Le corpus NLP4NLP pour l'analyse bibliométrique de 50 années de recherches en traitement automatique de la parole et du langage naturel

Joseph Mariani, Gil Francopoulo, Patrick Paroubek

Dans Document numérique 2017/2 (Vol. 20), pages 31 à 78
Éditions Lavoisier

ISSN 1279-5127
ISBN 9782746248533
DOI 10.3166/dn.2017.00012

Le corpus NLP4NLP pour l'analyse bibliométrique de 50 années de recherches en traitement automatique de la parole et du langage naturel

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

1. *CNRS, LIMSI, université Paris-Saclay, faculté des sciences, rue John von Neumann, Bâtiment 508, 91405 Orsay cedex, France {joseph.mariani, patrick.paroubek}@limsi.fr*

2. *Tagmatica, 126, rue de Picpus, 75012 Paris, France gil.francopoulo@wanadoo.fr*

RÉSUMÉ. *Nous avons constitué le corpus NLP4NLP pour étudier le contenu des publications scientifiques dans le domaine du traitement automatique de la parole et du langage naturel. Il contient les articles publiés dans 34 conférences et revues principales du domaine, sur une période de 50 ans (1965-2015), comprenant 65 000 documents, rassemblant 50 000 auteurs, incluant 325 000 références et représentant environ 270 millions de mots. Nous avons conduit différentes études sur ces données : évolution au fil du temps du nombre d'articles et d'auteurs, collaborations entre auteurs, citations entre papiers et entre auteurs, évolution des thèmes de recherche et identification des auteurs qui les ont introduits, détection des innovations et des ruptures épistémologiques, utilisation des ressources linguistiques, réutilisation des articles et plagiat, tout ceci dans le cadre d'une analyse globale ou comparative entre sources.*

ABSTRACT. *We have created the NLP4NLP corpus to study the content of scientific publications in speech and natural language processing. It contains articles published in 34 major conferences and journals in this field over a period of 50 years (1965-2015), comprising 65,000 documents, gathering 50,000 authors, including 325,000 references and representing approximately 270 million words. We have conducted various studies on this data: evolution over time of the number of articles and authors, collaborations between authors, citations between papers and authors, evolution of research topics and identification of the authors who introduced them, detection of innovations and epistemological ruptures, use of language resources, reuse of articles and plagiarism, all this in the context of a global or comparative analysis between sources.*

MOTS-CLÉS : *traitement de la parole, traitement du langage naturel, analyse de textes, bibliométrie, scientométrie.*

KEYWORDS: *speech processing, natural language processing, text analytics, bibliometrics, scientometrics.*

DOI: 10.3166/dn.2017.00012

1. Introduction

1.1. Objectif

Notre objectif est d'utiliser les outils du Traitement Automatique du Langage Naturel (TAL) pour analyser la bibliographie en Traitement Automatique du Langage Naturel. Nous avons initialisé de tels travaux dès 1991 avec la publication d'une étude de la conférence IEEE ICASSP sur une période de 15 ans (1976-1990) (Mariani, 1990). Cette étude a été utile dans le lancement de la conférence Eurospeech, à présent Interspeech (Mariani, 2013). L'ACL a constitué une Anthologie (Radev *et al*., 2013) et organisé le Workshop *Rediscovering 50 Years of Discoveries in Natural Language Processing*[1] (ACL, 2012) à l'occasion de son 50e anniversaire, lors de la conférence ACL à Jeju (Corée) en 2012. Nous avons été invités à faire une intervention plénière intitulée *Rediscovering 25 Years of Discoveries in Spoken Language Processing*, à l'occasion du 25e anniversaire de l'ISCA, lors de la conférence Interspeech à Lyon, en 2013 (Mariani *et al*., 2013), à partir des archives d'ISCA assemblées par Wolfgang Hess[2]. Puis une nouvelle intitulée *Rediscovering 15 Years of Discoveries in Language Resources and Evaluation, pour le* 15e anniversaire de la conférence LREC, en 2014 à Reykjavik (Mariani *et al*., 2014a). Cela s'est ensuite traduit par un article *Rediscovering 15 + 2 Years of Discoveries in LRE*, publié dans la revue *Language Resources and Evaluation* en Mars 2016 (Mariani *et al*., 2016). Et enfin une dernière présentation invitée *Rediscovering 10 to 20 Years of Discoveries in Language and Technology*, pour le 20e anniversaire de la conférence L&TC, à Poznan en 2015 (Mariani *et al*., 2015). Notre but a été ensuite d'étendre ces études à un demi-siècle de recherches en Traitement Automatique du Langage. Le présent article effectue un survol des principaux résultats obtenus, sans pouvoir prétendre ni à l'exhaustivité, ni à la description détaillée des données et des méthodes utilisées qui ont été ou seront plus finement explicitées dans d'autres publications.

