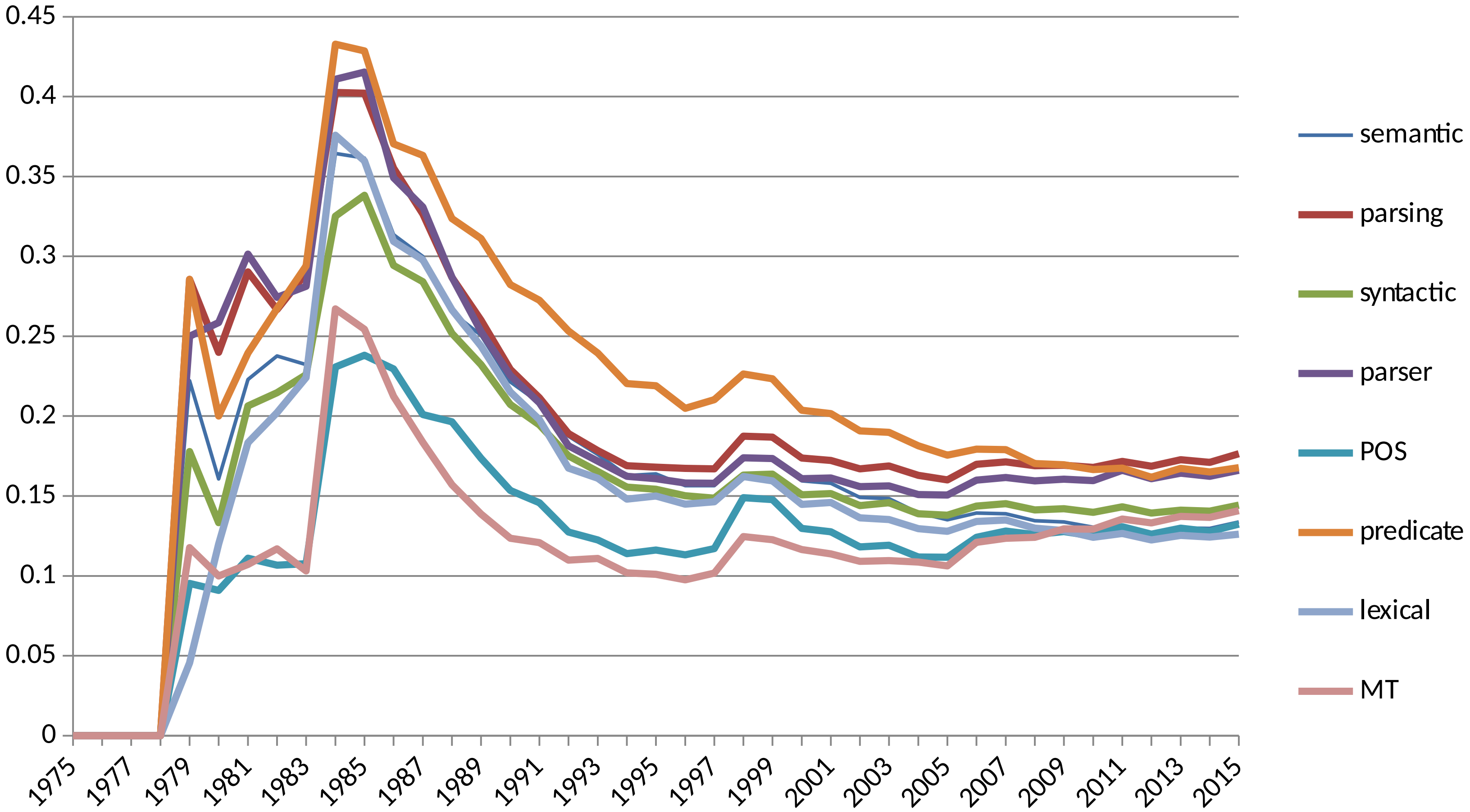
Cumulative presence of 10 most important terms



Authors' contributions to HMM (% papers)



Main domains within ACL
(% of papers)

