Constitution et exploitation du corpus NLP4NLP pour l'analyse bibliométrique de 50 ans de recherches en traitement automatique de la parole et du langage naturel.

Joseph Mariani[1], Gil Francopoulo[2], Patrick Paroubek[1]

[1]LIMSI-CNRS, [2]Tagmatica

Objectif

- Utiliser les outils du Traitement Automatique du Langage Naturel (TAL) pour analyser la bibliographie en Traitement Automatique du Langage Naturel
  - Etude de la conférence IEEE ICASSP (1976-1990) (1991)
  - Workshop *Rediscovering 50 Years of Discoveries in NLP* (50ème anniversaire de l'ACL, Conf. ACL, Jeju, 2012)
  - *Rediscovering **25 Years of Discoveries in SLP*** (25ème anniversaire de l'ISCA, Conf. Interspeech 2013, Lyon)
  - *Rediscovering **15 Years of Discoveries in LRE*** (15ème anniversaire de LREC, Conf. LREC 2014, Reykjavik)
    - *Rediscovering 15+2 Years of Discoveries in LRE* (Journal LRE, Mars 2016)
  - *Rediscovering **10 to 20 Years of Discoveries in L&TC*** (20ème anniversaire de la L&TC, Poznan, 2015)
- Etendre à un demi siècle de recherches en TAL

Sujet d'actualité

- *Workshop on **Mining Scientific Publications*** (WOSP'2015)
  - Fort Knox, 24-25 Juin 2015
  - D-Lib Magazine **(**Nov./Dec. 2015, Vol. 21, N° 11/12)
- *Workshop on **Computational Linguistics and Bibliometrics*** (CLBib)
  - 15ème *Intal Society of Scientometrics and Informetrics Conference* (ISSI)
  - Istanbul, 29 Juin 2015
- *BIRNDL: Joint Workshop on **Bibliometric-enhanced IR** (BIR) and **NLP for digital libraries** (NLPIR4DL)*
  - ACM/IEEE Joint Conference on Digital Libraries'2016
  - Newark, 23 Juin 2016
  - Numéro spécial IJDL (Mars 2017)
- *Language Resources and Evaluation Conference* (LREC)
  - 16 articles à LREC 2016
  - Workshop spécifique à LREC 2018 ?

Corpus NLP4NLP

- Analyse du domaine du TAL (langue écrite, langue parlée, langue signée, Recherche d'Information)
- 34 publications sur 50 années (1965-2015)
- Conférences (ACL, IEEE-ICASSP, ISCA-Interspeech, ELRA-LREC, etc.) et revues (IEEE-TASLP, CL, SpeechCom, CSAL, LRE, etc.)
- 558 évènements
  - Tenue d'une conférence
  - Numéro de revue
- 65,003 articles
- 48,894 auteurs différents
- 270 Mmots
- 324,422 références bibliographiques

Corpus NLP4NLP

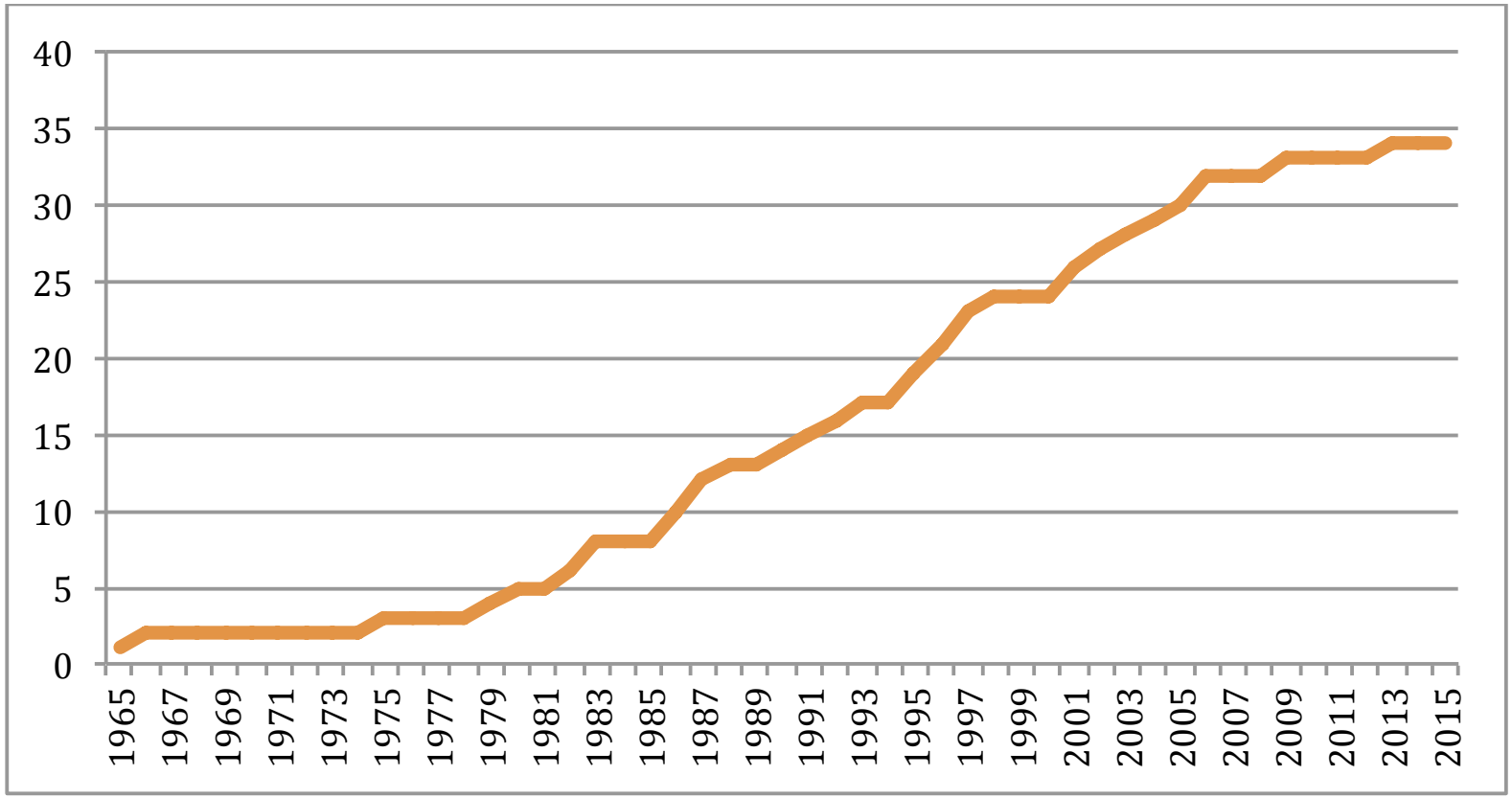
Nom court	# docs	Type	Nom long	Langue	Accès contenu	Période	# évènements
acl	4264	Conférence	Association for Computational Linguistics Conférence	Anglais	Libre accès *	1979-2015	37
acmtslp	82	Revue	ACM Transaction on Speech and Language Processing	Anglais	Propriétaire	2004-2013	10
alta	262	Conférence	Australasian Language Technology Association	Anglais	Libre accès *	2003-2014	12
anlp	278	Conférence	Applied Natural Language Processing	Anglais	Libre accès *	1983-2000	6
cath	932	Revue	Computers and the Humanities	Anglais	Propriétaire	1966-2004	39
cl	776	Revue	American Revue of Computational Linguistics	Anglais	Libre accès *	1980-2014	35
coling	3813	Conférence	Conférence on Computational Linguistics	Anglais	Libre accès *	1965-2014	21
conll	842	Conférence	Computational Natural Language Learning	Anglais	Libre accès *	1997-2015	18
csal	762	Revue	Computer Speech and Language	Anglais	Propriétaire	1986-2015	29
eacl	900	Conférence	European Chapter of the ACL	Anglais	Libre accès *	1983-2014	14
emnlp	2020	Conférence	Empirical methods in natural language processing	Anglais	Libre accès *	1996-2015	20
hlt	2219	Conférence	Human Language Technology	Anglais	Libre accès *	1986-2015	19
icassps	9819	Conférence	IEEE International Conférence on Acoustics, Speech and Signal Processing - Speech Track	Anglais	Propriétaire	1990-2015	26
ijcnlp	1188	Conférence	International Joint Conférence on NLP	Anglais	Libre accès *	2005-2015	6
inlg	227	Conférence	International Conférence on Natural Language Generation	Anglais	Libre accès *	1996-2014	7
isca	18369	Conférence	International Speech Communication Association	Anglais	Libre accès	1987-2015	28
jep	507	Conférence	Journées d'Etudes sur la Parole	Français	Libre accès *	2002-2014	5
lre	308	Revue	Language Resources and Evaluation	Anglais	Propriétaire	2005-2015	11
lrec	4552	Conférence	Language Resources and Evaluation Conférence	Anglais	Libre accès *	1998-2014	9
ltc	656	Conférence	Language and Technology Conférence	Anglais	Propriétaire	1995-2015	7
modulad	232	Revue	Le Monde des Utilisateurs de L'Analyse des Données	Français	Libre accès	1988-2010	23
mts	796	Conférence	Machine Translation Summit	Anglais	Libre accès	1987-2015	15
muc	149	Conférence	Message Understanding Conférence	Anglais	Libre accès *	1991-1998	5
naacl	1186	Conférence	North American Chapter of the ACL	Anglais	Libre accès *	2000-2015	11
paclic	1040	Conférence	Pacific Asia Conférence on Language, Information and Computation	Anglais	Libre accès *	1995-2014	19
ranlp	363	Conférence	Recent Advances in Natural Language Processing	Anglais	Libre accès *	2009-2013	3
sem	950	Conférence	Lexical and Computational Semantics / Semantic Evaluation	Anglais	Libre accès *	2001-2015	8
speechc	593	Revue	Speech Communication	Anglais	Propriétaire	1982-2015	34
tacl	92	Revue	Transactions of the Association for Computational Linguistics	Anglais	Libre accès *	2013-2015	3
tal	177	Revue	Revue Traitement Automatique du Langage	Français	Libre accès	2006-2015	10
taln	1019	Conférence	Traitement Automatique du Langage Naturel	Français	Libre accès *	1997-2015	19
taslp	6612	Revue	IEEE/ACM Transactions on Audio, Speech and Language Processing	Anglais	Propriétaire	1975-2015	41
tipster	105	Conférence	Tipster DARPA text program	Anglais	Libre accès *	1993-1998	3
trec	1847	Conférence	Text Retrieval Conférence	Anglais	Libre accès	1992-2015	24
Total avec doublons	67,937					1965-2015	577
Total sans doublons	65,003						558

Traitements des données

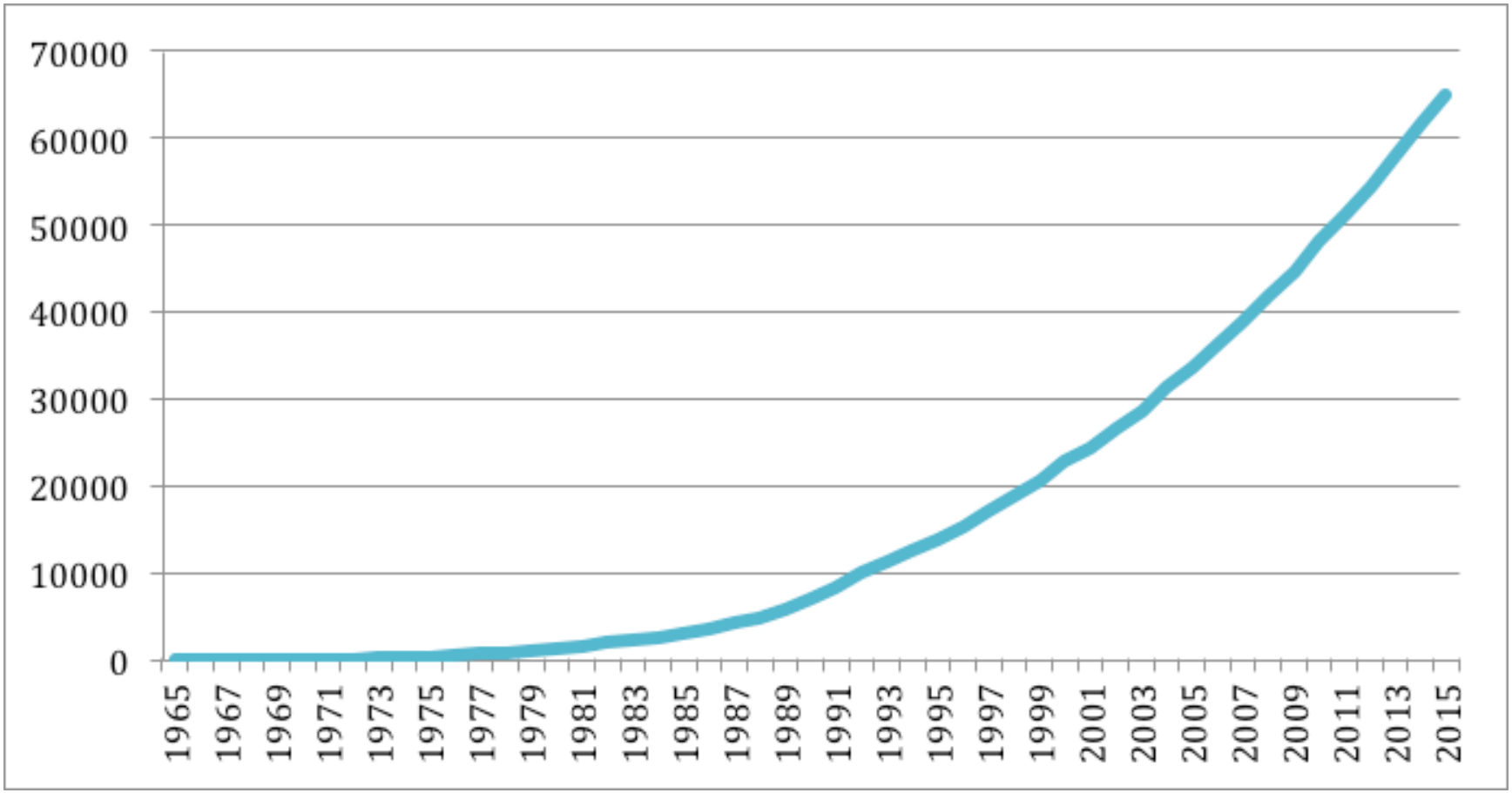
- Texte sous forme électronique ou papier
  - OCR
- Existence de méta-données
- Extraction automatique
  - Noms d'auteurs différents
    - Affiliation, nationalité, genre
  - Termes scientifiques
  - Ressources linguistiques
  - Citations
    - Auteurs, titres, sources
  - Agences de financement,…