1.2. Un sujet actuel

L'application des méthodes d'analyse de textes aux articles scientifiques est un sujet très actuel (voir par exemple Li *et al*. (2006), Tang *et al*. (2008), Dunne *et al*. (2012), Osborne *et al*. (2013), Ding *et al*. (2014), Gollapalli et Li (2015), Jha *et al*. (2016)), le Stanford Large Network Dataset Collection (SNAP)[3] ou le projet Saffron[4]. Notre contribution nous semble cependant originale en ce qu'elle traite spécifiquement le domaine du traitement du langage, et qu'elle y inclut à la fois le langage écrit et la parole en rassemblant une grande masse de publications du domaine. Outre les analyses que nous avons été invités à présenter dans les conférences précitées, nous avons publié une

1. http://aclweb.org/anthology/.
2. http://www.isca-speech.org/iscaweb/index.php/archive/online-archive.
3. http://snap.stanford.edu/data/.
4. http://saffron.deri.ie.

partie de nos travaux dans des conférences sur l'infométrie, la scientométrie ou la bibliométrie : au Workshop on Mining Scientific Publications[5] (WOSP'2015) à Fort Knox (USA), les 24 et 25 juin 2015, qui a donné lieu à un numéro spécial du D-Lib Magazine (Nov./Dec. 2015, vol. 21, n° 11/12) (G. Francopoulo *et al*., 2015c), et, à peu près au même moment, au Workshop on Computational Linguistics and Bibliometrics (CLBib)[6], organisé en marge de la 15e Intal Society of Scientometrics and Informetrics Conference (ISSI) à Istanbul (Turquie), le 29 juin 2015. Plus récemment, nous avons participé à BIRNDL : Joint Workshop on Bibliometric-enhanced IR (BIR) and NLP for digital libraries (NLPIR4DL)[7], organisé dans le cadre de l'ACM/IEEE Joint Conference on Digital Libraries 2016 à Newark (USA) le 23 juin 2016, qui a donné lieu à un numéro spécial de l'International Journal on Digital Libraries en mars 2017 (Mariani *et al*., 2017).

2. Le corpus NLP4NLP

Nous utilisons donc les outils du Traitement Automatique du Langage Naturel (TAL) pour analyser la bibliographie en Traitement Automatique du Langage Naturel, d'où le nom que nous avons donné à notre corpus : NLP4NLP (Francopoulo *et al*., 2015a, 2015b). Nous entendons ici le traitement du langage naturel dans un sens large qui inclut le traitement de la langue écrite, de la langue parlée, de la langue des signes et l'extraction d'Information.

Le corpus contient 34 publications sur une période de 50 années (1965-2015), dont les principales conférences (ACL, IEEE-ICASSP (uniquement la partie consacrée au traitement automatique de la parole), ISCA-Interspeech, ELRA-LREC…) et revues (IEEE-TASLP, *Computational Linguistics*, *Speech Communication*, *Computer Speech and Language*, *Language Resources and Evaluation*…) du domaine. Cela représente 558 « évènements » (par « évènement », nous entendons la tenue d'une conférence, qui peut être annuelle ou à fréquence variable, ou bien la publication d'un numéro d'une revue, coïncidant souvent avec une année calendaire). Cela regroupe 65,003 articles écrits par 48,894 auteurs différents, rassemblant environ 270 millions de mots, et contenant 324,422 références bibliographiques.

Le tableau 1 donne la liste des éléments du corpus, avec le nom et l'acronyme éventuel de la publication, conférence ou revue, le nombre de documents qu'elle contient, la langue (la plupart sont en anglais, mais quelques unes sont en français, certains articles pouvant également être en allemand ou en russe), le mode d'accès, libre ou propriétaire (dans ce cas, nous avons obtenu la permission d'utiliser les données pour cette étude de la part de la maison d'édition), la période couverte et le nombre d'évènements. Pour obtenir la somme des documents et des évènements contenus dans

5. https://wosp.core.ac.uk/jcdl2015/.
6. http://ceur-ws.org/Vol-1384/ et https://arxiv.org/abs/1506.05402.
7. http://wing.comp.nus.edu.sg/birndljcdl2016/.

Tableau 1. Le corpus NLP4NLP des Conférences (24) et Revues (10)[a]

Nom court	# docs	Type	Nom long	Langue	Accès contenu	Période	# évènements
acl	4 264	Conférence	Association for Computational Linguistics Conference	Anglais	Libre accès[b]	1979-2015	37
acmtslp	82	Revue	ACM Transactions on Speech and Language Processing	Anglais	Propriétaire	2004-2013	10
alta	262	Conférence	Australasian Language Technology Association	Anglais	Libre accès[b]	2003-2014	12
anlp	278	Conférence	Applied Natural Language Processing	Anglais	Libre accès[b]	1983-2000	6
cath	932	Revue	Computers and the Humanities	Anglais	Propriétaire	1966-2004	39
cl	776	Revue	Computational Linguistics	Anglais	Libre accès[b]	1980-2014	35
coling	3 813	Conférence	Conference on Computational Linguistics	Anglais	Libre accès[b]	1965-2014	21
conll	842	Conférence	Computational Natural Language Learning	Anglais	Libre accès[b]	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[b]	1983-2014	14
emnlp	2 020	Conférence	Empirical methods in natural language processing	Anglais	Libre accès[b]	1996-2015	20

Tableau 1 – (suite)

Nom court	# docs	Type	Nom long	Langue	Accès contenu	Période	# évènements
hlt	2 219	Conférence	Human Language Technology	Anglais	Libre accès[b]	1986-2015	19
icassps	9 819	Conférence	IEEE International Conference on Acoustics, Speech and Signal Processing - Speech Track	Anglais	Propriétaire	1990-2015	26
ijcnlp	1 188	Conférence	International Joint Conference on NLP	Anglais	Libre accès[b]	2005-2015	6
inlg	227	Conférence	International Conference on Natural Language Generation	Anglais	Libre accès[b]	1996-2014	7
isca	18 369	Conférence	International Speech Communication Association	Anglais	Libre accès	1987-2015	28
jep	507	Conférence	Journées d'Etudes sur la Parole	Français	Libre accès[b]	2002-2014	5
lre	308	Revue	Language Resources and Evaluation	Anglais	Propriétaire	2005-2015	11
lrec	4 552	Conférence	Language Resources and Evaluation Conference	Anglais	Libre accès[b]	1998-2014	9
ltc	656	Conférence	Language and Technology Conference	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