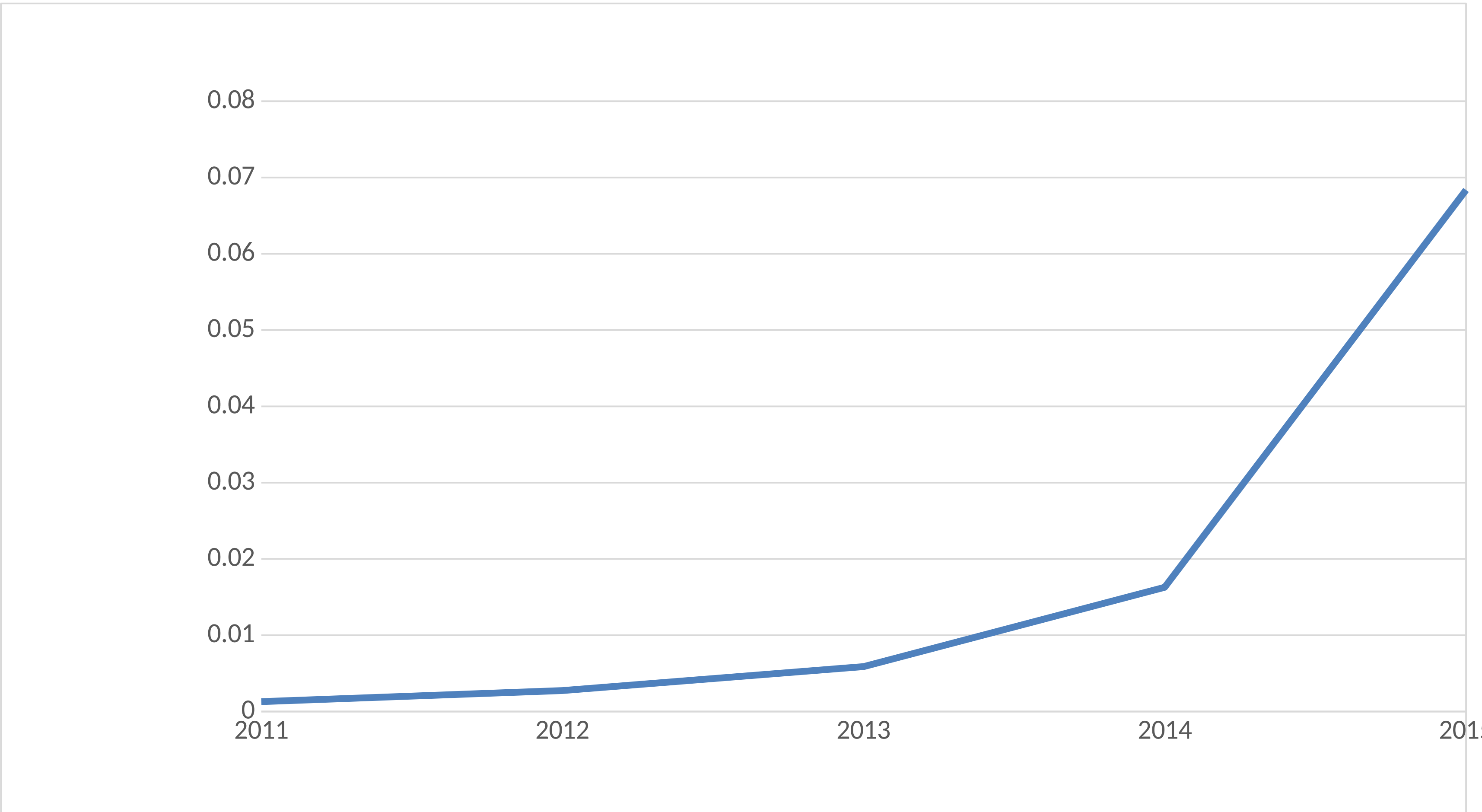


Prediction of research topics

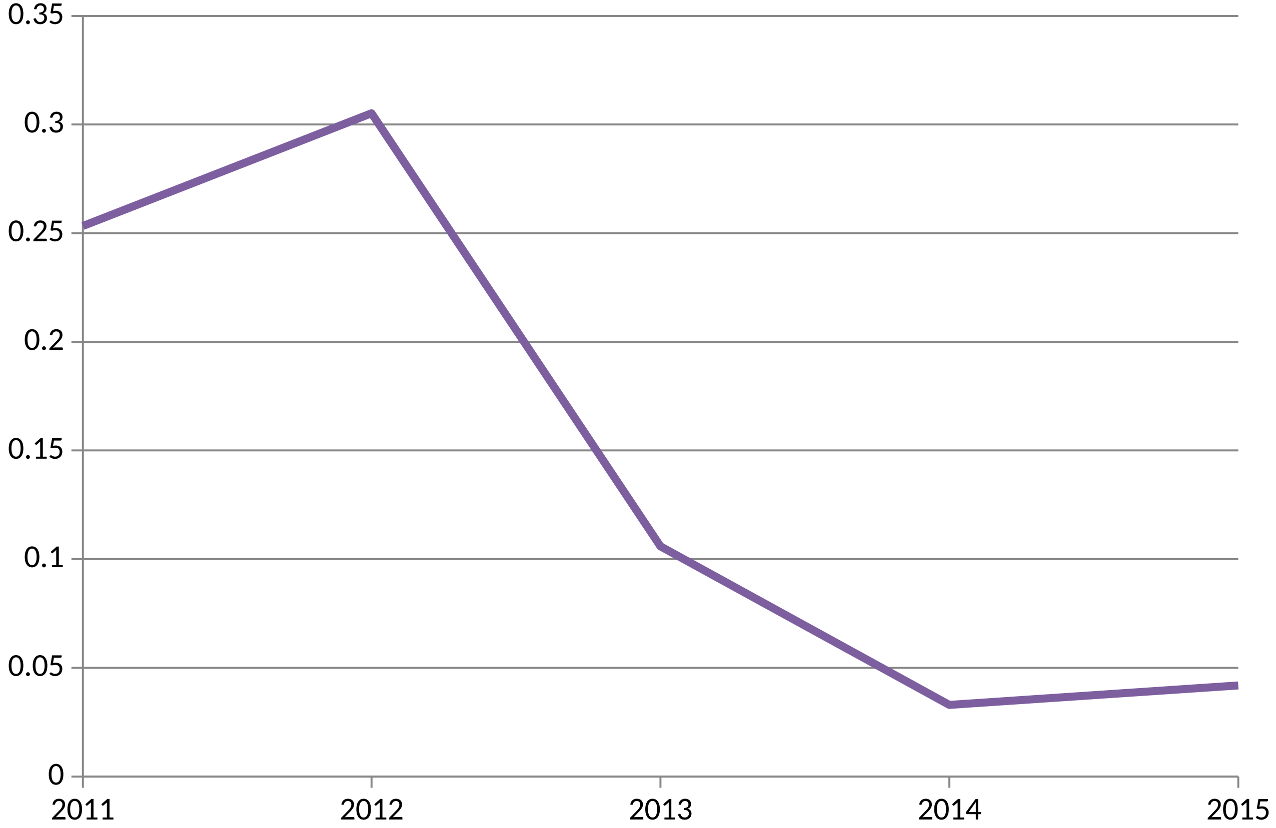
Topic Prediction
(Weka ML software package)