Production

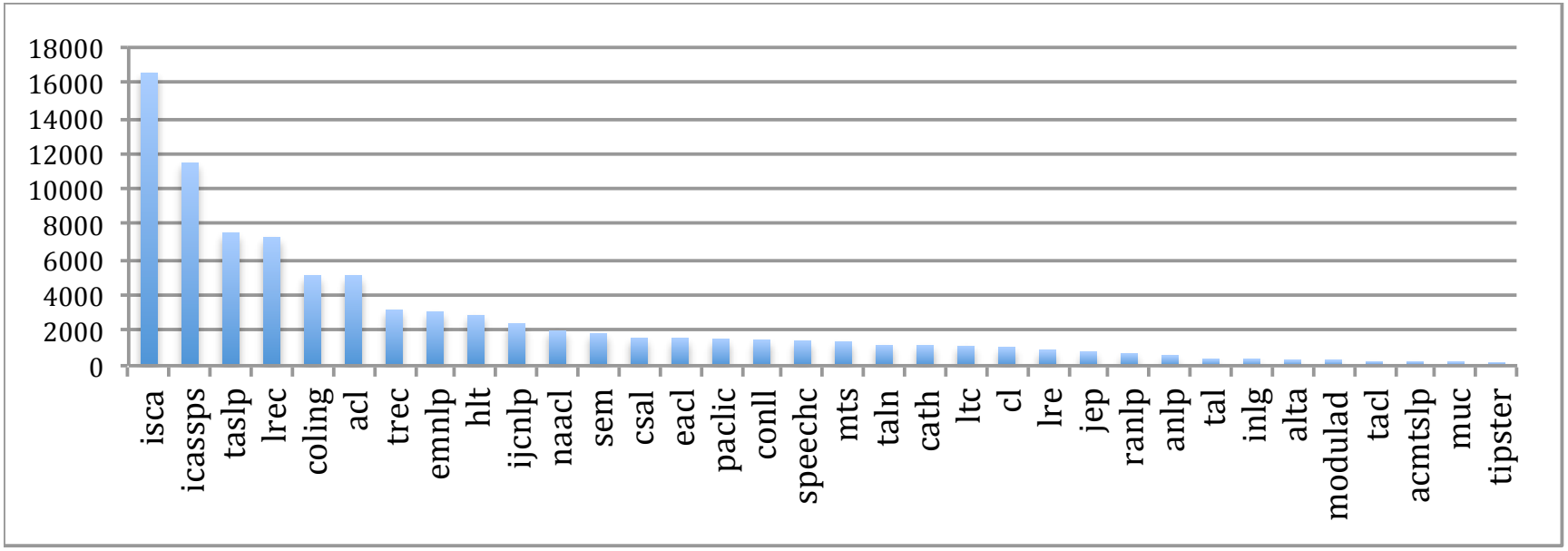
Nombre cumulé de publications différentes



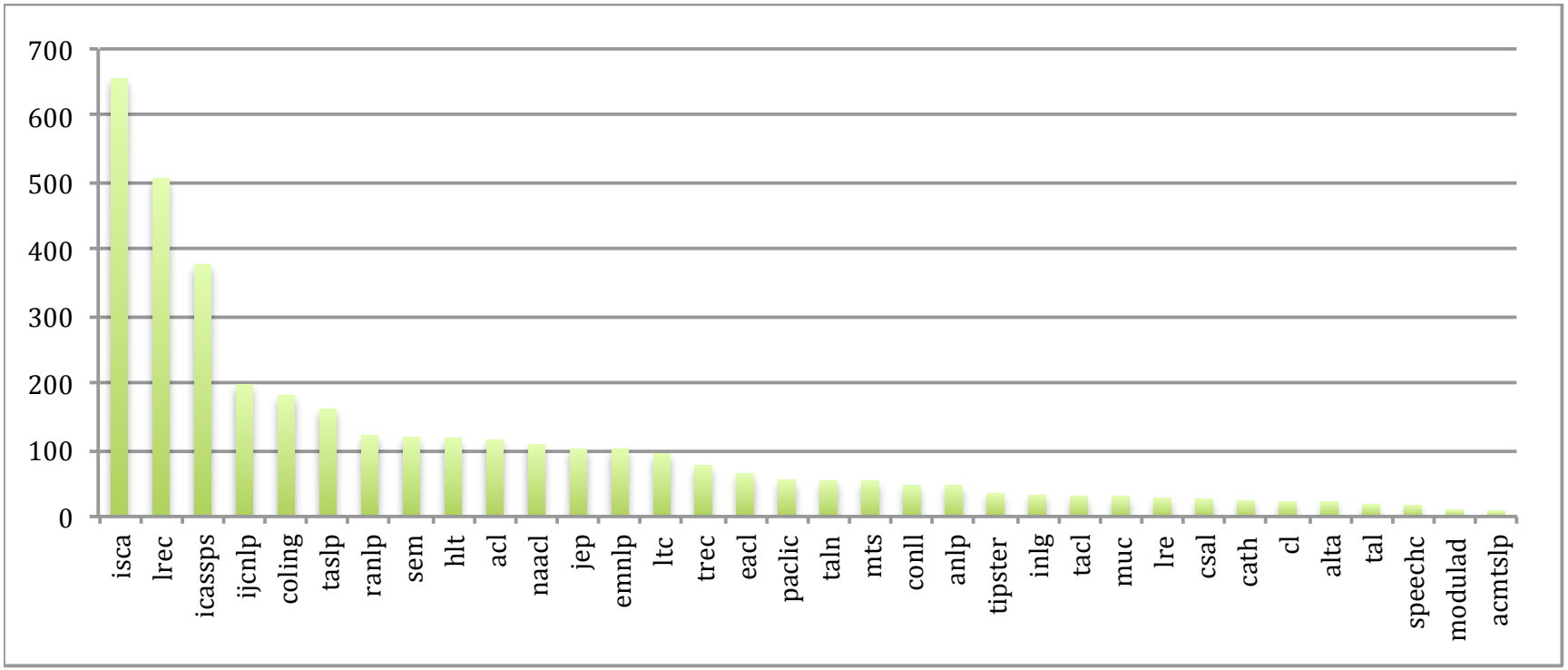
Nombre cumulé d'articles



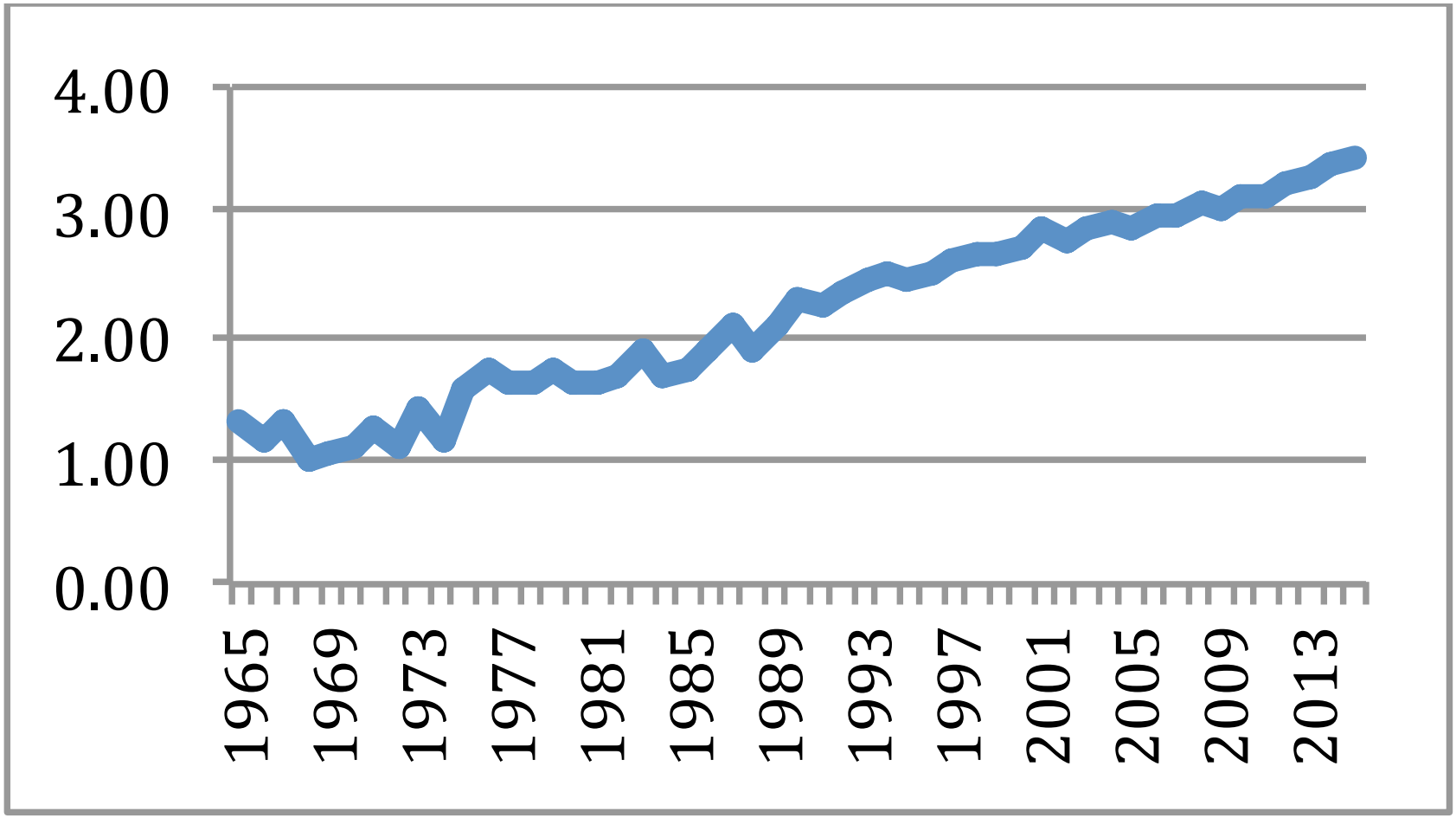
Nombre d'articles pour chacune des publications



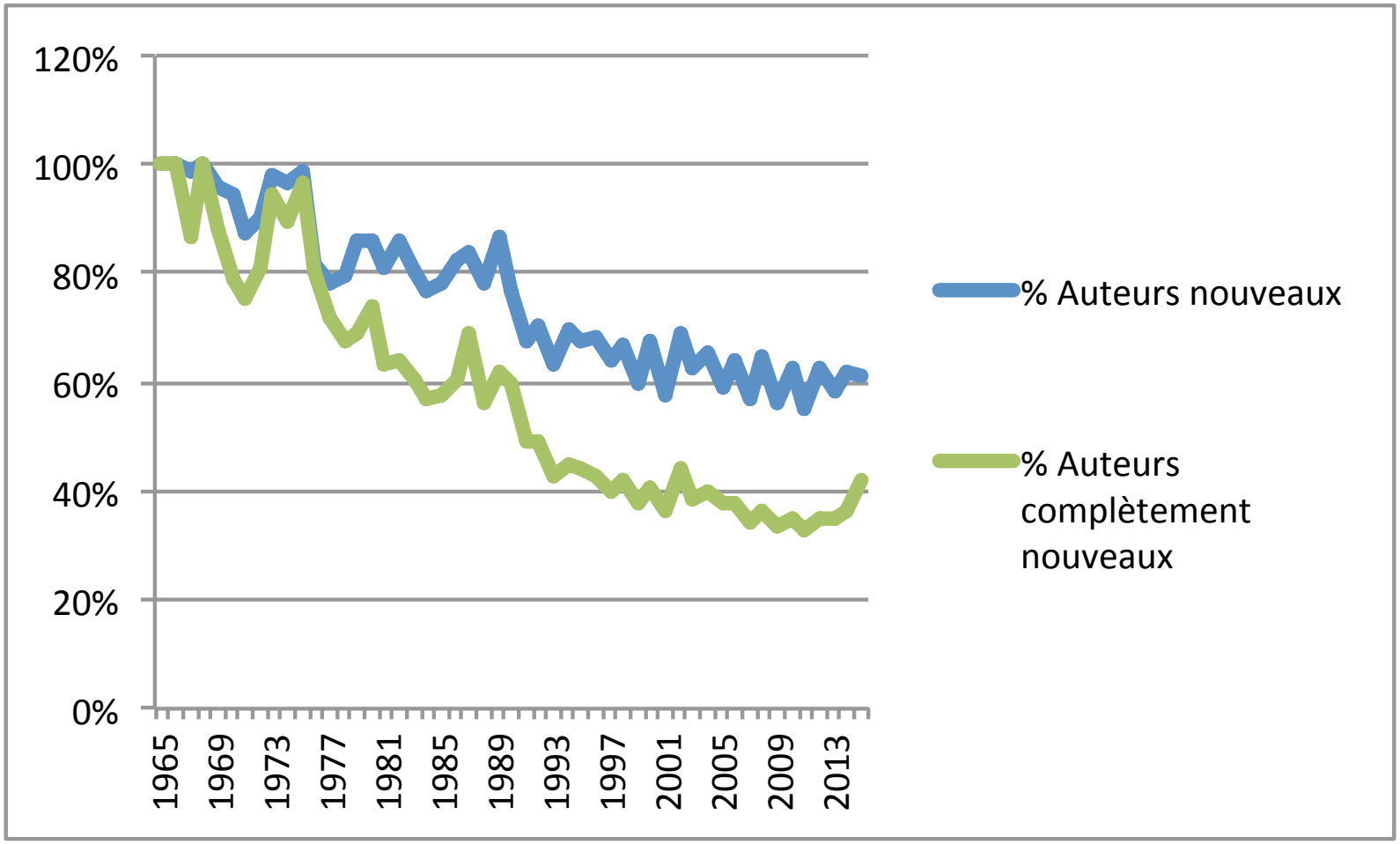
Nombre d'articles à chaque évènement



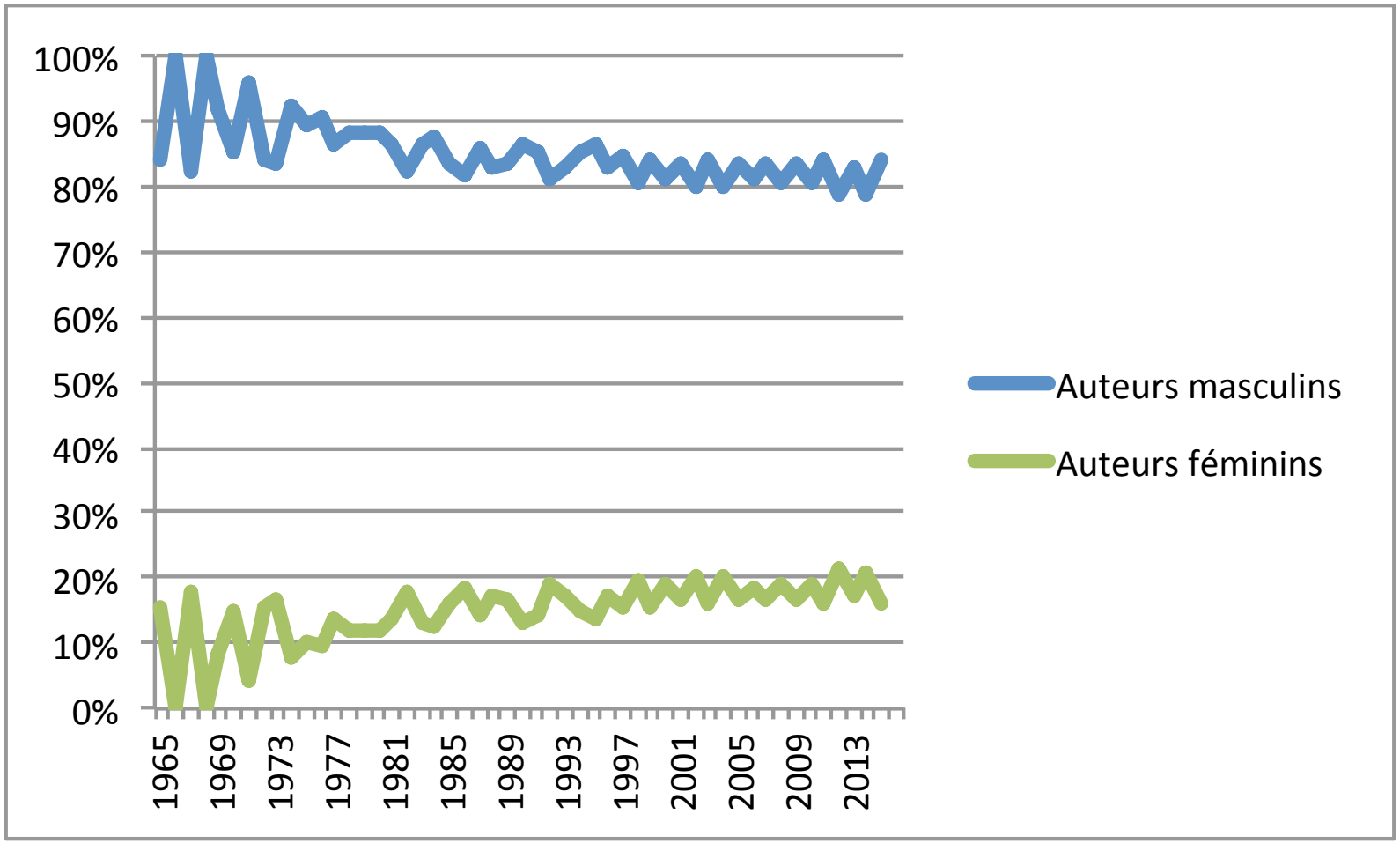
Nombre moyen d'auteurs par article au fil du temps