Tableau 1 – (suite)

Nom court	# docs	Type	Nom long	Langue	Accès contenu	Période	# évènements
muc	149	Conférence	Message Understanding Conference	Anglais	Libre accès[b]	1991-1998	5
naacl	1 186	Conférence	North American Chapter of the ACL	Anglais	Libre accès[b]	2000-2015	11
paclic	1 040	Conférence	Pacific Asia Conference on Language, Information and Computation	Anglais	Libre accès[b]	1995-2014	19
ranlp	363	Conférence	Recent Advances in Natural Language Processing	Anglais	Libre accès[b]	2009-2013	3
sem	950	Conférence	Lexical and Computational Semantics / Semantic Evaluation	Anglais	Libre accès[b]	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[b]	2013-2015	3
tal	177	Revue	Revue Traitement Automatique du Langage	Français	Libre accès	2006-2015	10
taln	1 019	Conférence	Traitement Automatique du Langage Naturel	Français	Libre accès[b]	1997-2015	19
taslp	6 612	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[b]	1993-1998	3

Tableau 1 – (suite)

Nom court	# docs	Type	Nom long	Langue	Accès contenu	Période	# évènements
trec	1 847	Conférence	Text Retrieval Conference	Anglais	Libre accès	1992-2015	24
Total avec doublons	67 937					1965-2015	577
Total sans doublons	65 003						558

a. Les conférences jointes et les articles correspondants sont comptabilisés une seule fois dans le nombre total d'évènements et d'articles.
b. Inclus dans l'anthologie de l'ACL.

corpus, il faut enlever les doublons liés au fait que certaines conférences sont parfois organisées conjointement certaines années.

3. Traitement des données

Les documents du corpus ont été obtenus sous forme d'images scannées ou sous forme de document numériques. Dans le premier cas, il a été nécessaire de les transformer en utilisant un logiciel de reconnaissance de caractères. Dans certains cas, les documents sont accompagnés de métadonnées. Dans d'autres cas, il a fallu les extraire automatiquement des textes. L'extraction automatique d'informations a porté sur plusieurs sujets : noms des différents auteurs, avec leur affiliation, leur nationalité, leur genre, termes scientifiques, ressources linguistiques, citations (auteurs, titres, sources), agences de financement… Plusieurs traitements ont été effectués à l'aide du logiciel d'analyse syntaxique profonde TagParser (Francopoulo, 2007), basé sur un vaste lexique multilingue et sur la base de connaissance *Global Atlas*, provenant du contenu de 18 Wikipedias (Francopoulo *et al*., 2013).

4. Analyse de la production

4.1. Évolution du nombre de publications

Comme on le voit dans la figure 1, le nombre de publications a globalement augmenté au fil des ans, mais semble à présent se stabiliser.

4.2. Évolution du nombre d'articles

Le nombre d'articles augmente pour sa part constamment et de manière quasi exponentielle, pour atteindre plus de 65,000 documents en 2015 (figure 2).

Le nombre de documents pour chacune des sources est aussi très variable, allant de 18 369 documents pour la série des conférences ISCA jusqu'à 82 dans le cas des *ACM Transactions on Speech and Language Processing* (ACM-TLSP) (figure 3).

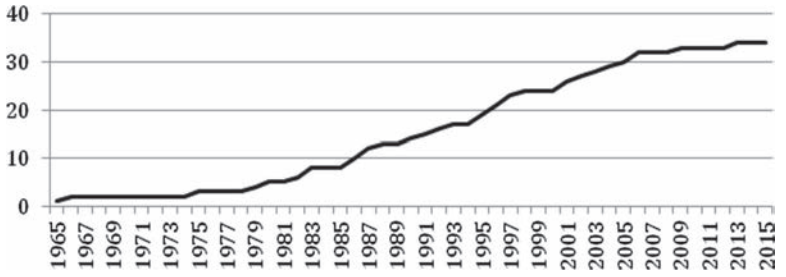

Figure 1. Nombre cumulé de publications (conférences et revues) différentes au fil des ans