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

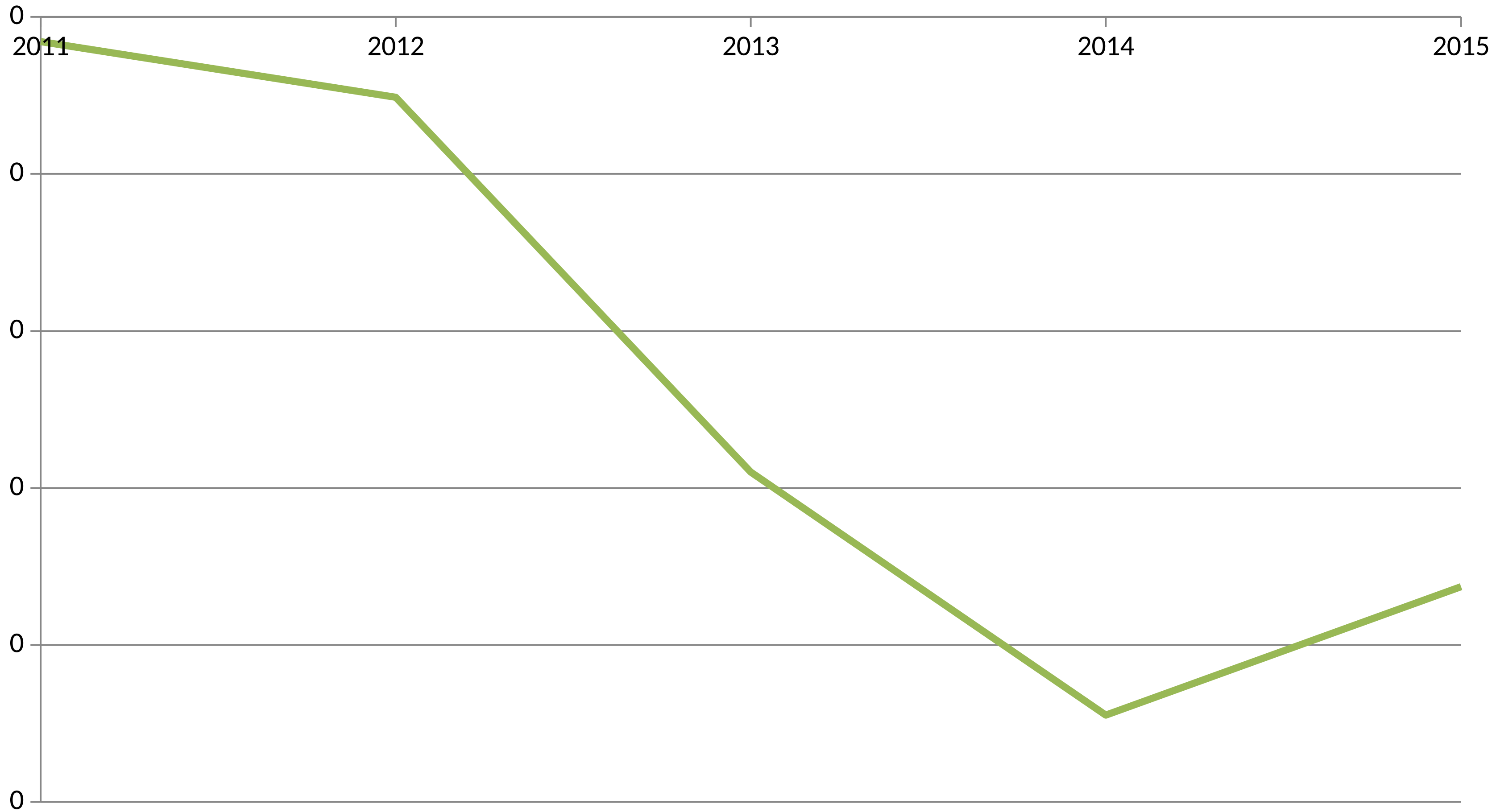
Prediction reliability:
Prediction errors from 2010