% de nouveaux auteurs



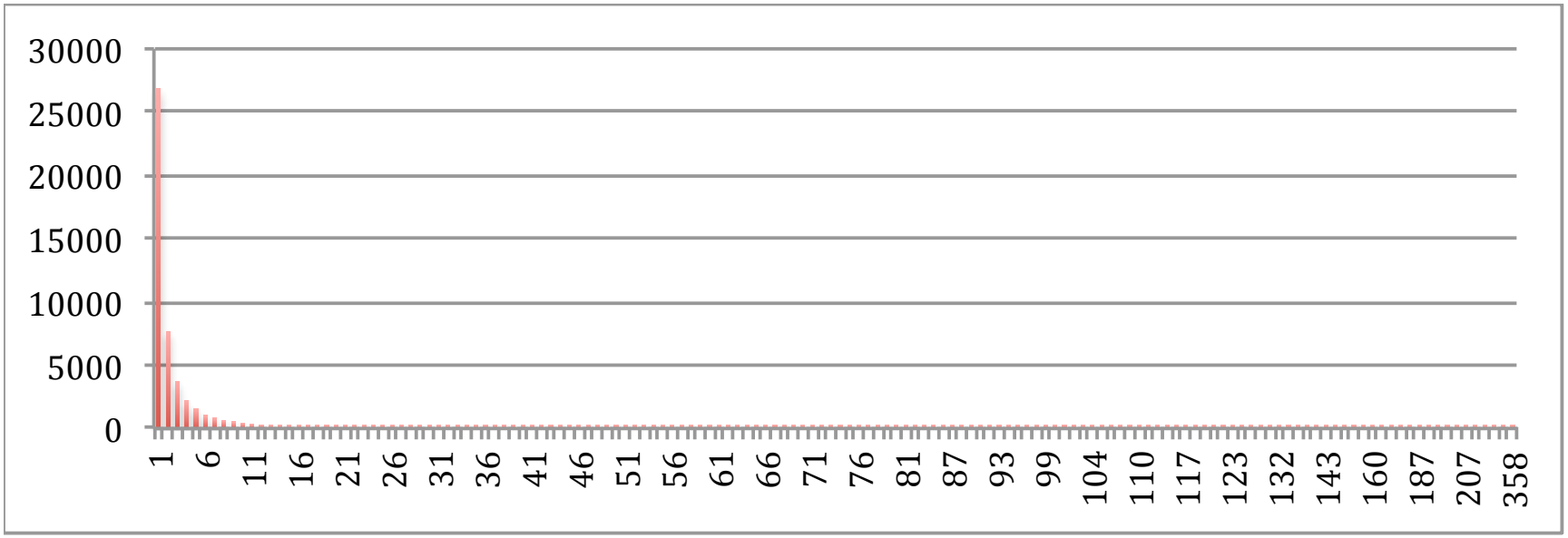
Evolution du genre des auteurs



Genre des auteurs selon les publications : % auteurs masculins

Collaborations entre auteurs

Nombre d'articles par auteur



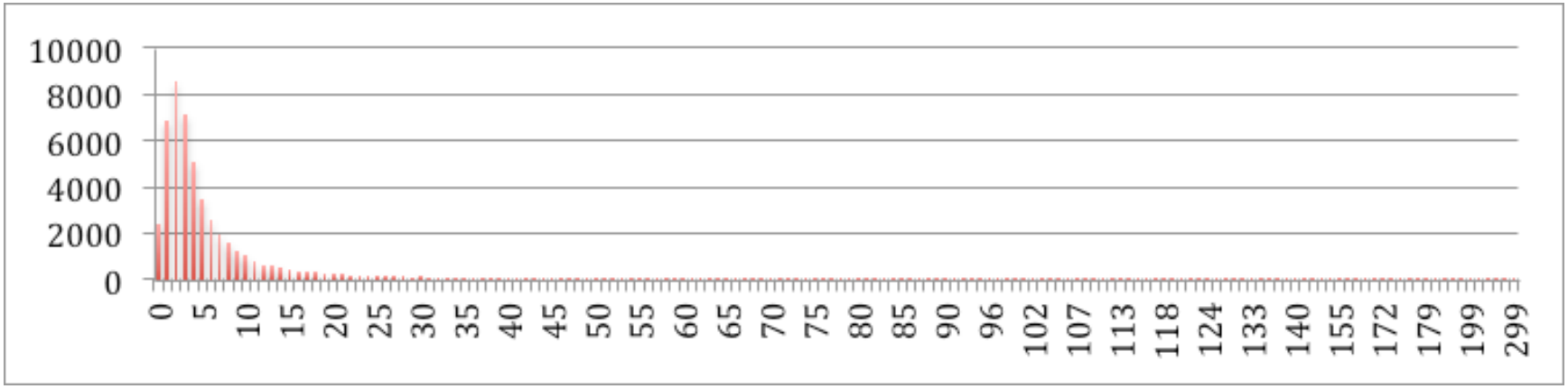
Nombre d'articles

Nom	Nombre d'articles	Nombre d'articles comme auteur unique
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

Nombre d'articles signés comme seul auteur

# d'articles	# d'auteurs	Nom des auteurs
0	42,471	…
1	4402	…
2	1038	…
3	416	…
4	211	…
5	131	…
6	76	…
7	49	…
8	27	…
9	24	…
10	10	Aravind K Joshi, Eckhard Bick, Hermann Ney, Hugo Van Hamme, Joshua T Goodman, Karen Spärck Jones, Kuldip K Paliwal, Mark Hepple, Raymond S Tomlinson, Roger K Moore
11	10	Dekang Lin, Eduard H Hovy, Jörg Tiedemann, Marius A Pasca, Michael Schiehlen, Olov Engwall, Patrick Saint-Dizier, Philippe Blache, Stephanie Seneff, Tomek Strzalkowski
12	9	David S Pallett, Harvey F Silverman, Jen-Tzung Chien, Kenneth Ward Church, Lynette Hirschman, Martin Kay, Reinhard Rapp, Ted Pedersen, Yorick Wilks
13	4	John Makhoul, Paul S Jacobs, Rens Bod, Robert C Moore
14	2	Dominique Desbois, Sadaoki Furui
15	2	Donna Harman, Takayuki Arai
16	2	Jerry R Hobbs, Steven M Kay
17	2	Beth M Sundheim, Kenneth C Litkowski
18	3	Douglas B Paul, Mark A Johnson, Rathinavelu Chengalvarayan
20	1	Olivier Ferret
21	1	Ralph Grishman
25	1	Ellen M Voorhees
26	1	Jerome R Bellegarda
27	1	W Nick Campbell

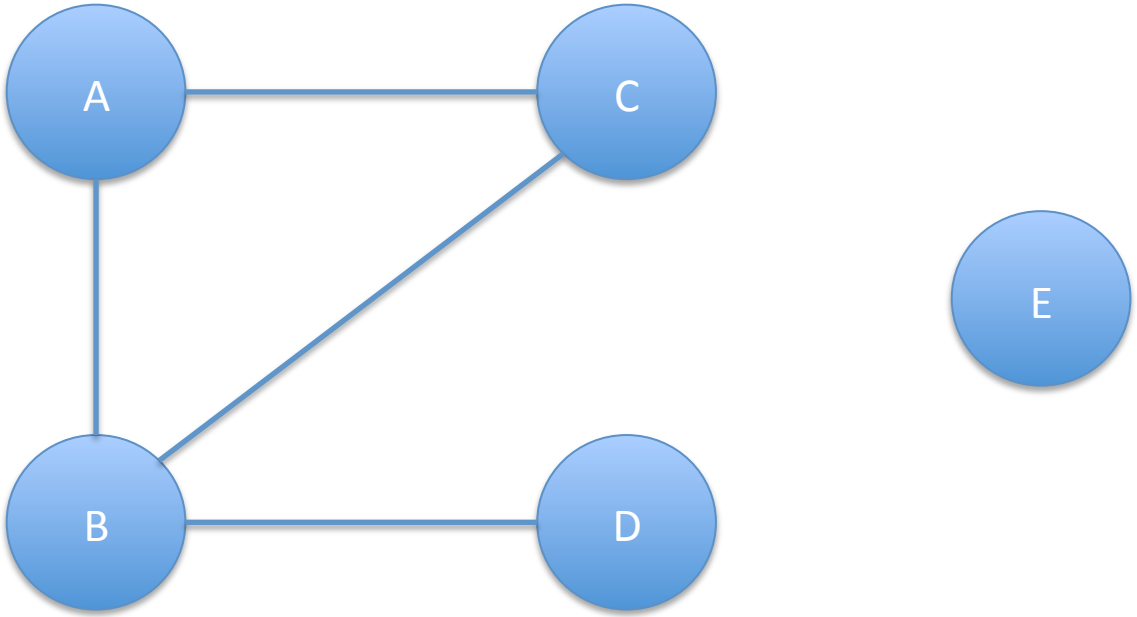
Nombre de co-auteurs



Nombre de co-auteurs

Nom	# Co-auteurs
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

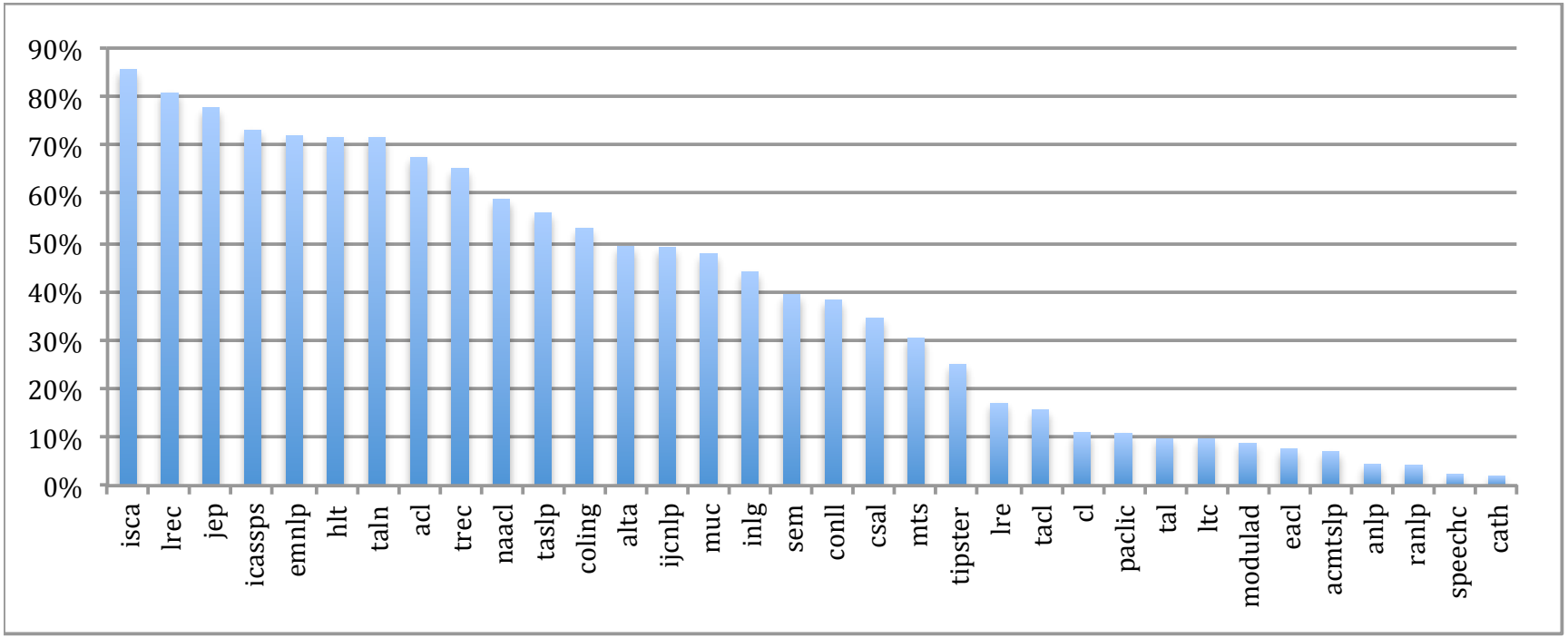
Graphe de collaboration



Graphe de collaboration: Composantes connexes

Taille de la Composante Connexe	Nombre de Composantes Connexes	Nombre d'auteurs dans les composantes connexes	% d'auteurs dans les composantes connexes	% des Composantes Connexes
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%

Graphes de collaboration : % d'auteurs dans la plus grande composante connexes par publication

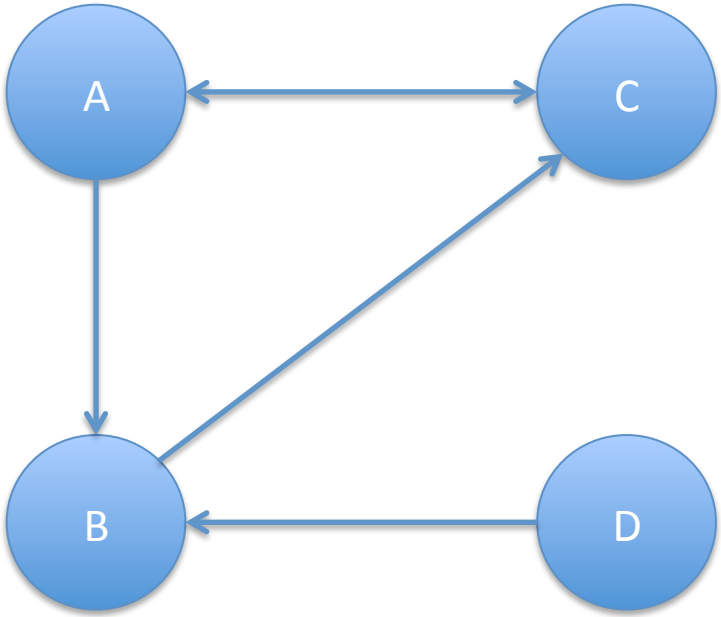


Graphe de collaboration : Centralité (LREC)