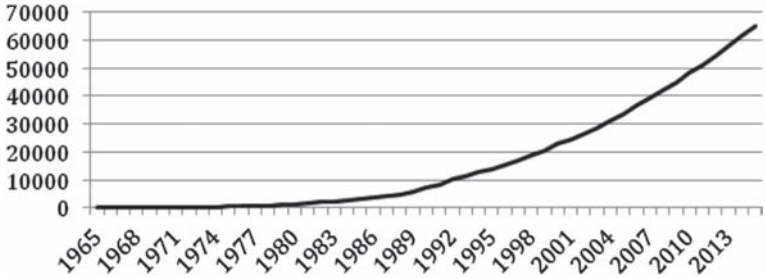

Figure 2. Nombre cumulé d'articles au fil des ans

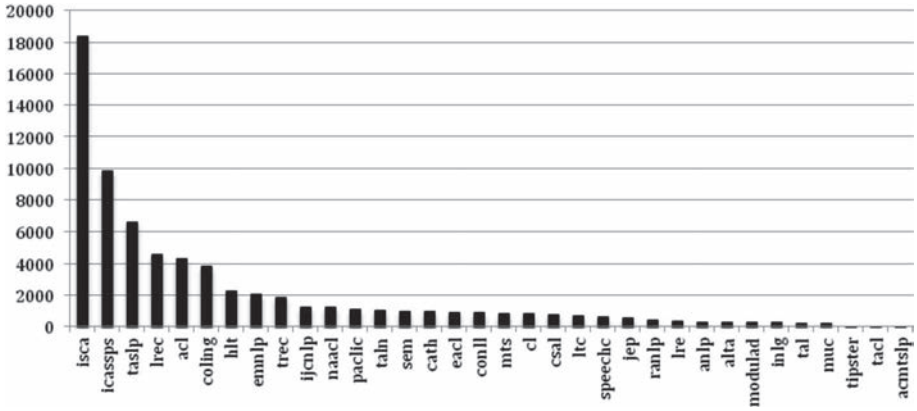

Figure 3. Nombre de documents pour chacune des sources

Cela est lié à l'ancienneté de la publication, à sa fréquence et aux nombres de documents qui sont publiés pour chaque évènement, qui est aussi très variable (figure 4). Ce sont les conférences Interspeech d'ISCA qui publient le plus d'articles à chaque évènement (656 en moyenne), suivies par LREC (506), ICASSP-Speech (378) IJCNLP

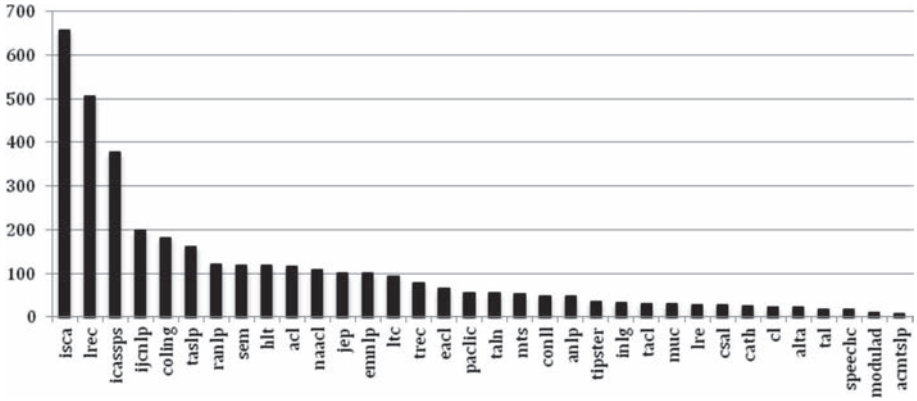

Figure 4. Nombre moyen de document pour chaque tenue de conférence ou numéro de revue

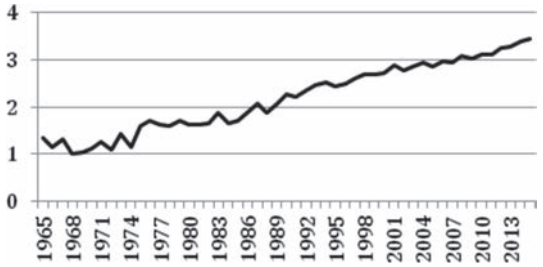

Figure 5. Évolution du nombre moyen d'auteurs par article au fil du temps

(198) et Coling (182). Les *ACM-TLSP* n'ont que 8 articles en moyenne à chaque numéro.

4.3. Analyse des auteurs

L'étude des auteurs est difficile du fait des variantes d'un même nom (nom de famille et prénom, initiales, initiales médianes, ordre, nom de femme mariée…). Cela nécessite donc une opération de nettoyage semi-automatique très fastidieuse (Mariani *et al*., 2014b) qui a abouti à une liste de 48 894 auteurs différents. Cela incite à trouver une façon d'identifier de manière unique les chercheurs (Joerg *et al*., 2012), à laquelle s'attachent également des organismes comme ORCID[8].

4.3.1. Évolution du nombre moyen de co-auteurs par article

Le nombre moyen de co-auteurs par article varie au cours du temps : de 1,33 en 1965 à 3,45 en 2015 (c'est-à-dire une augmentation moyenne de deux co-auteurs par article) (figure 5). Cela montre clairement le changement dans la façon de mener les recherches, qui est passée progressivement d'une recherche individuelle à des projets de taille importante conduits par des équipes ou en collaboration au sein de consortia, souvent dans des projets ou des programmes internationaux.

4.3.2. Renouvellement des auteurs

Nous avons ensuite étudié le renouvellement des auteurs au fil du temps (figure 6), soit le taux d'auteurs nouveaux d'une conférence à la suivante, pourcentage qui a diminué au fil du temps pour atteindre 61 % en 2015, soit en tant qu'auteurs n'ayant jamais publié dans la publication, qui a diminué jusqu'à atteindre 42 % en 2015. Cela exprime bien sûr une stabilisation dans le temps de la communauté scientifique, mais cela reflète aussi l'existence de « sang frais » dans cette communauté.

8. https://orcid.org/.