Surprises: Epistemological Ruptures



Topic emergence: DNN



Predictions 2015 for next 5 years

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

Use of Language Resources

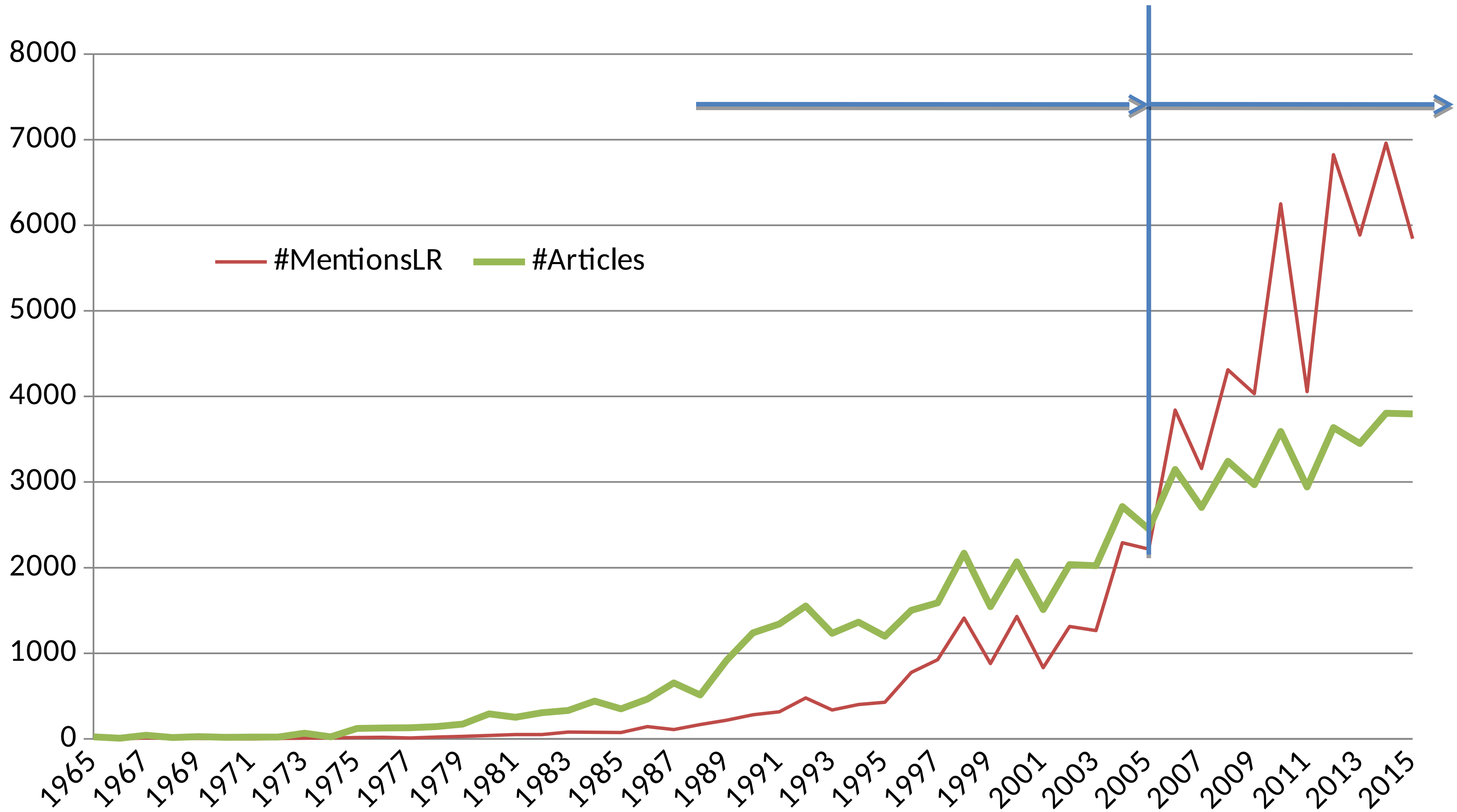
LRE Map

- Language Resources and Evaluation Map
  - Launched in 2010 to identify LRs (data, tools, evaluation, meta-resources) and their use
  - Contains actual data provided by the community at conferences through an online questionnaire
- Use of LRE Map 2014
  - 12 conferences (LREC, COLING, EMNLP, ACL-HLT, IJCNLP, Interspeech, LTC, Oriental-Cocosda, RANLP)
  - 4395 entries, 3121 different LRs, 2747 families of LRs