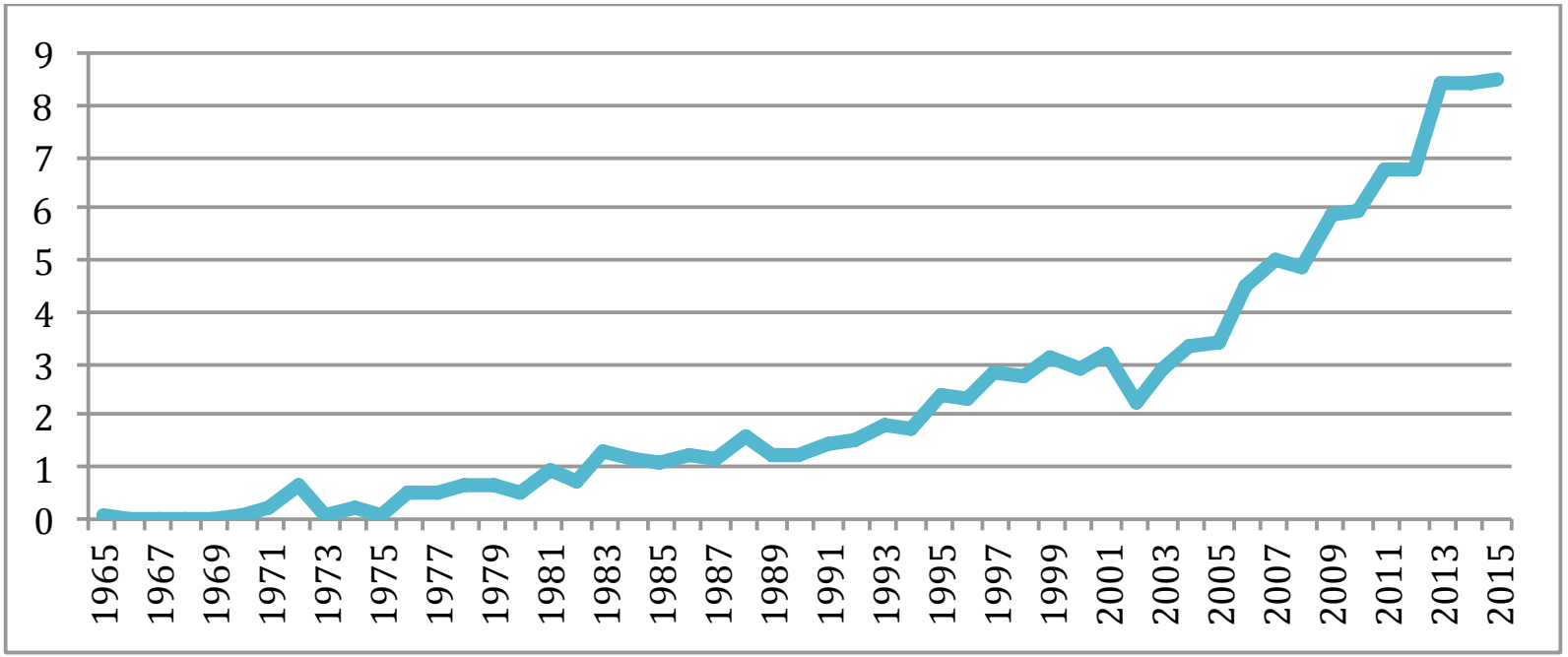
Centralité de Proximité			Centralité de Degré		Centralité d'Intermédiarité		
Nom de l'auteur	Centralité Harmonique	Valeur relative	Nom de l'auteur	Index & valeur relative	Nom de l'auteur	Index	Valeur relative
Nicoletta Calzolari	2,076	**1.000**	Nicoletta Calzolari	**1.000**	Khalid Choukri	269,538	**1.000**
Monica Monachini	1,996	**0.961**	Khalid Choukri	**0.944**	Nicoletta Calzolari	202,365	**0.751**
Khalid Choukri	1,983	**0.955**	Monica Monachini	**0.814**	Hans Uszkoreit	180,854	**0.671**
Núria Bel	1,980	**0.954**	Núria Bel	**0.739**	Núria Bel	158,669	**0.589**
Bernardo Magnini	1,941	**0.935**	Bernardo Magnini	**0.708**	Bernardo Magnini	157,090	**0.583**
Stelios Piperidis	1,933	**0.931**	Asunción Moreno	**0.689**	Asunción Moreno	151,440	**0.562**
Asunción Moreno	1,910	**0.920**	Hans Uszkoreit	**0.658**	Monica Monachini	144,944	**0.538**
Dan Tufiş	1,903	**0.917**	Stelios Piperidis	**0.609**	Martha Palmer	133,788	**0.496**
Joseph Mariani	1,893	**0.912**	Dan Tufiş	**0.540**	Ulrich Heid	133,446	**0.495**
Hans Uszkoreit	1,889	**0.910**	Jan Hajič	**0.534**	Stephanie M Strassel	120,573	**0.447**

Citations

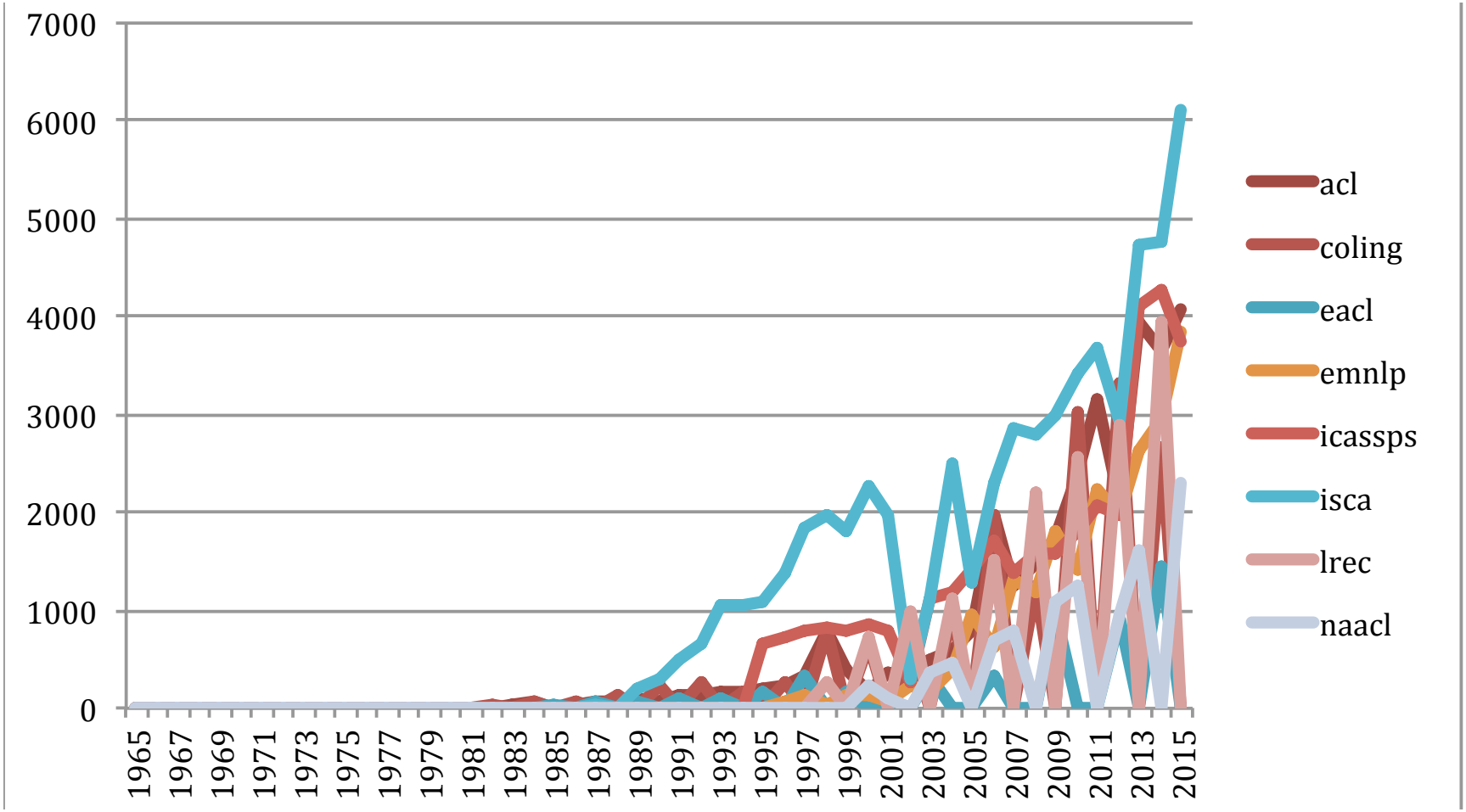
Graphe de citation (auteurs ou articles)



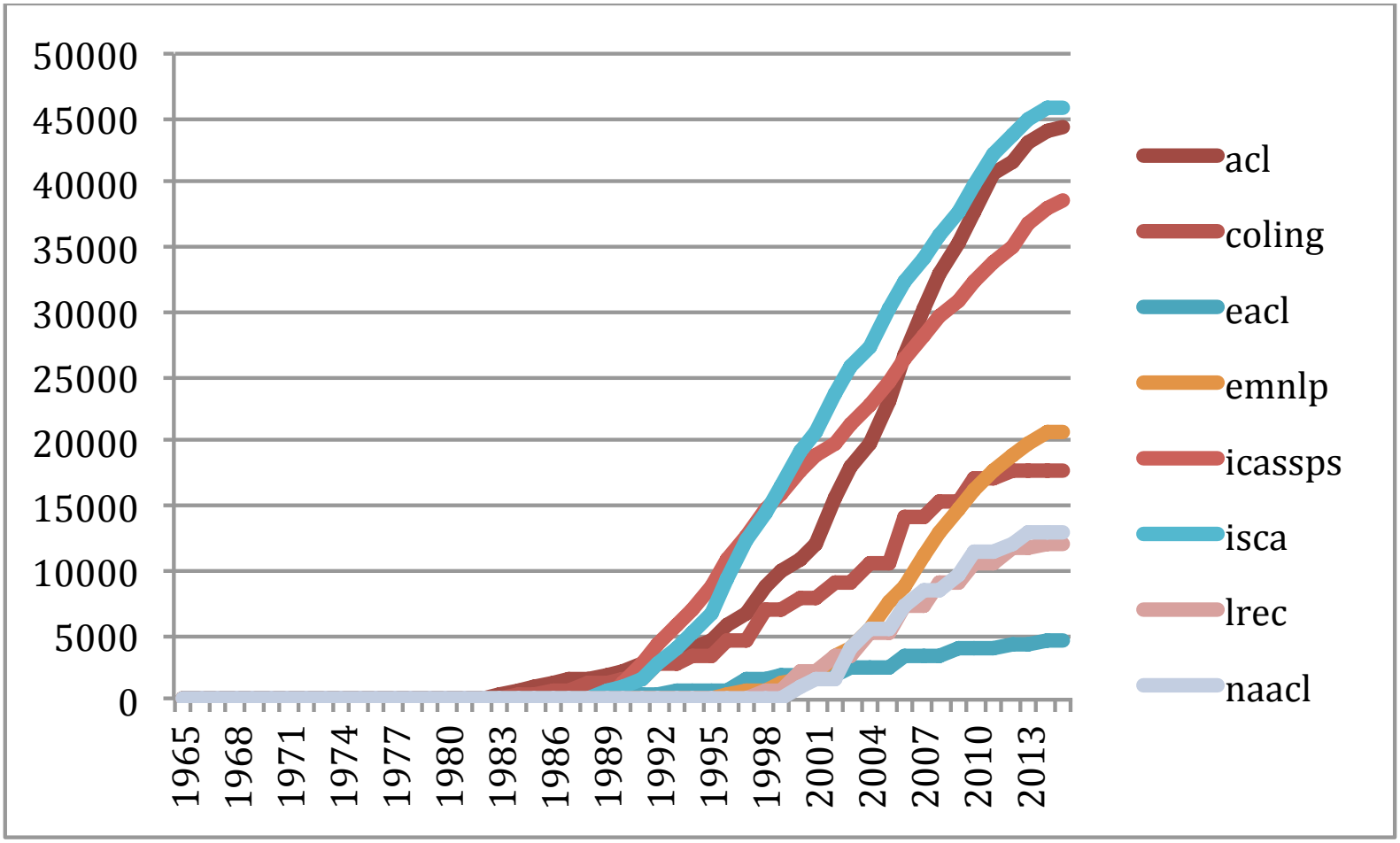
Nombre moyen de références bibliographiques par article au fil du temps



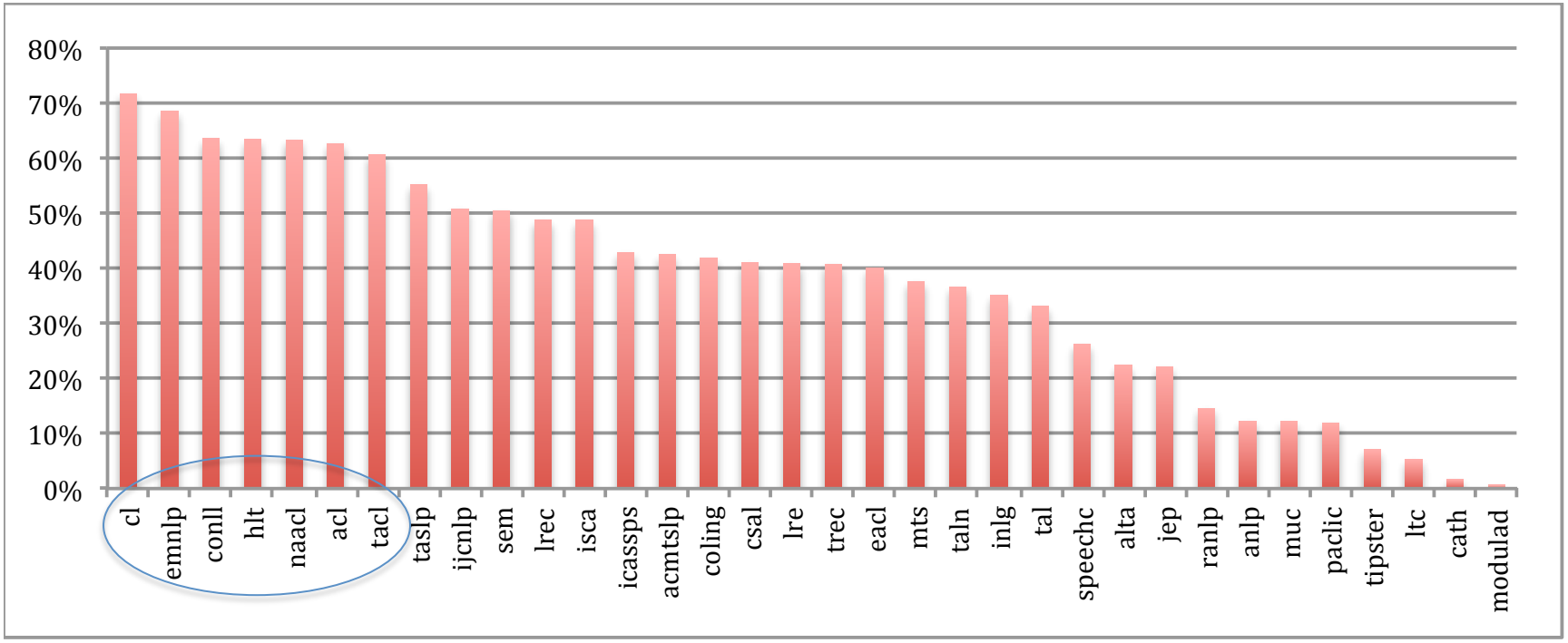
Nombre total de références bibliographiques au fil des ans (8 conférences principales)



Nombre cumulé d'articles cités au fil des ans (8 conférences principales)



Graphe de citation des auteurs : % d'auteurs dans la plus grande Composante Fortement Connexe