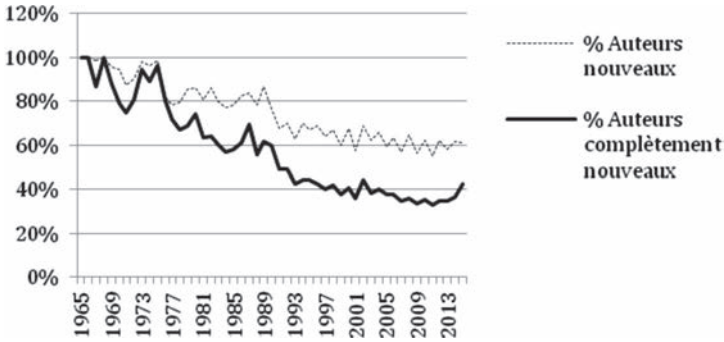

Figure 6. Pourcentage d'auteurs nouveaux et complètement nouveaux au fil du temps

4.3.3. Genre des auteurs

Nous avons conduit une étude du genre des auteurs à l'aide d'un lexique de 27 509 prénoms avec information sur leur genre (66 % masculin, 31 % féminin, 3 % épicène[9]). Comme on l'a déjà noté, les variations liées aux habitudes culturelles de nommer les gens (prénoms uniques ou multiples, nom de famille ou nom de clan, inclusion de particules honorifiques, ordre des composantes du prénom…) (Yu Fu *et al*., 2010), les changements dans les pratiques éditoriales et le partage d'un même nom par plusieurs personnes contribuent à rendre l'identification des individus par leur nom difficile (Vogel et Jurafsky, 2012). Dans certains cas, on n'a que les initiales du prénom, ce qui rend impossible l'identification du genre, à moins que cette même personne n'apparaisse avec la totalité de son prénom dans une autre publication. Bien que les résultats aient été vérifiés à la main par un expert du domaine, au moins pour les noms les plus « publiant », les résultats doivent être considérés comme possiblement entachés d'erreurs.

L'analyse de l'ensemble des publications montre que 49 % des auteurs sont masculins (22 858), 14 % féminins (6 746) et 37 % de genre indécidable (17 138), parce que leur prénom est épicène, ou parce que nous n'avons que leurs initiales. Si l'on estime que les auteurs de genre indécidable ont la même distribution que les autres, on obtient un pourcentage global d'auteurs masculins de 77 % et d'auteurs féminins de 23 %.

Si l'on observe la situation selon les sources (figure 7), on voit que les publications dans les domaines du traitement du signal (*IEEE Transactions on Speech and Language Processing* et ICASSP-S) ont le plus grand pourcentage d'auteurs masculins (respectivement 90 et 88 %), alors que les conférences et revues francophones, la revue LRE et la conférence LREC, ont le plus faible (de 63 à 70 %). L'analyse du genre

9. « Épicène » signifie que le prénom est de genre ambigü.

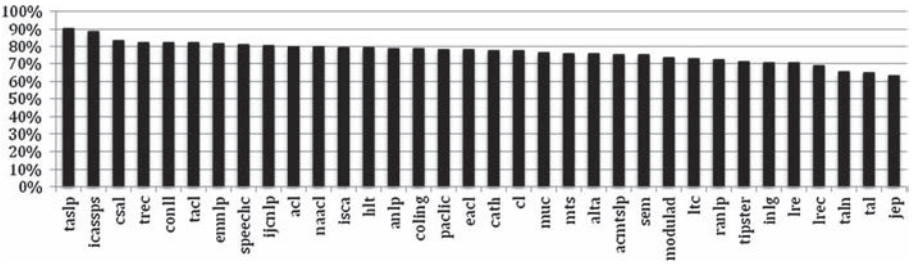

Figure 7. Pourcentage d'auteurs masculins selon les publications

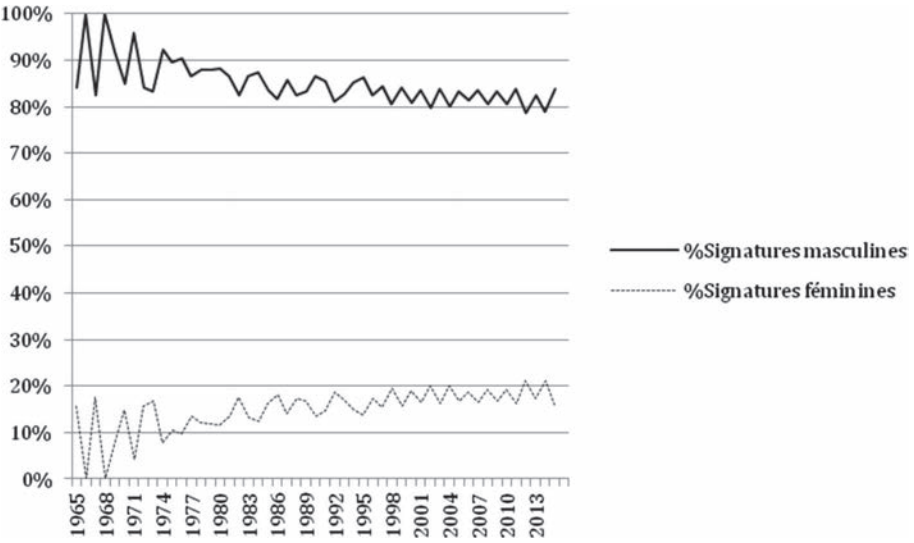

Figure 8. Genre des auteurs selon leurs contributions au fil du temps

des auteurs au fil du temps (figure 8) montre que le pourcentage d'auteurs féminins[10] a progressé lentement pour doubler, allant de 10 % à environ 20 %.