Mentions of the LRE Map LR in papers

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

LRE Map LR citation over time



Reuse and Plagiarism

Documents Comparison

- 67,937 x 67,937 papers

- After automatic text processing
  - handling typographical (hyphen, caesura, case…), lexical (orthographic variants, abbreviations,…), syntactic information (parsing)
- Analyze paper content and References (Authors' names, Title, Source)
- Use *Jaccard Distance* to compare documents

Documents Comparison

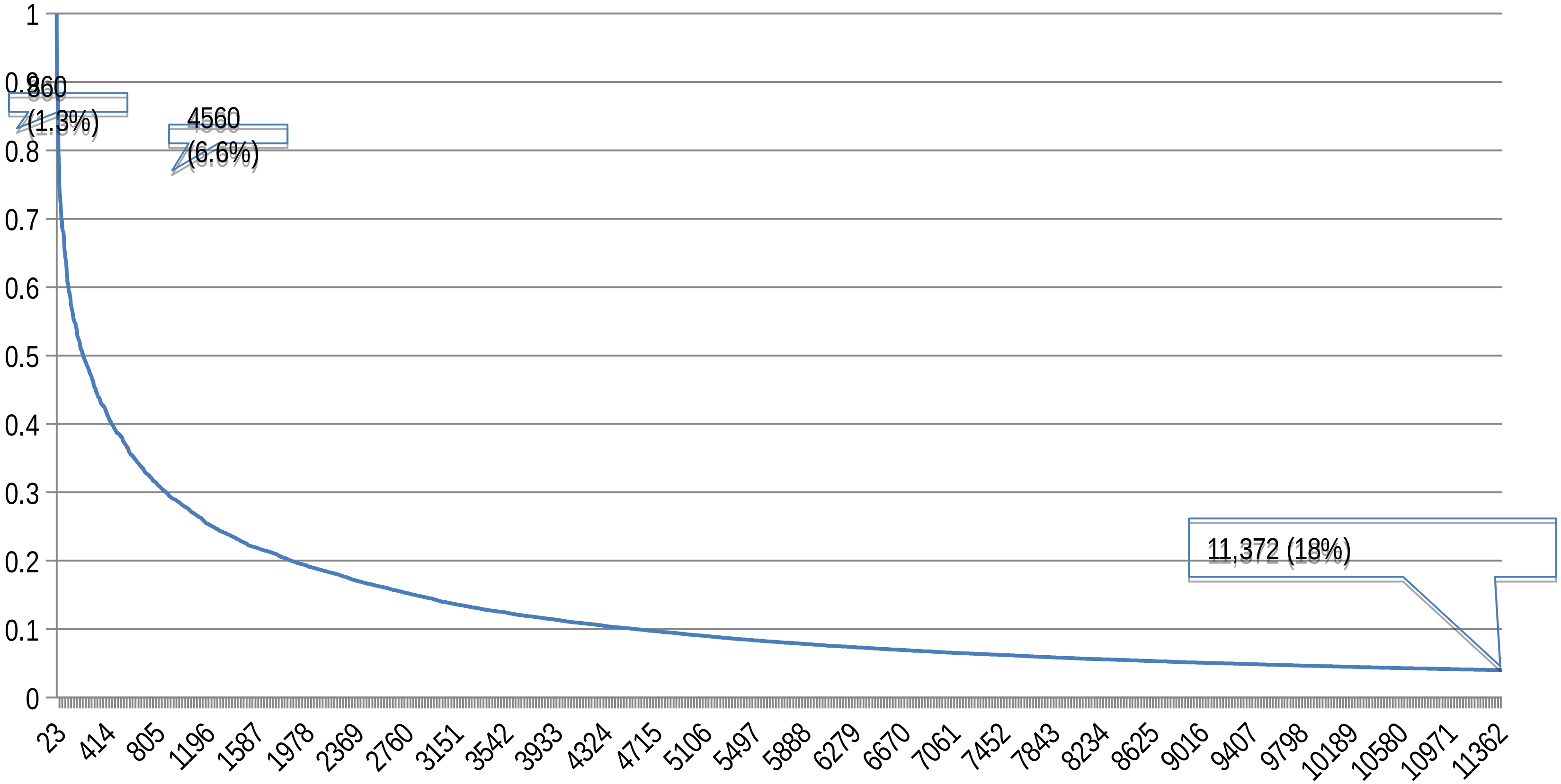
ıuous dynamics of the signal within a state An alternative approach is segmental modeling where the basic modeling unit is not a
ıetic unit this family of models relax both the stationarity and the independence within a state assumptions of standard HMM s in
eview major variants of segmental models A more detailed survey of segmental models can be found in 20 Goldberger et al
ling 265 Deng et al 1 used a regression polynomial function of time to model the trajectory of the mean in each state A similar
ested by Gish and Ng 9 for a keywords spotting task in that model the observation vectors within a state are generated according
t to zero at the beginning of the state and then incremented with each new incoming frame are state dependent vector parameters
ın Gaussian with a state dependent diagonal covariance matrix the case corresponds to standard HMM this model assumes that
ı a state are independently although not identically distributed Russell and Holmes 12 14 23 and Gales and Young 6 7 extended
sted by Deng by assuming a parametric segmental model with random coefficients that are sampled once per segment realization
an trajectory is a stochastic process instead of a fixed parameter more precisely this model is defined by 1 and by the PDF s of
ł stage we create the observations by sampling along the parametric curve that was determined in the first stage this sampling is
the PDF of Diagonal covariance Gaussian PDF s are typically attributed to and in addition is assumed to have zero mean the
's can be normalized according to the segment length in order to achieve better performance and to simplify the parameter
enny et al 15 have used a state conditioned linear prediction coefficients LPC model to remove correlation between successive
ors i the observation vectors within a state are generated according to where are diagonal matrices so that a LPC model applies to
of the vector A disadvantage of the model is that it assumes stationarity within a state the two approaches of 1 and 15 were
ralized in 2 Digalakis 4 proposed a dynamical system model which generalizes the Gauss Markov model 2 to a Kalman filter
suming noisy observations the special case where the hidden Gauss Markov process is assumed to be constant was named target
arget state model is similar to the model proposed by Russell 23 therefore the dynamical system model can also be considered a
the hidden constant Gaussian mean target state model several authors have proposed nonparametric segment models A major
ıparametric models is that they are not sensitive to the shape of the feature trajectory that needs to be approximated consequently
sensitive to the segment partitioning problem that was explained in Section II and demonstrated in Fig 3 for a horizontal line
ximation on the other hand nonparametric models might require more data to train the model on since they are less constrained
models the first nonparametric approach to a nonstationary state HMM was the stochastic segment model SSM suggested by
oukos 18 in 1989 the SSM assigns a Gaussian distribution to the entire segment which is resampled to a fixed length A
proach to a nonstationary state HMM with an additional step of time warping was suggested by Ghitza and Sondhi 8 in 8 the
mean in a given state is set equal to that state realization in the training set whose dynamic time warping DTW distance 24 from
es in the ensemble is minimal more recently Kimball et al 16 20 suggested a nonparametric approach that models each segment
ture of nonparametric mean trajectories Direct implementation of segmental models is typically computationally demanding this
that the exact beginning and ending points of the segment must be given in order to compute an acoustic score the best paradigm
on to this problem by using the following two stage recognition procedure at the first stage a standard HMM recognition system
e a list of size of best hypothesized candidate strings with the associated acoustic segmentation of each hypothesis at the second
ormative segmental acoustic model is used to rescore these candidates essentially the best paradigm takes advantage of the
ficiency of standard HMM recognition Continuous mixture of Nonparametric Segmental models in this section we present a new