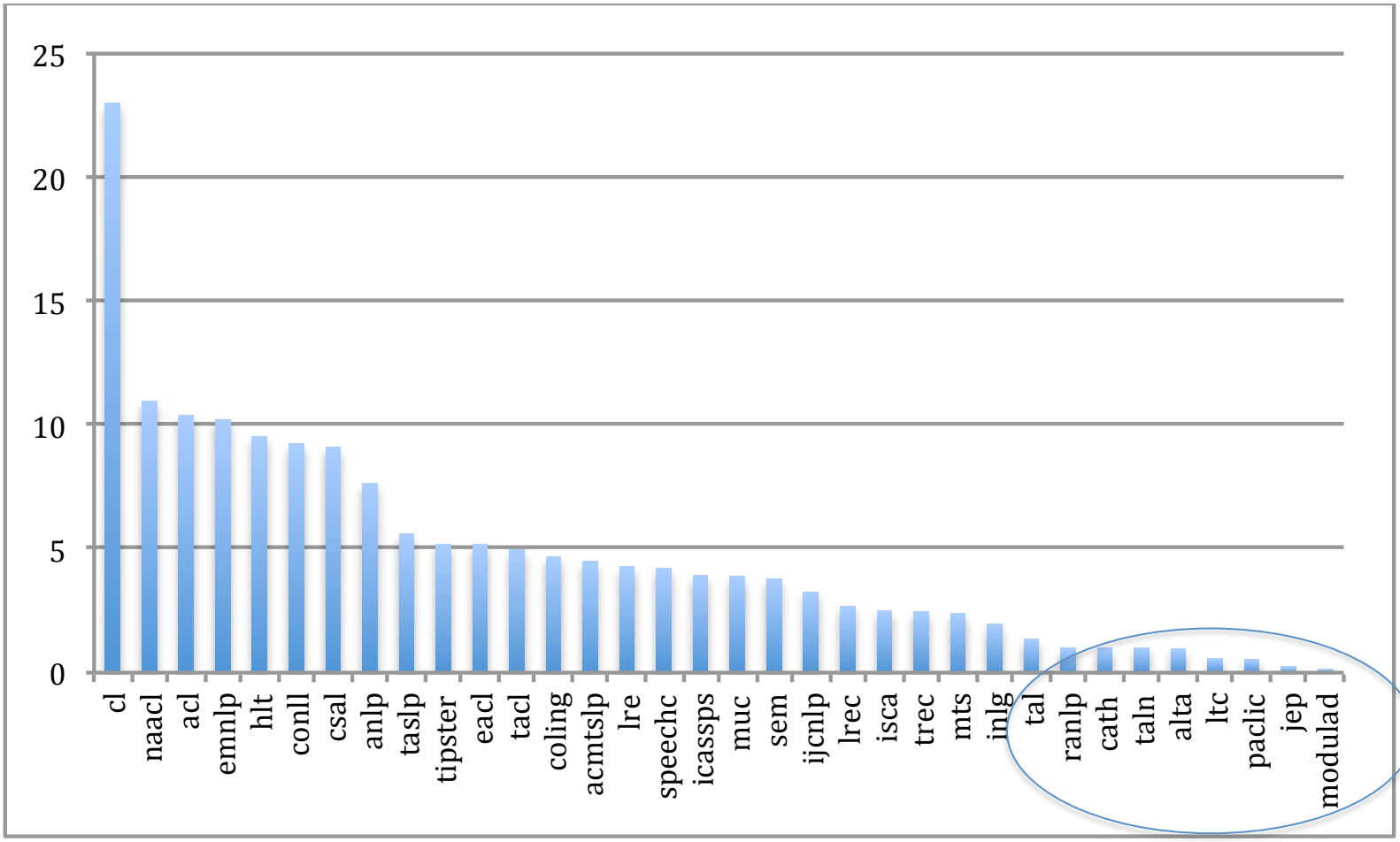
10 auteurs les plus cités

Nom	Nbre de citations	Nbre d'articles écrits par l'auteur	Ratio nbre de citations / nbre d'articles écrits par l'auteur	Pourcentage d'auto-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

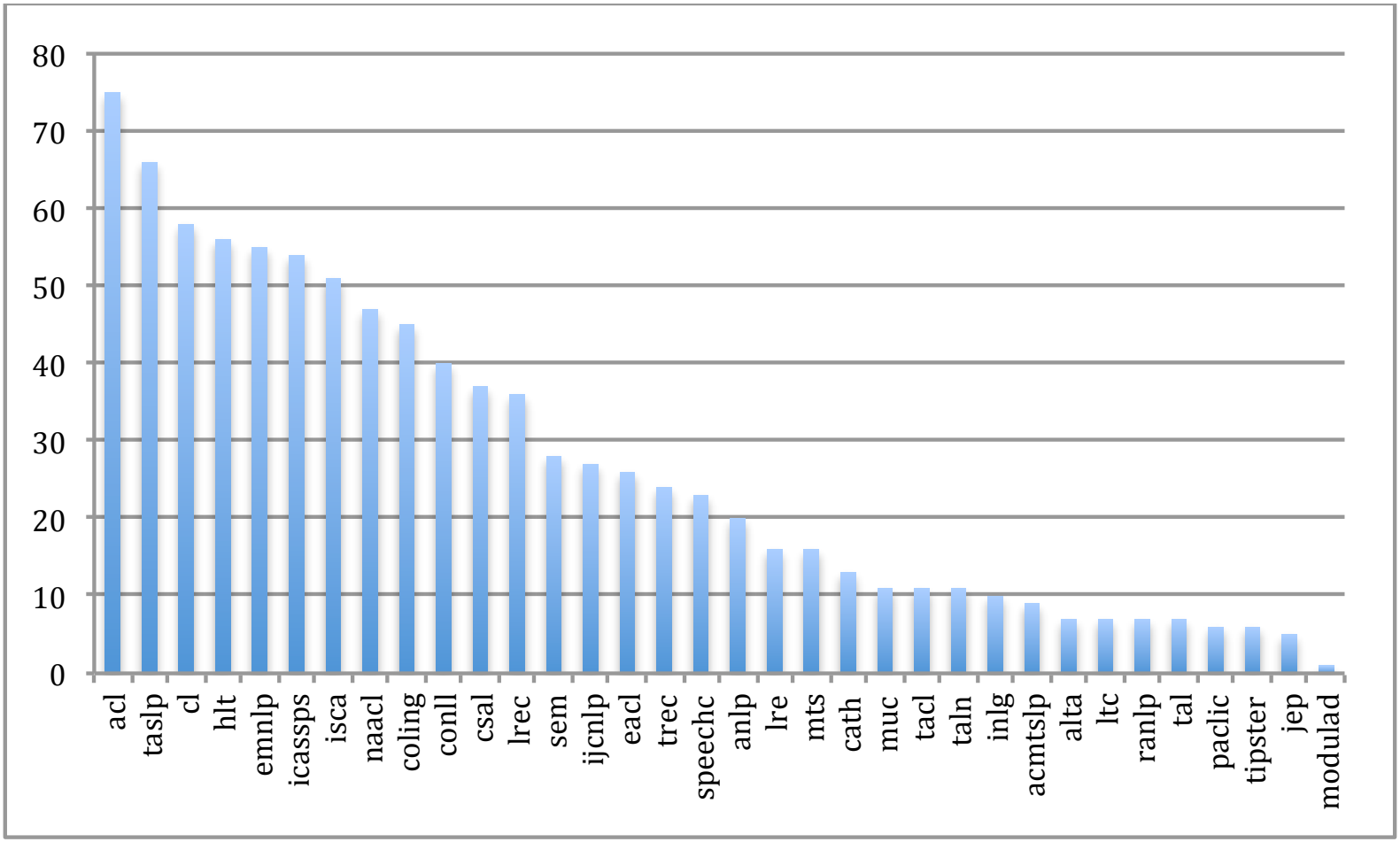
Citations

	Nombre	%
Articles jamais cités	27,183	42%
Auteurs jamais cités	19,740	40%

Degré moyen d'articles cités par publication



H-Index des 34 publications

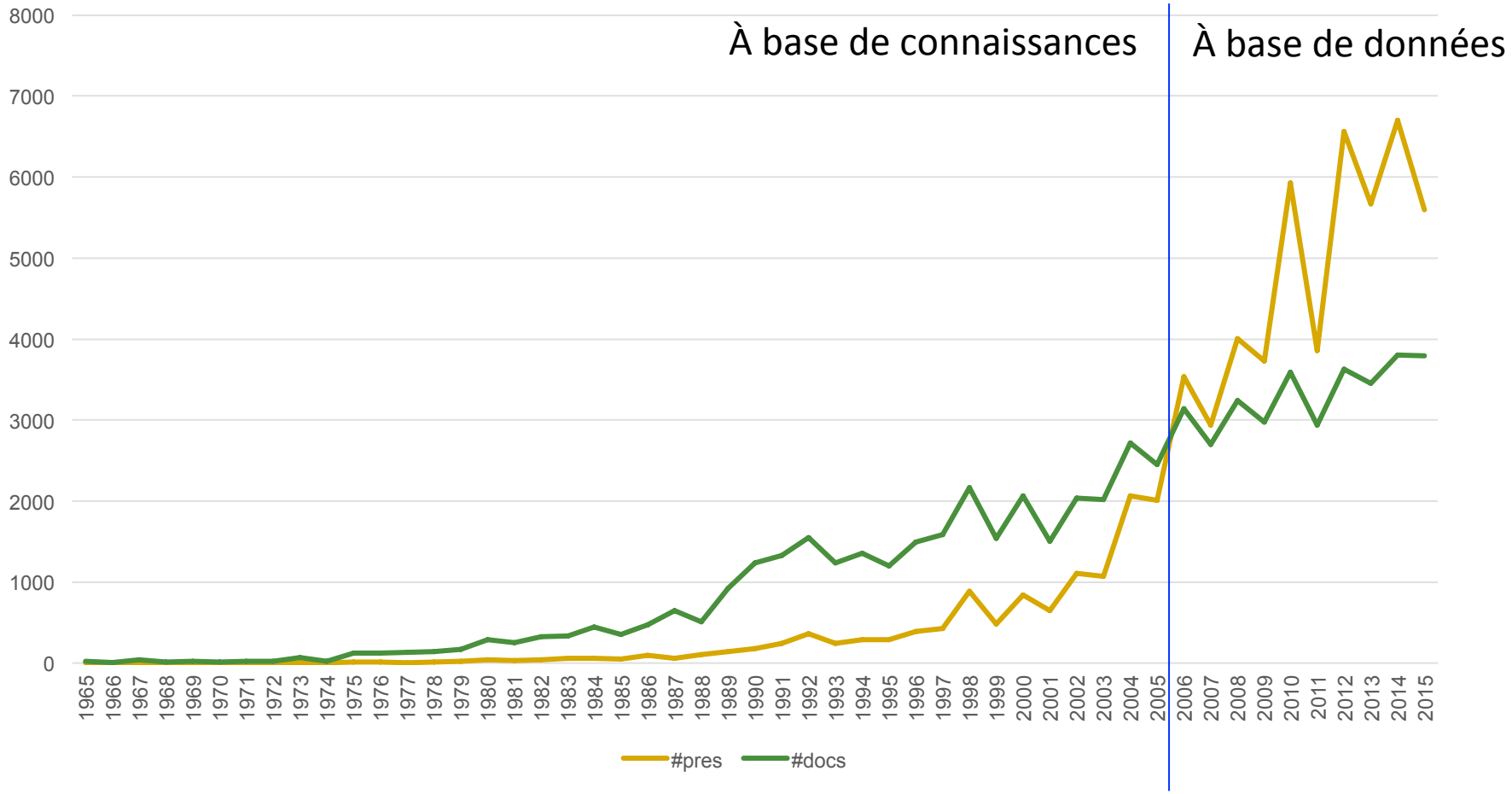


Utilisation des Ressources Linguistiques (LRE Map)

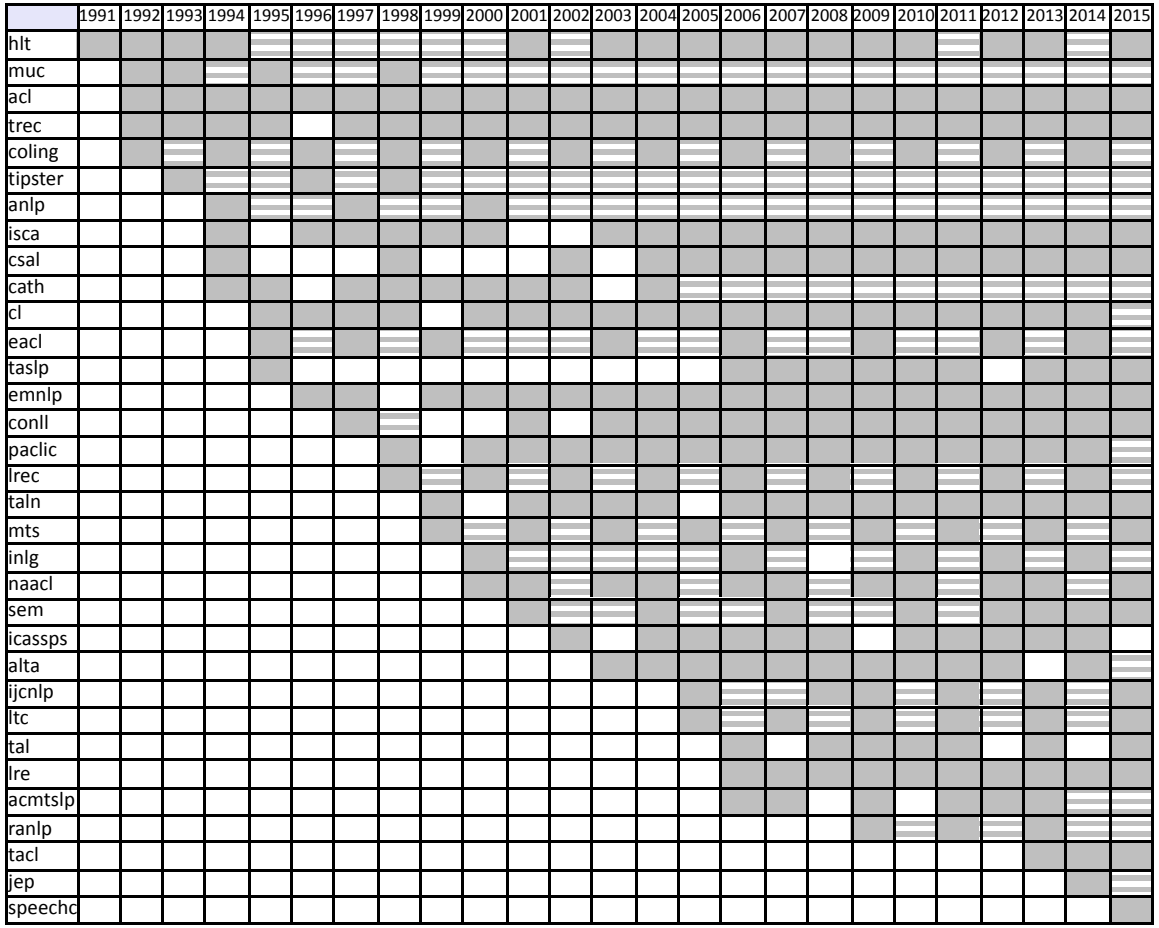
Ressources Linguistiques de la LRE-Map mentionnées dans les articles

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

Evolution de l'utilisation des Ressources Linguistiques au fil du temps (présence)



Propagation dans les publications de “Wordnet”



Facteur d'impact d'une RL (données)

Ressource	# présence
WordNet	4203
Timit	3005
Wikipedia	2824
Penn Treebank	1993
Europarl	855
FrameNet	824

Facteur d'impact d'une RL (outils)