5. Collaborations entre auteurs

5.1. Production et co-production

L'auteur le plus productif a publié 358 articles, alors que 26 870 auteurs (55 % des 48.894 auteurs) n'ont publié qu'un seul article (figure 9). Le tableau 2 donne la liste des 10 auteurs les plus productifs, accompagné du nombre d'articles qu'ils ont signé comme unique auteur. Le tableau 3 donne le nombre d'auteurs suivant le nombre

10. On considère ici les pourcentages en tenant compte du nombre d'articles signés.

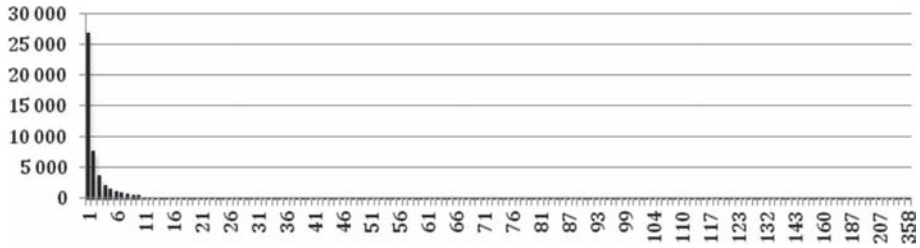

Figure 9. Nombre d'articles suivant le nombre de co-auteurs

d'articles qu'ils ont publiés seuls. 42 471 auteurs (87 % des auteurs) n'ont jamais publié un article comme seul auteur.

L'auteur le plus collaboratif a publié avec 299 co-auteurs différents, alors que 2 401 auteurs ont toujours publié seuls (figure 10). On retrouve aux deux premières places dans le tableau 4 les deux auteurs les plus productifs du tableau 2. Par contre, le troisième auteur le plus productif n'apparaît pas dans le classement. 108 auteurs ont publié avec 100 co-auteurs ou plus.

Tableau 2. Liste des 10 auteurs les plus productifs, incluant le nombre d'articles qu'ils ont publiés seuls

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

Tableau 3. Nombre d'articles signés comme auteur unique

# d'articles	# d'auteurs	Nom des auteurs
0	42 471	…
1	4 402	…
2	1 038	…
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

Tableau 3 – (suite)

# d'articles	# d'auteurs	Nom des auteurs
20	1	Olivier Ferret
21	1	Ralph Grishman
25	1	Ellen M Voorhees
26	1	Jerome R Bellegarda
27	1	W Nick Campbell

Figure 10. Nombre d'auteurs en fonction du nombre de co-auteurs différents

5.2. Graphe de collaboration

Un graphe de collaboration[11] (CollG) est un modèle de réseau social où les nœuds représentent les participants à ce réseau (habituellement des individus) et où deux participants distincts sont liés par un *arc* s'il y a une relation de collaboration (la co-signature d'un article) entre eux. Un CollG n'est pas orienté et ne contient ni boucle (un auteur ne collabore pas avec lui-même) ni arcs multiples (un seul arc relie deux auteurs, quelque soit le nombre d'articles qu'ils ont signés en commun). Les nœuds d'un CollG ne sont pas forcément tous connectés : les auteurs qui n'ont jamais eu de co-auteurs sont représentés par un nœud isolé (E). Les nœuds qui sont connectés forment une composante connexe (c'est le cas des auteurs A, B, C, D). Quand une composante connexe contient la majorité des nœuds, elle peut être qualifiée de composante géante. Les *Cliques* sont des composantes connexes où tous les auteurs ont publié entre eux : les nœuds sont alors tous reliés. Le graphe de collaboration de NLP4NLP contient 48.894 nœuds correspondant aux 48 894 auteurs différents, et 162 497 arcs (figure 11).

Comme on le voit dans le tableau 5, le graphe de collaboration contient 4.585 composantes connexes. La plus grande de ces composantes regroupe 39.744 auteurs, ce qui signifie que 81 % des 48 894 auteurs sont reliés à travers le réseau de collaboration.

11. http://en.wikipedia.org/wiki/Collaboration_graph.

Tableau 4. Liste des 12 auteurs ayant le plus grand 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

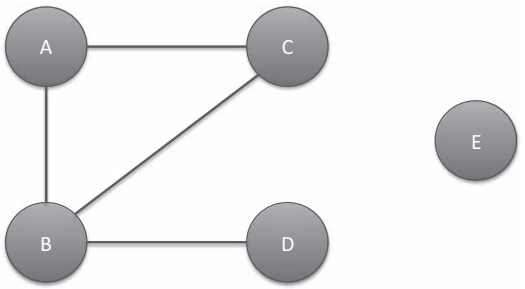

Figure 11. Graphe de collaboration

La composante connexe suivante regroupe 29 auteurs, qui ont publié ensemble mais n'ont jamais publié avec aucun des 39 744 auteurs précédents. Les composantes connexes suivantes regroupent un petit nombre d'auteurs n'ayant jamais publié avec les autres : ce sont les représentants de petites communautés rassemblées autour de l'étude d'une langue spécifique ou de thèmes périphériques. Comme on l'a déjà signalé, 5 % des auteurs (2.401) n'ont jamais publié avec un autre co-auteur, et il est apparu que la

Tableau 5. Composantes connexes dans le graphe de collaboration

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
39 744	1	39 744	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	1 008	2	5
3	535	1 605	3	12
2	1 113	2 226	5	24
1	2 401	2 401	5	52
39 963	4 585	48 894	100	100