assumption the joint observation probability can be rewı
TT qopqqoopqop although the frame independence assu
clearly inappropriate for speech sounds the standard HN
has worked extremely well for various types of speech ı
tasks review of Research efforts ON frame Correlation ı
maximum likelihood Ml criteria the performance of a H
system relies on how well the HMMs can characterize t
real speech for this reason various approaches have beeı
account of frame correlation for more realistic modeling
are generally known by the name of frame correlation n
family of segment models tries to directly express speec
trajectories the basic modeling unit is not a frame but a
this family of models relaxes both the stationarity and tł
independence assumptions within a standard HMM stat
seem to be successful in extracting dynamic cues for spe
recognition under a suitable trajectory assumption they
on widely availiable HMM technology Deng et al 6 use
polynomial function of time to model the trajectory of tł
each state A similar model was suggested by Gish and N
keyword spotting task Russell and Holmes and Gales ar
extended the model suggested by Deng by assuming a p
segmental model with random coefficients that are samp
segment realization therefore the mean trajectory is a st
process instead of a fixed parameter Digalakis 9 propos
system model which generalizes the Gauss Markov moc
filter framework by assuming noisy observations severa
proposed nonparameteric segment models A major adva
nonparametric models is that they are not sensitive to th
feature trajectory that needs to be approximated consequ
also not sensitive to the segment partition problem on th
nonparameteric models might require more data to train
since they are less constrained that parametric models tł
nonparametric approach to a nonstationary state HMM
stochastic segment model SSM suggested by Ostendorf