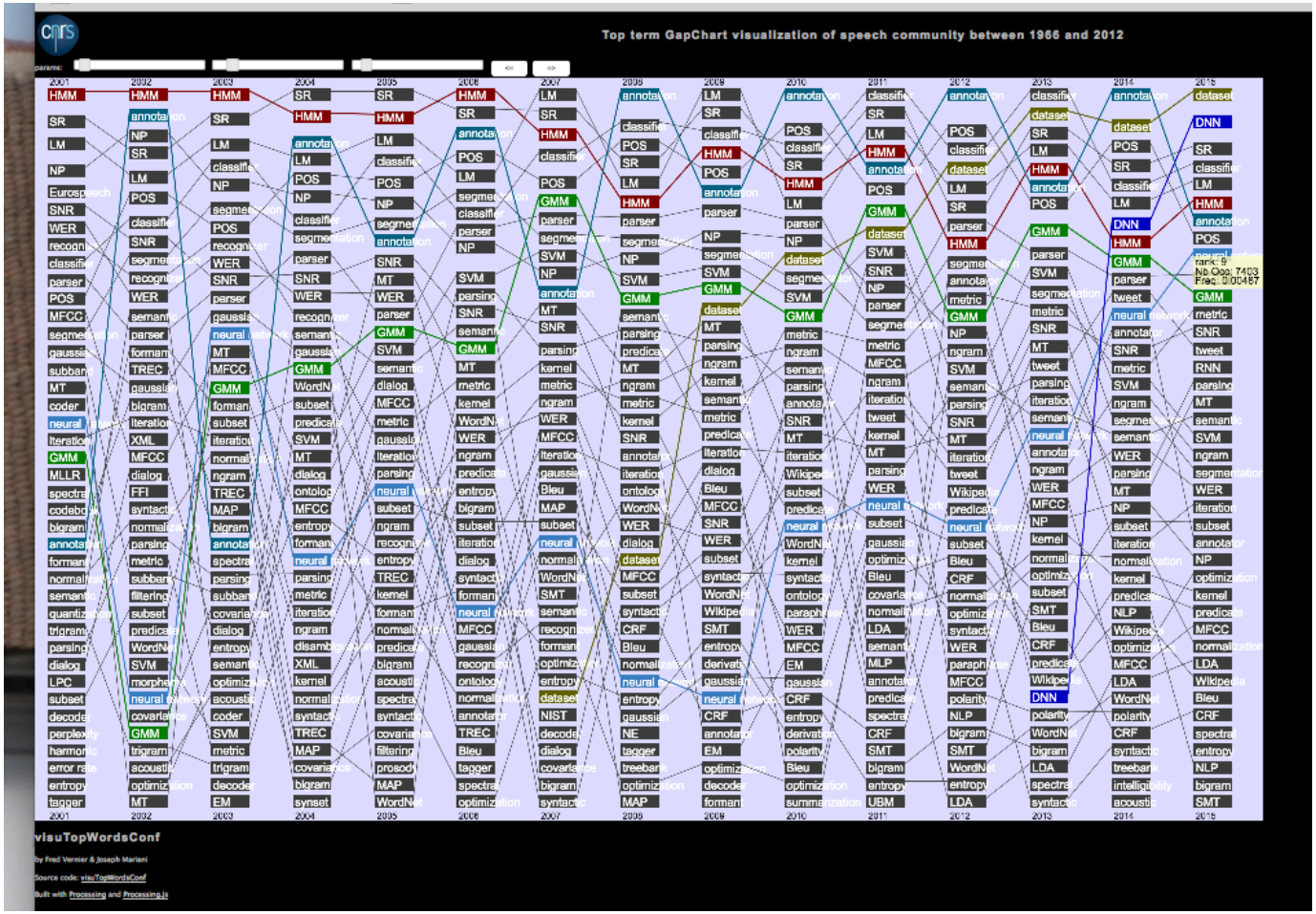
Ressource	# présence
Praat	1245
SRI Language Modeling Toolkit	1029
Weka	957
GIZA++	758

Thèmes de recherche

Termes les plus fréquents

Terme	Variantes	Nombre d'occurrences	Fréquence	Nombre de présences	Présence relative	Rang
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	134060	0.00609	14353	0.22671	1
SR	ASR, ASRs, Automatic Speech Recognition, SRs, Speech Recognition, automatic speech recognition, speech recognition	128590	0.00584	20324	0.32102	2
LM	LMs, Language Model, Language Models, language model, language models	111582	0.00507	12809	0.20232	3
annotation	annotations	111142	0.00505	11992	0.18942	4
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	101333	0.0046	13803	0.21802	5
classifier	classifiers	98092	0.00446	11513	0.18185	6
NP	NPs, noun phrase, noun phrases	94808	0.00431	9584	0.15138	7
parser	parsers	86901	0.00395	9636	0.1522	8
segmentation	segmentations	76232	0.00346	10850	0.17138	9
SNR	SNRs, Signal Noise Ratio, Signal Noise Ratios, signal noise ratio, signal noise ratios	68722	0.00312	6848	0.10817	10

Evolution des thèmes de recherche au fil du temps (ISCA-Interspeech 2001-2015)



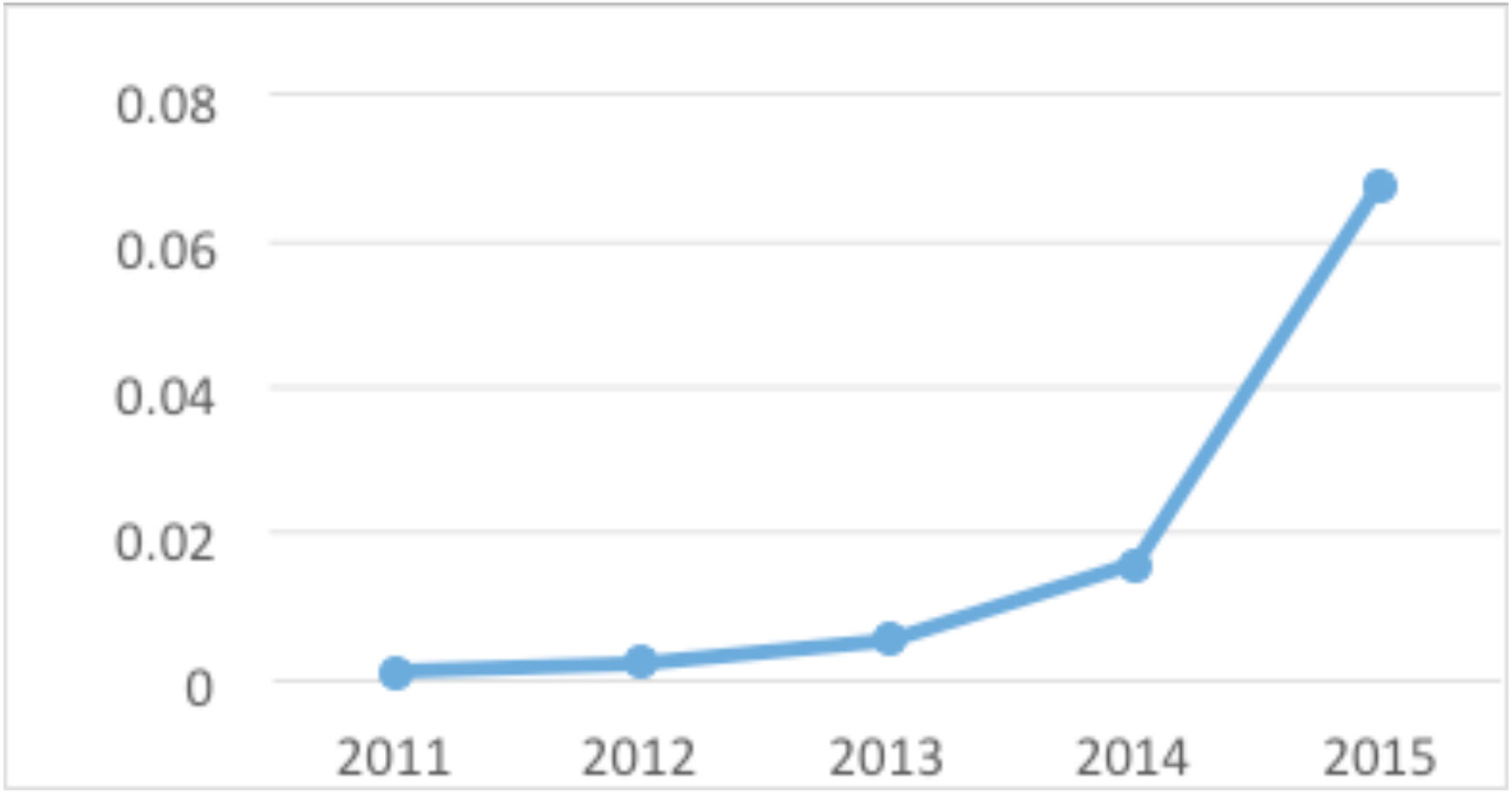
Innovation: introduction de nouveaux termes

Terme	Année où le terme est apparu	Auteur(s) ayant introduit le terme	Document(s) où le terme est apparu	Nombre d' occurrences du terme en 2015	Nombre de présences du terme en 2015 ("Impact Global du terme")
dataset	1966	Laurence Urdang	cath1966-3	14060	1474
classifier	1967	Aravind K Joshi, Danuta Hiz	C67-1007	8202	999
linear	1967	Marian W Cobin	cath1967-5	2065	918
optimization	1967	Ellis B Page	C67-1032	3317	901
normalization	1967	Bruce A Beatie	cath1967-16	2974	774
HMM	1982	Cory S Myers, Stephen E Levinson	taslp1982-86	7575	688
acoustic	1967	David Shillan	C67-1018	1906	646
spectral	1971	Arne Zettersten	cath1971-12	2486	608
filtering	1973	Eugenio Morreale, Massimo Mennucci	C73-2024	1719	604
toolkit	1980	C Raymond Perrault	J80-3003	1155	603

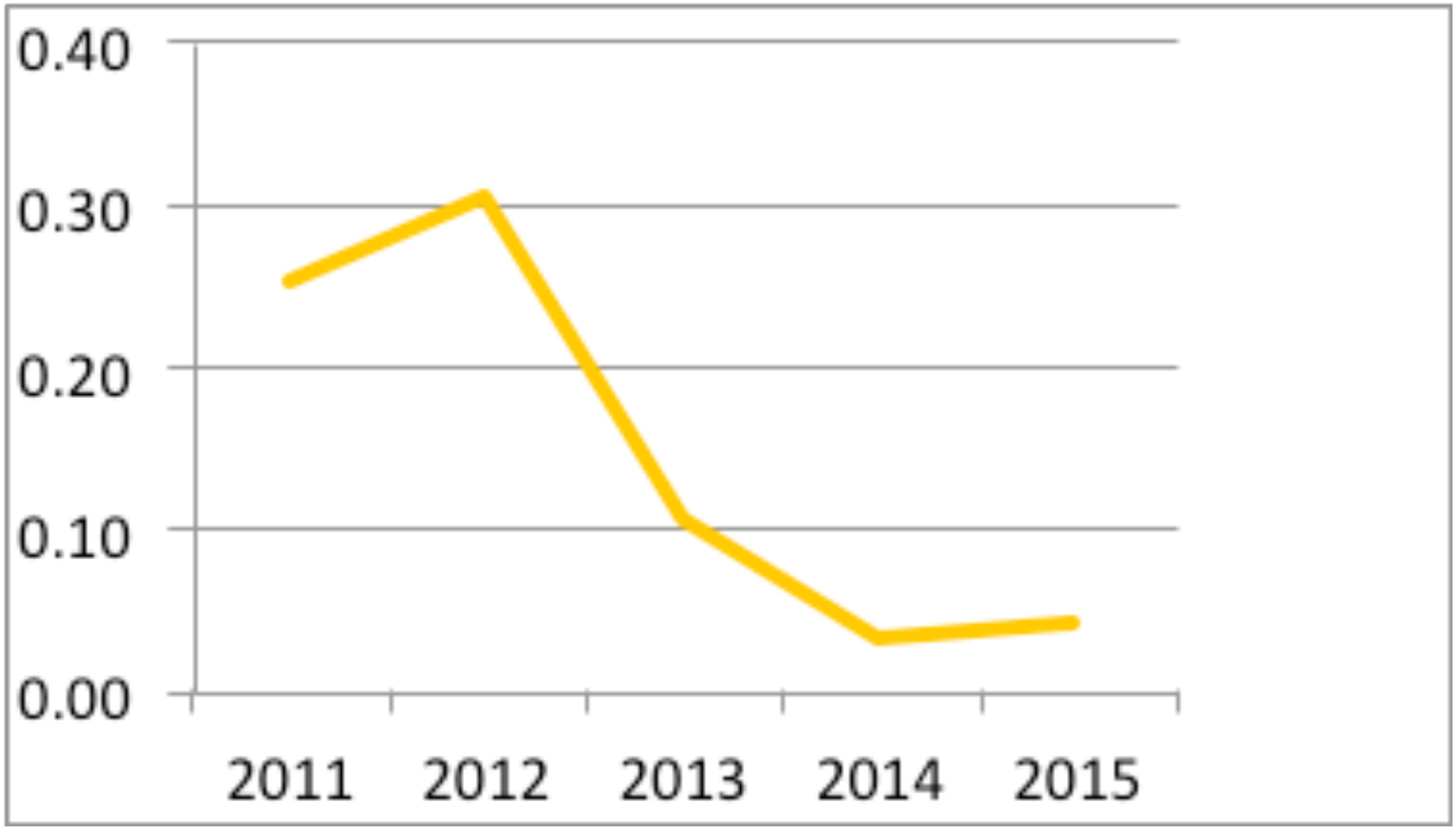
Prédiction de thèmes de recherche (Weka)

Observé pour 2013	Observé pour 2014	Prédit pour 2015	Observé pour 2015	Rang
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