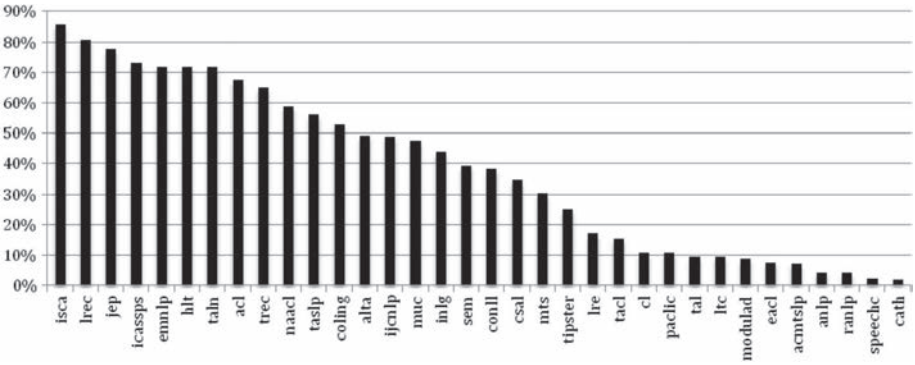

Figure 12. Pourcentage d'auteurs dans la plus grande composante connexe des graphes de collaboration pour les 34 publications

détermination des plus grandes cliques dans notre corpus s'obtient simplement en identifiant les auteurs de l'article qui a le plus grand nombre de co-auteurs (44 co-auteurs dans NLP4NLP). La figure 12 donne le pourcentage d'auteurs dans la plus grande composante connexe pour les 34 publications. On voit que certaines conférences internationales (ISCA, LREC, ICASSP-S, EMNLP, HLT) ou nationales (jep, taln) sont plus denses que d'autres où les collaborations sont moins fortes (EACL, ANLP, RANLP).

5.3. Mesures de centralité

Nous avons exploré le rôle des auteurs dans les CollG pour évaluer leur centralité. En théorie des graphes, il existe différents types de mesures de centralité (Freeman, 1978). La distance de proximité a été introduite dans les Sciences Humaines pour mesurer l'efficacité d'un réseau de communication (Bavelas, 1948, 1950). Elle est basée sur la distance géodésique la plus courte entre deux auteurs indépendamment du nombre de collaborations entre ces auteurs. La centralité de proximité est calculée comme étant la distance de proximité moyenne d'un auteur avec tous les autres auteurs appartenant à la même composante connexe. Nous utilisons plus précisément la centralité harmonique qui est un raffinement de la formule initiale introduit récemment par (Rochat, 2009) pour prendre en considération la totalité du graphe en une seule étape au lieu de prendre en compte chaque composante connexe séparément. La centralité de degré est simplement le nombre de co-auteurs d'un auteur, c'est-à-dire le nombre d'arcs attachés au nœud correspondant. La centralité d'intermédiarité est basée sur le nombre de chemins passant par un nœud et reflète l'importance d'un auteur comme passerelle entre différents groupes de co-auteurs (ou sous-communautés).

En regardant le tableau 6 qui est relatif au sous-corpus LREC, nous voyons que certains auteurs peuvent apparaître parmi les 10 les plus centraux pour les différents

types de centralité, éventuellement avec un rang différent, alors que d'autres n'apparaissent que pour une mesure de centralité.

6. Citations

6.1. Graphes de citations

Contrairement au graphe de collaboration, le graphe de citation (CitG) est orienté (figure 13). Dans un graphe de citation d'auteurs (ACG), les nœuds représentent des auteurs individuels. On peut considérer le graphe des auteurs citant *(CgAG)*, dans lequel un auteur citant est relié à tous les auteurs des articles qu'il cite par un arc orienté vers ces auteurs ; et un graphe des auteurs cités (CgAG), dans lequel un auteur cité est relié à tous les auteurs des articles qui le citent par un arc orienté vers cet auteur. Ces graphes peuvent avoir des boucles, un auteur pouvant se citer lui-même, mais ils ne peuvent avoir d'arcs multiples : il y a un seul arc reliant deux auteurs, quel que soit le nombre de fois où un auteur cite ou est cité par un autre auteur.

Dans un Graphe de citation d'articles (PCG), les nœuds représentent des articles individuels. Ici aussi, on peut considérer les graphes d'articles citant (CgPG), dans lesquels un article est relié à tous les papiers qu'il cite par un arc orienté vers ces articles ; et les graphes d'articles cités (CdPG), où chaque article est relié à tous les articles qui le citent par un arc orienté vers cet article. Ces graphes ne contiennent ni boucle, un article ne pouvant se citer lui-même, ni arcs multiples.

Les graphes de citation ne sont pas nécessairement connexes, un auteur pouvant ne citer aucun autre auteur ou n'être cité par aucun autre auteur, pas même lui-même, tout comme un article peut ne citer ou n'être cité par aucun autre article (E). Dans ces cas, les auteurs ou les articles apparaissent comme des nœuds isolés dans le graphe. Les nœuds qui sont reliés par un chemin orienté (c'est le cas pour A, B, C, D dans la figure 13 où l'auteur A cite l'auteur B et l'auteur C, l'auteur B cite l'auteur C, l'auteur C cite l'auteur A et l'auteur D cite B), constituent une composante fortement connexe. Les nœuds qui sont de plus reliés dans les deux directions constituent une composante symétrique fortement connexe. Elles sont communes dans les ACGs (l'auteur A cite l'auteur B et l'auteur B cite l'auteur A, par exemple), mais peu communes dans les PCGs (si l'article M cite l'article N, il y peu de chance que l'article N cite l'article M, dans la mesure où les articles citent d'autres articles qui sont déjà parus. Cela peut cependant arriver dans le cas de publications simultanées.