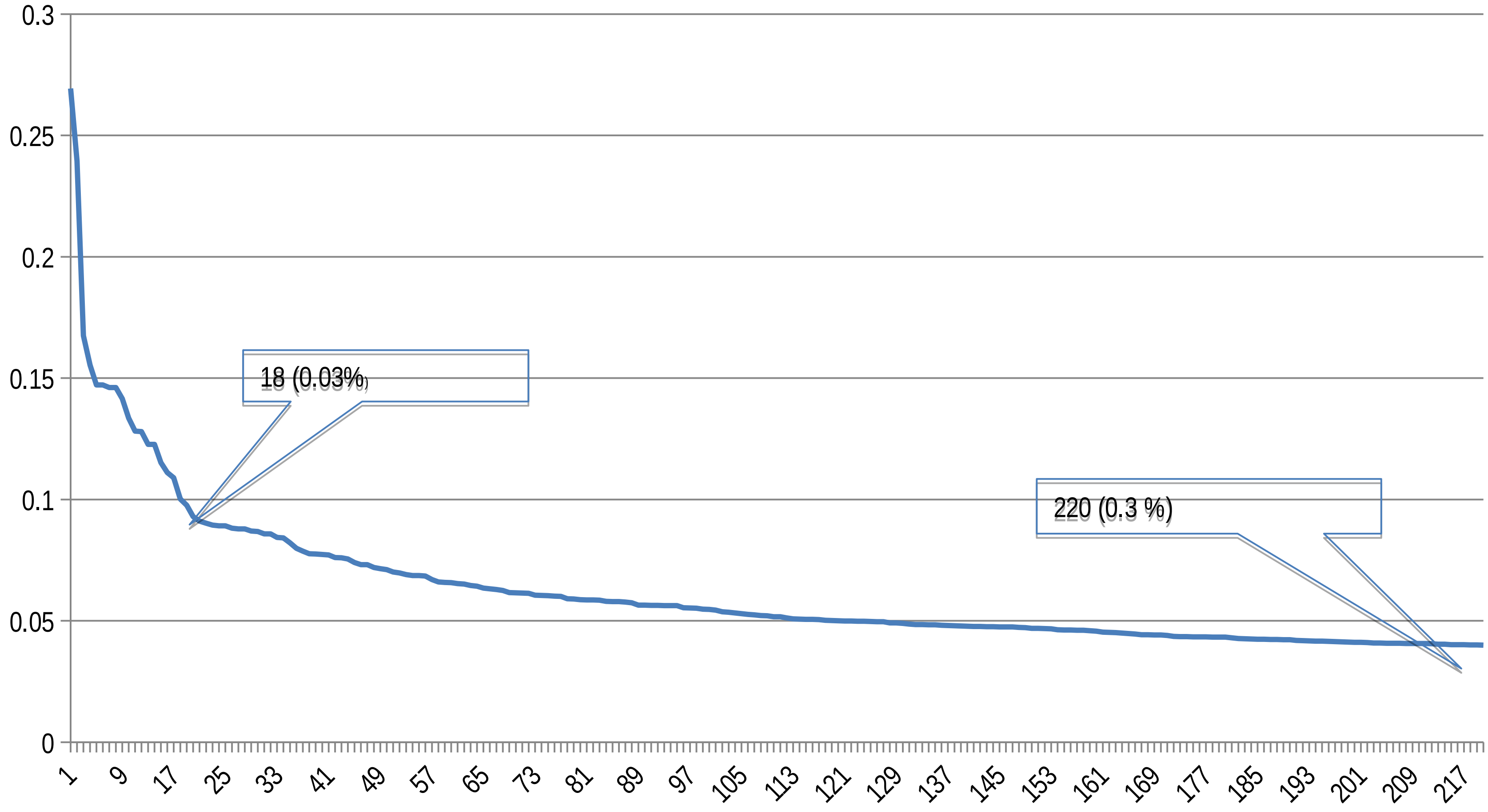
(Self-) Reuse and Plagiarism

>4% similarity	Source is quoted	Source is not quoted	Legally / Ethically acceptable
At least one author in both papers	Self-Reuse	Self-Plagiarism	30%
No author in common	Reuse	Plagiarism	10% (FR, CAN)

Self-Reuse and Self-Plagiarism



Reuse and Plagiarism



Manual checking

- Qing Guo, Fang Zheng, Jian Wu, and Wenhu Wu, Non-Linear Probability Estimation Method Used in HMM for Modeling Frame Correlation (ISCA-Interspeech 1998)
- Guo Qing, Zheng Fang, Wu Jian and Wu Wenhu, An New Method Used in HMM for Modeling Frame Correlation (IEEE-ICASSP 1999)

-

- Quoted: Graham W. (2007) “an OWL Ontology for HPSG”, proceeding of the ACL 2007 demo and poster sessions, 169-172.
- Correct: Graham Wilcock (2007), “An OWL Ontology for HPSG”, proceeding of the ACL 2007 demo and poster sessions, 169-172.