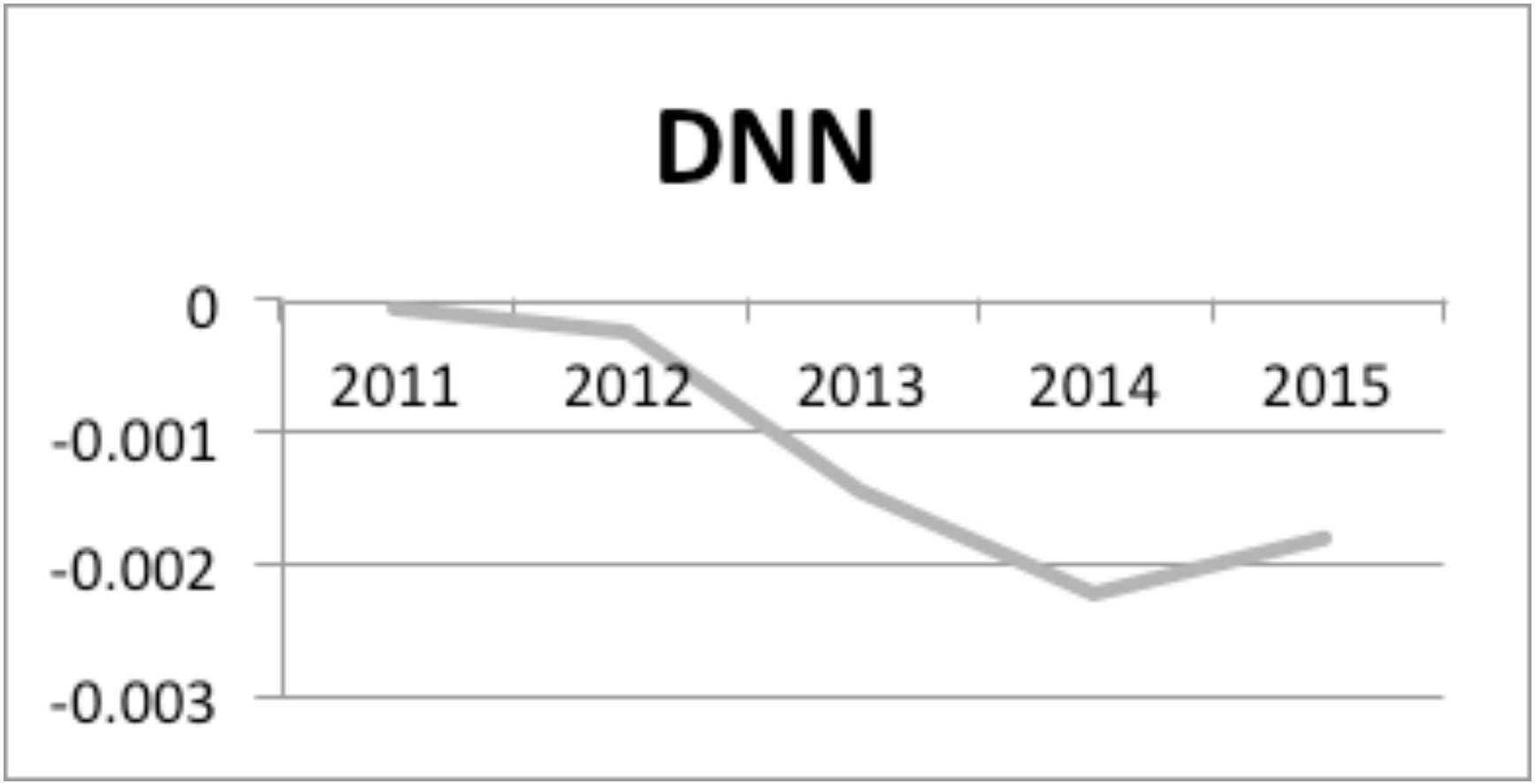
Fiabilité des prédictions : erreur de prédiction à partir de 2010



« Surprises » : Ruptures épistémologiques



Emergence de thème de recherche



Prédiction pour les 5 années à venir

Observé 2014	Observé 2015	Prédiction pour 2016	Prédiction pour 2017	Prédiction pour 2018	Prédiction pour 2019	Prédiction pour 2020	Rang
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

Réutilisation et Plagiat

(Auto-)réutilisation et plagiat

>4% de ressemblance	La source est citée	La source n'est pas citée
Au moins un auteur en commun	Auto-réutilisation	Auto-plagiat
Aucun auteur en commun	Réutilisation	Plagiat

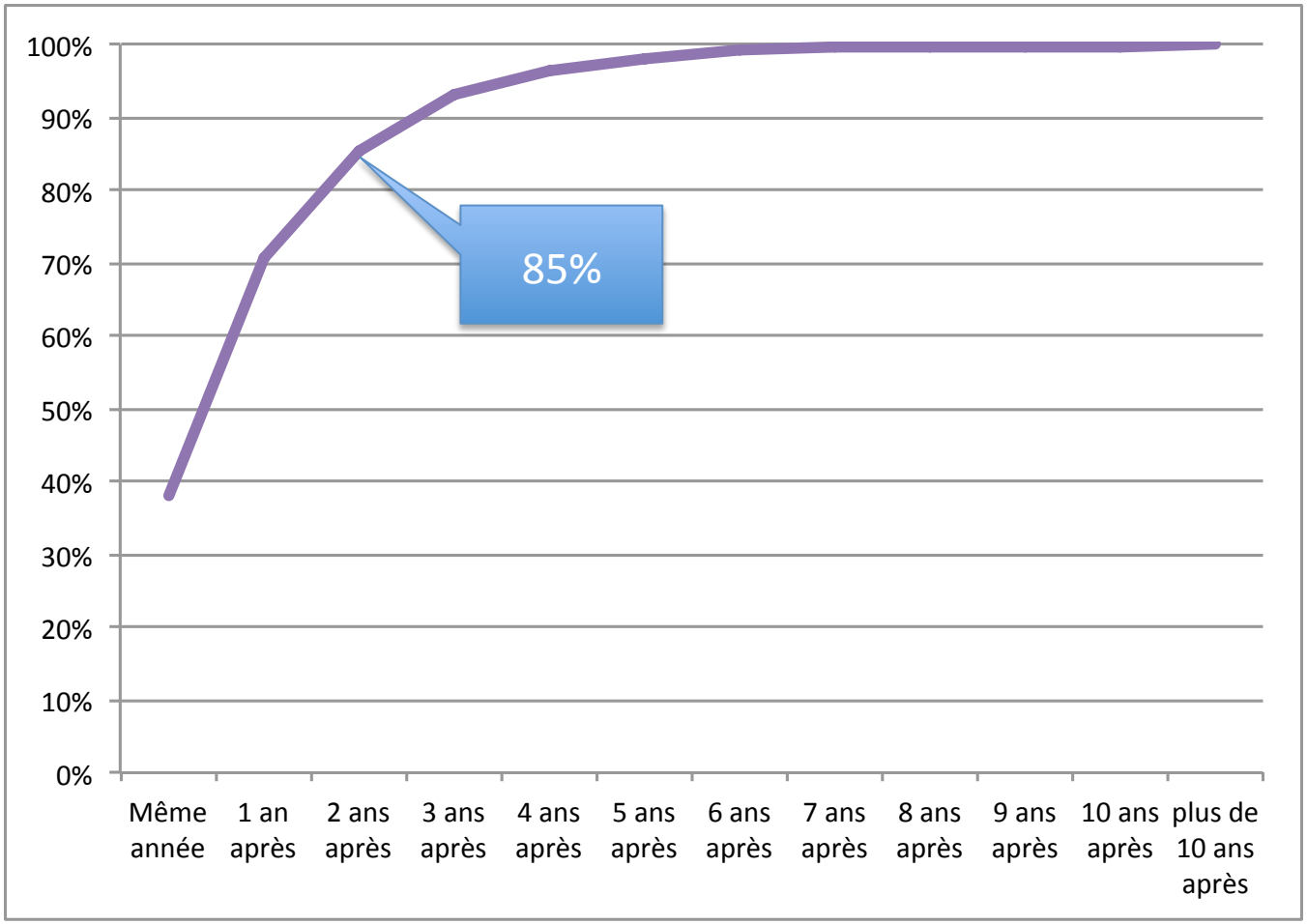
Réutilisation et Plagiat (0,3%)

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

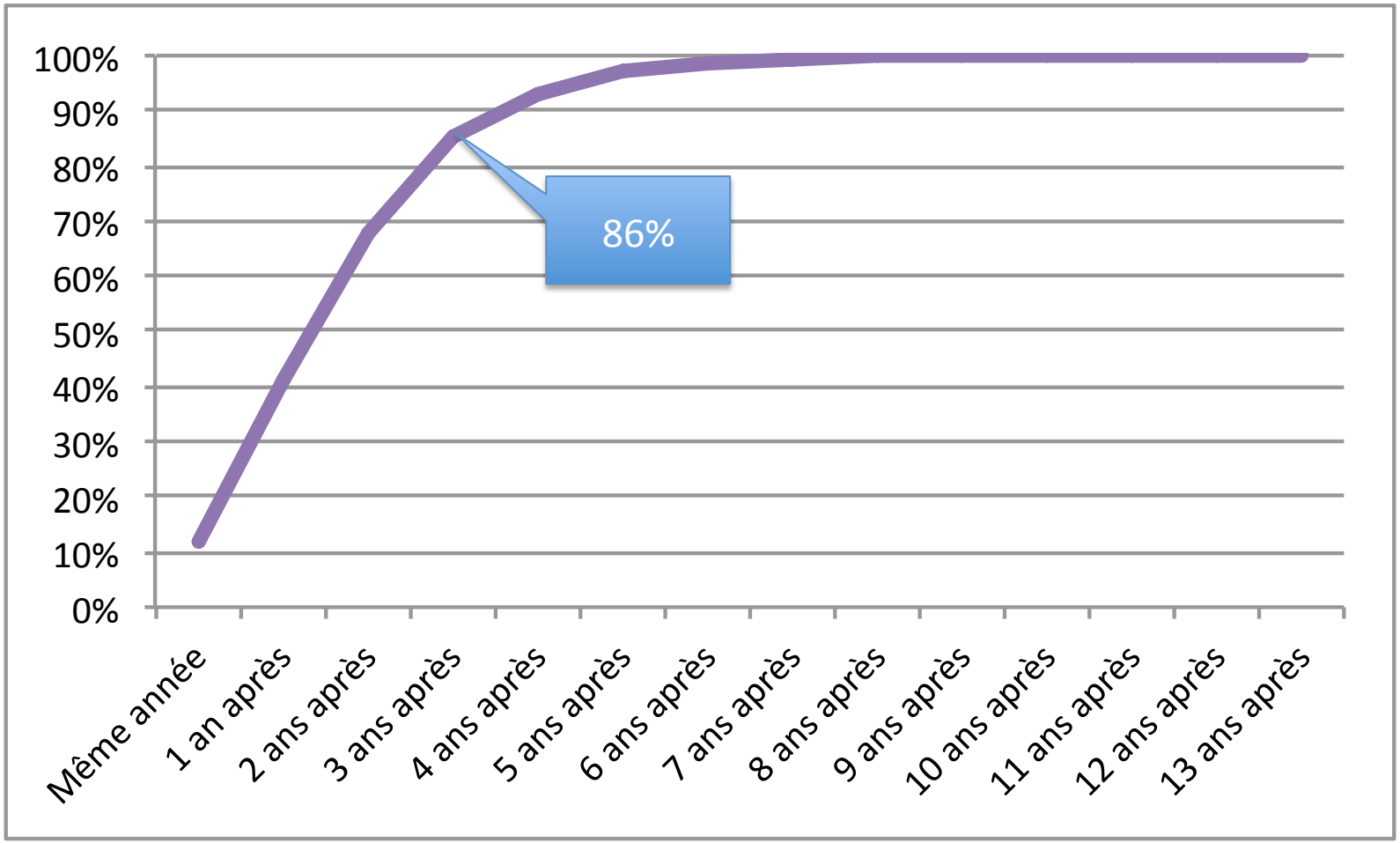
Auto-réutilisation et Auto-plagiat (20%)

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

Délai entre publication et réutilisation



Délai entre publication dans conférence et réutilisation dans revue

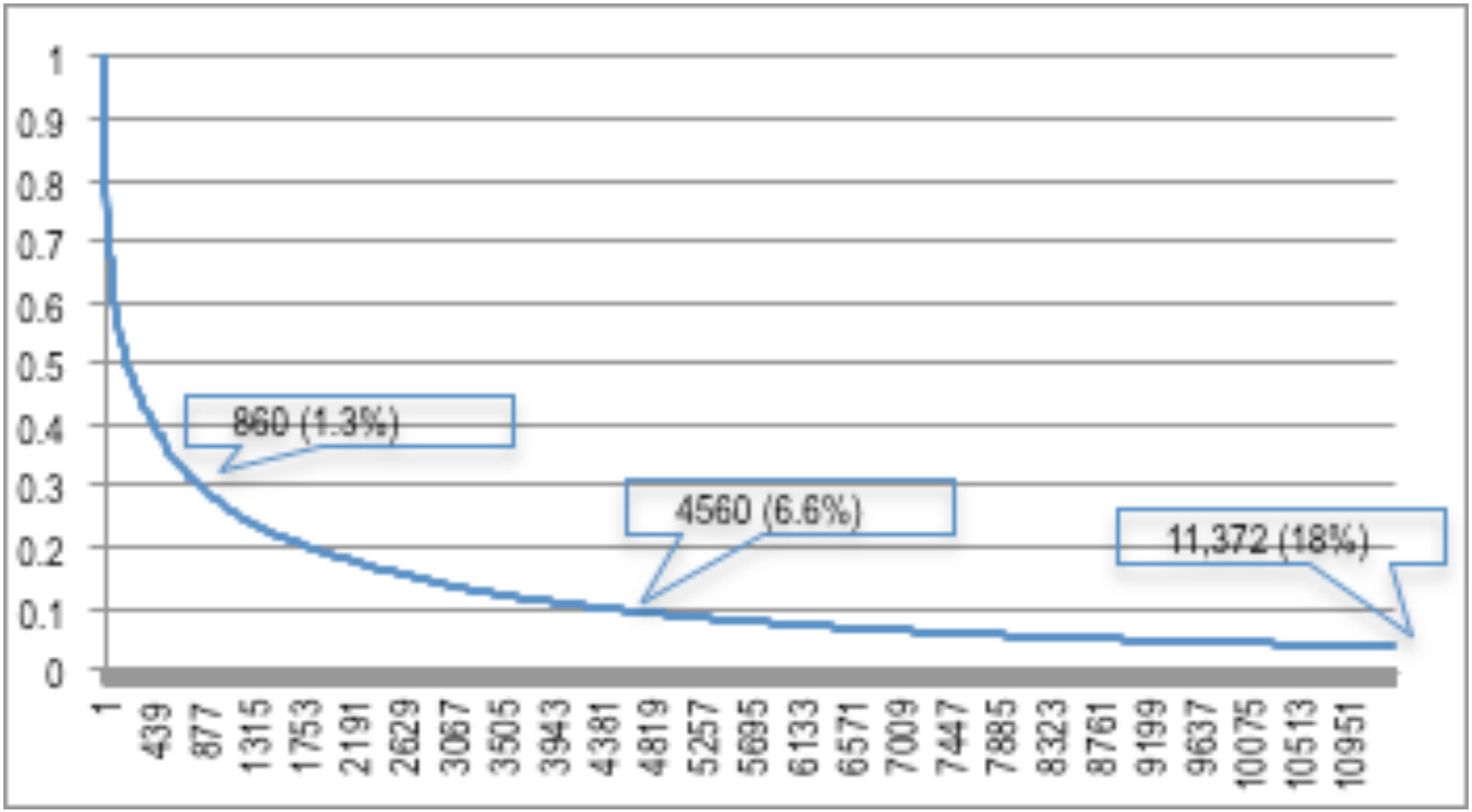


Conclusions

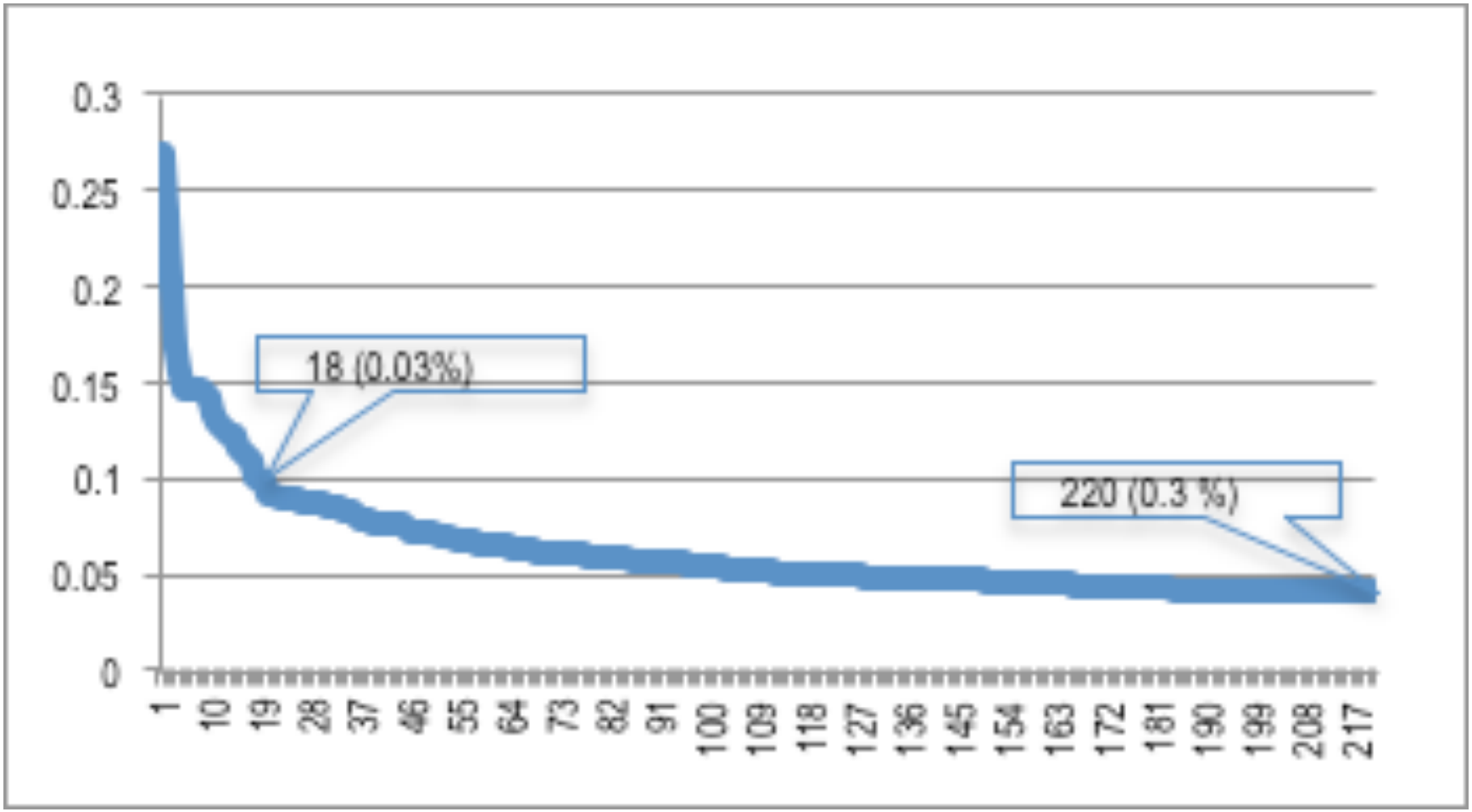
- Vaste analyse des données bibliographiques dans un domaine scientifique particulier (TAL)
- Problème de l'identification des informations :
  - Noms d'auteurs,
  - Noms de laboratoires,
  - Noms de revues, de conférences,
  - Noms de ressources linguistiques…
- Nécessite un nettoyage manuel fastidieux
- Besoin d'une action de coordination internationale pour affecter des identifiants uniques et persistants à ces informations

Mais c’est aussi beaucoup de fun !

Auto-réutilisation/plagiat



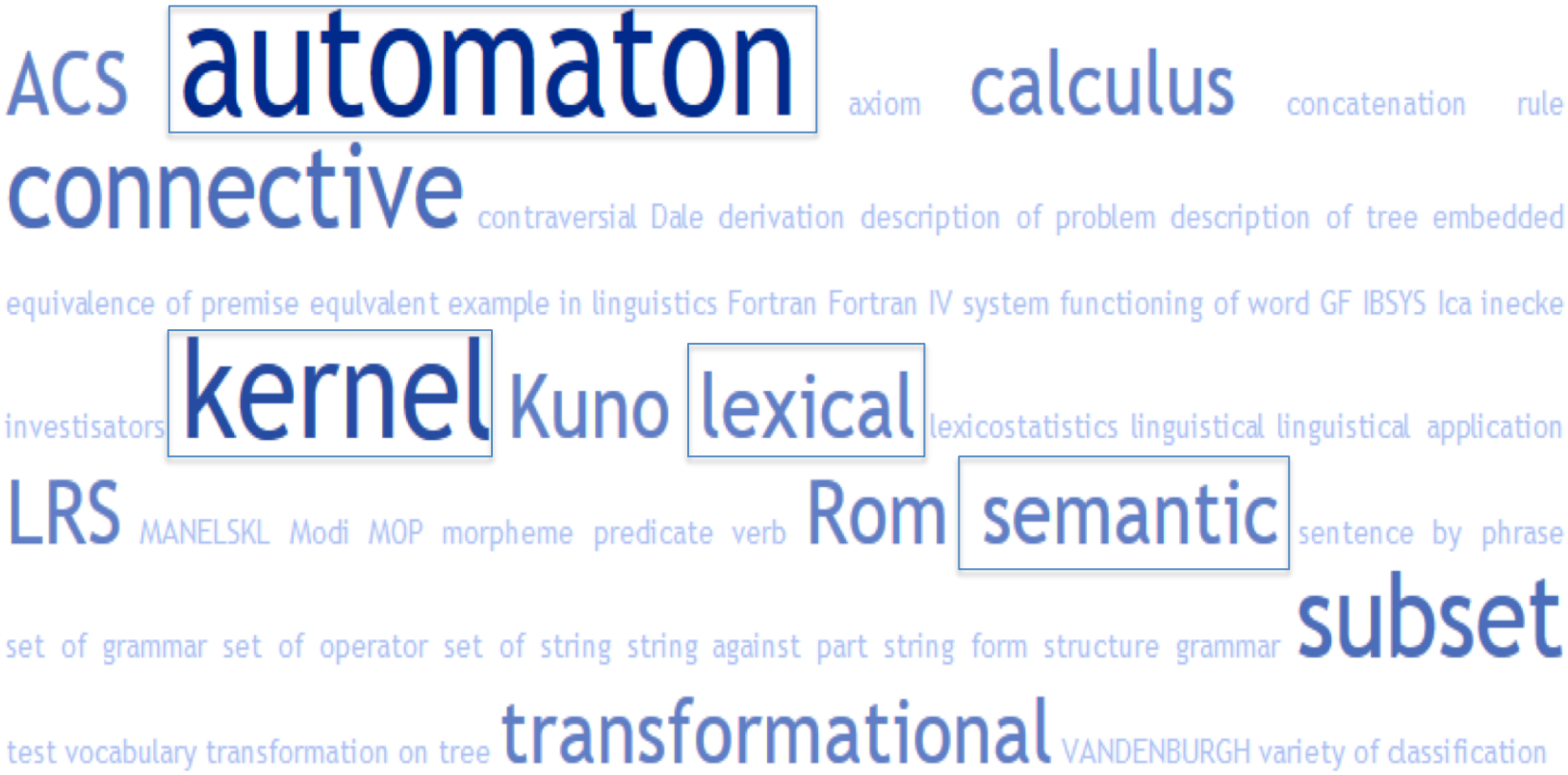
Réutilisation/Plagiat



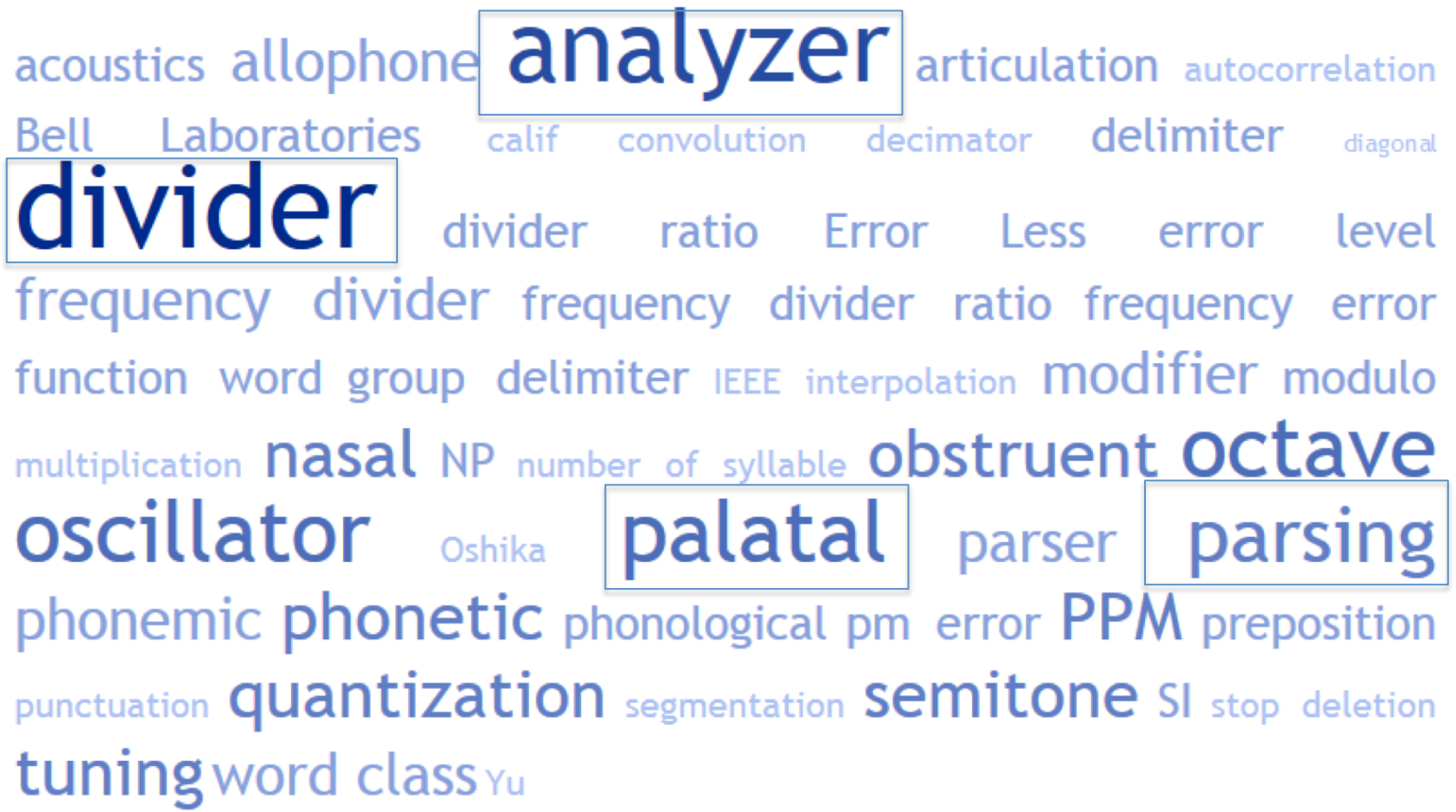
H-Index Google Scholar en linguistique computationnelle (2011-2015)

Rang	Publication	h5-index	h5-median
1	Meeting of the Association for Computational Linguistics (ACL)	70	112
2	Conference on Empirical Methods in Natural Language Processing (EMNLP)	55	105
3	arXiv Computation and Language (cs.CL)	52	86
4	Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL)	35	57
5	International Conference on Language Resources and Evaluation (LREC)	35	53
6	Computational Linguistics	30	60
7	Computer Speech & Language	29	48
8	Conference of the European Chapter of the Association for Computational Linguistics (EACL)	26	45
9	International Conference on Computational Linguistics (COLING)	26	32
10	Language Resources and Evaluation	24	39
11	International Joint Conference on Natural Language Processing (IJCNLP)	23	35
12	Conference on Computational Natural Language Learning (CoNLL)	22	43
13	Transactions of the Association for Computational Linguistics	21	43
14	Workshop on Statistical Machine Translation	21	27
15	IEEE Spoken Language Technology Workshop (SLT)	19	34
16	International Conference on Computational Linguistics and Intelligent Text Processing	19	26
17	IEEE International Conference on Semantic Computing	15	23
18	Recent Advances in Natural Language Processing (RANLP)	15	22
19	Text Analysis Conference	14	24
20	Natural Language Engineering	14	20

Nuage de mots-clefs (1965)



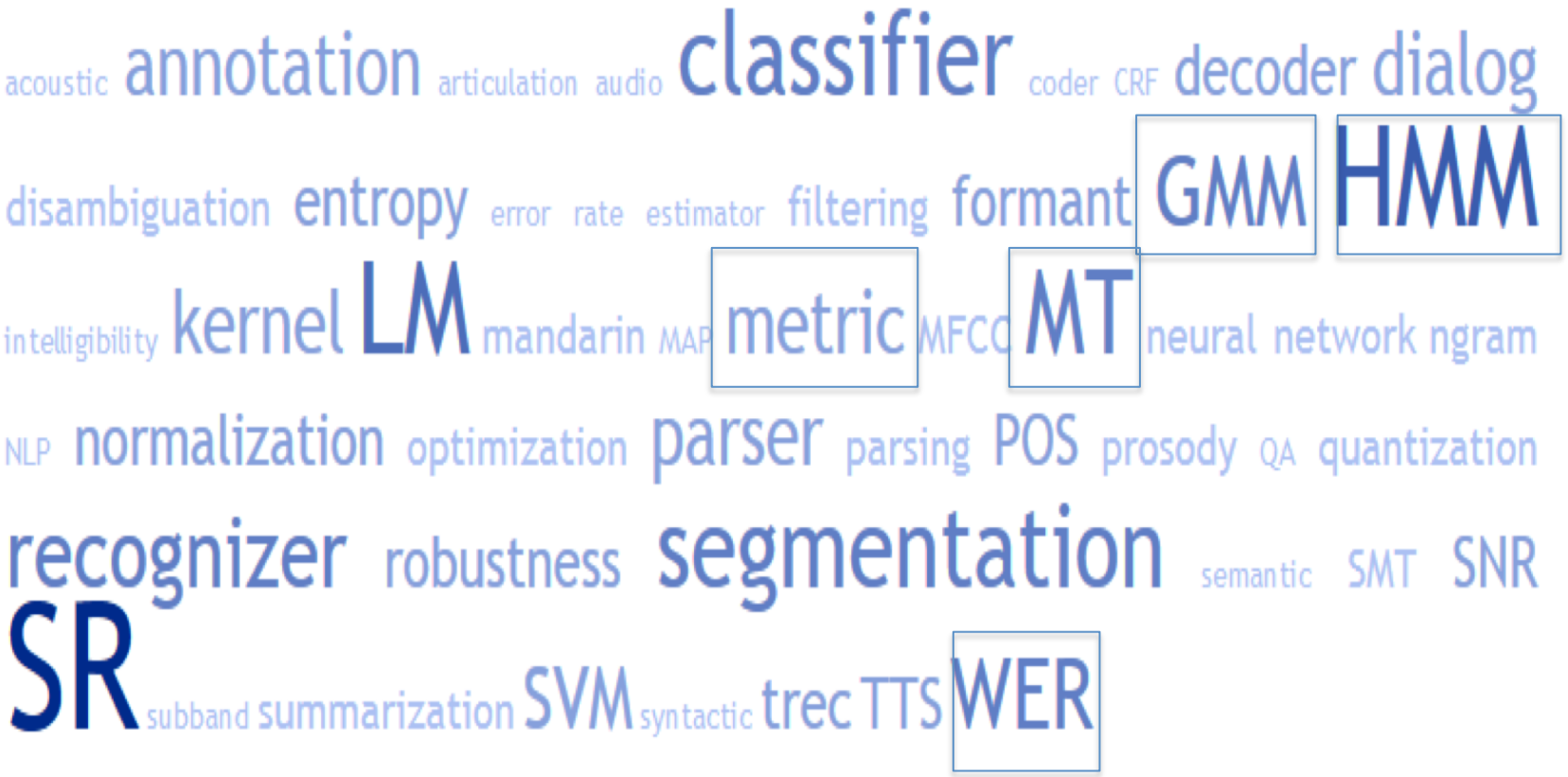
Nuage de mots-clefs (1975)



Nuage de mots-clefs (1985)

Nuage de mots-clefs (1995)

Nuage de mots-clefs (2005)



Nuage de mots-clefs (2015)