-

- Quoted: Li Liu, Jianglong He, “On the use of orthogonal GMM in speaker verification”
- Correct: Li Liu and Jialong He, “On the use of orthogonal GMM in speaker recognition”

Marie Skłodowska-Curie Actions
Innovative Training Networks (ITN)

European Joint doctorate (EJD)

H2020-2015

METHODS IN RESEARCH ON RESEARCH
MIROR

ESR 11. Assisted authoring for avoiding inadequate claims in scientific reporting

Anna Koroleva

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 676207

Inadequate reporting: why is it an important problem?

- Focus: randomized controlled trials (RCTs) assessing an intervention
- Inadequate reporting (spin): presentation of the experimental treatment as more effective/safe than the research has proved.
- Impact: overestimation of the beneficial effect of the experimental treatment by physicians, patients, media[1,2].
- Prevalence: present in abstracts of 60% of reported randomized controlled trials (RCTs)[1].

Main project objective: create Natural Language Processing (NLP) algorithms to detect spin automatically.

[1] Boutron I., Altman D.G., Hopewell S., Vera-Badillo F., Tannock I., Ravaud P. Impact of spin in the abstracts of articles reporting results of randomized controlled trials in the field of Cancer: the SPIIN randomized controlled trial. J Clin Oncol. 2014;32:4120–4126.

[2] Yavchitz A., Boutron I., Bafeta A., Marroun I., Charles P., Mantz J., et al. Misrepresentation of randomized controlled trials in press releases and news coverage: a cohort study. PLoS Med. 2012;9:e1001308.

Types of spin

Misleading reporting of results:

- not reporting adverse events;
- **selective reporting of outcomes** (omission of primary outcome; focus on statistically significant secondary outcomes, subgroup or within-group analyses);
- misleading reporting of study design;
- **linguistic spin**;
- no consideration of limitations;
- selective citation of other studies.

Inadequate interpretation of results:

- **claiming a beneficial or equivalent effect of the intervention for statistically non-significant results**;
- claiming that the treatment is safe for statistically non-significant safety outcomes;
- concluding a beneficial effect despite no comparison test performed;
- interpretation of the results according to statistical significance instead of clinical relevance.

Inadequate extrapolation:

- inadequate extrapolation from the population, interventions or outcome actually assessed in the study to a larger population, different interventions or outcomes;
- inadequate implications for clinical practice.

NLP algorithms: our results

Information extraction: claims supporting information

- Methods: rule-based approach; finite state automata
- Baseline approach implemented; to be used for corpus pre-annotation

1. Outcomes / objectives

*We chose <Out Type="Prim">**housing status**</Out> as the main effectiveness measure.*

*The primary efficacy scale was <Out Type="Prim">**the CGI Severity of Illness scale**</Out> (CGI- Severity).*

*<Out Type="Prim">**The BPRS Anxiety/Depression factor (ANDP)**</Out> was used as the primary measure of depression in this study.*

2. Patient population / subgroups

*Carbamazepine as adjunctive treatment in <Subj>**nonepileptic chronic inpatients with EEG temporal lobe abnormalities**</Subj>.*

*The first author of this paper defined a treatment manual for BPT with <Subj>**schizophrenia patients suffering from persistent negative symptoms**</Subj>.*

3. Statistical measures (p-value, confidence intervals)

*There was a significant difference in the mean endpoint CGI-I score, with modafinil-treated subjects having greater improvement (mean CGI-I score, 3.2 vs. 4.1; t = 3.35, df = 18, <StatMeas Type="Pval">**p = .004**</StatMeas>).*

Corpus creation

1. Corpus of PMC articles collected in LIMSI (3,938 RCTs / 65,396 articles)

2. Secondment in the Cochrane Schizophrenia group (Nottingham, the UK)

22,978 PDF files (full-text articles, abstracts, registry entries)

Extract text from PDF

Extraction quality assessment

Identifying full-text articles that include abstracts

~5,000 articles

~1,000 articles with trial registration numbers

~1,300 articles with outcomes identifiable automatically

Conclusions & Perspectives

- Large analysis of bibliographical data in a specific scientific domain (NLP)
- Problem with quality of data
  - Early papers (1960s)
  - Contextual Term extraction
- Improve measure of innovation
- Analyze citation polarity

Conclusions & Perspectives

- Problem with information identification
  - Authors Names
  - Laboratories Names
  - Papers Title
  - Journals and Conferences Names
  - Names of Funding agencies
  - Language Resources Names, etc.
- Needs a tedious manual cleaning
- Would necessitate an international coordination action for assigning unique and persistent identifiers to data (cf ISLRN for LR)

le dernier mot

- *croissance dans toutes les dimensions (articles, auteurs, citations...)*
- *besoin de normalisation (identifiants, auteurs, ressources)*
- *vers une automatisation de l'évaluation des articles*
- *MAIS l'expertise humaine est toujours requise pour la validation*