Nous avons étudié l'ensemble de ces graphes de citation pour l'ensemble du corpus et pour chacune des publications de manière interne ou dans le contexte du corpus NLP4NLP complet.

Nous donnons quelques éléments de comparaison entre les publications, sachant qu'il faut garder à l'esprit que les durées de vie et les fréquences sont différents, pour les conférences (9 conférences en 17 ans pour LREC, 28 en 27 ans pour ISCA et 36 en 35 ans pour ACL), comme pour les revues. Les 65 003 articles de NLP4NLP contiennent 314,042 références.

Tableau 6. Calcul et comparaison des mesures de centralité de proximité, centralité de degré et centralité d'intermédiarité pour les 10 auteurs les plus centraux dans la conférence LREC

Centralité de proximité			Centralité de degré		Centralité d'intermédiarité		
Nom de l'auteur	Centralité harmonique	Valeur relative	Nom de l'auteur	Index et valeur relative	Nom de l'auteur	Index	Valeur relative
Nicoletta Calzolari	2076	**1,000**	Nicoletta Calzolari	**1,000**	Khalid Choukri	269 538	**1,000**
Monica Monachini	1996	**0,961**	Khalid Choukri	**0,944**	Nicoletta Calzolari	202 365	**0,751**
Khalid Choukri	1983	**0,955**	Monica Monachini	**0,814**	Hans Uszkoreit	180 854	**0,671**
Núria Bel	1980	**0,954**	Núria Bel	**0,739**	Núria Bel	158 669	**0,589**
Bernardo Magnini	1941	**0,935**	Bernardo Magnini	**0,708**	Bernardo Magnini	157 090	**0,583**
Stelios Piperidis	1933	**0,931**	Asunción Moreno	**0,689**	Asunción Moreno	151 440	**0,562**
Asunción Moreno	1910	**0,920**	Hans Uszkoreit	**0,658**	Monica Monachini	144 944	**0,538**
Dan Tufis	1903	**0,917**	Stelios Piperidis	**0,609**	Martha Palmer	133 788	**0,496**
Joseph Mariani	1893	**0,912**	Dan Tufis	**0,540**	Ulrich Heid	133 446	**0,495**
Hans Uszkoreit	1889	**0,910**	Jan Hajič	**0,534**	Stephanie M Strassel	120 573	**0,447**

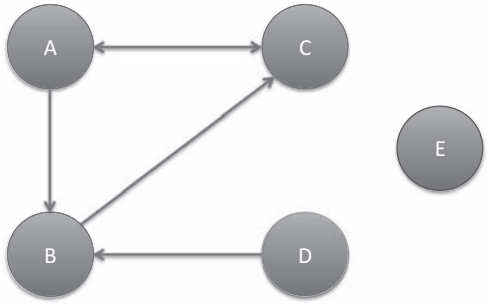

Figure 13. Graphe de citation

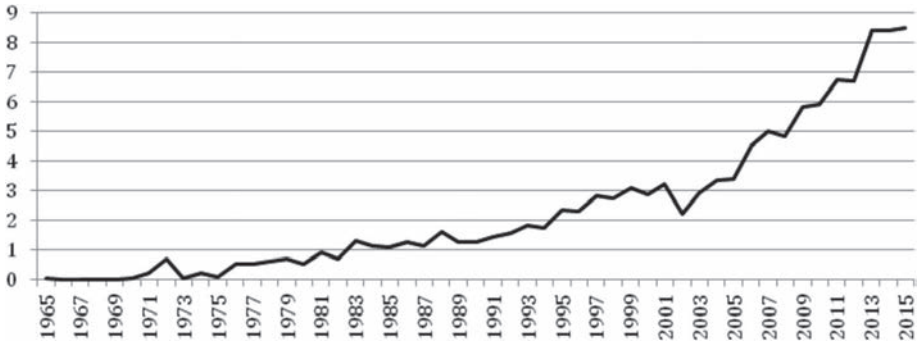

Figure 14. Nombre moyen de références par article au fil des ans

6.2. Nombre de citations au fil du temps

Nous avons étudié les citations dans les articles qui existent sous forme numérique originale (non scannés). 58.204 de ces articles contiennent des références.

Si l'on regarde le nombre moyen de références par article, nous voyons qu'il augmente au fil du temps de près de 0 en 1965 à 8,5 en 2015 (figure 14). C'est une évolution générale qui résulte des usages de citation et du nombre d'articles publiés dans la littérature scientifique[12].

Il est difficile de conduire l'étude comparative de l'évolution au fil du temps du nombre de références et du nombre de citations pour les 34 publications. Si on limite cette étude à 8 conférences importantes (ACL, COLING, EACL, EMNLP, ICASSP, ISCA, LREC, NAACL), on voit que le nombre de références augmente fortement au fil du temps pour ISCA (figure 15). Cela est directement en relation avec la règle décidée

12. Il faut cependant rappeler que nous ne considérons ici que les données du corpus NLP4NLP.

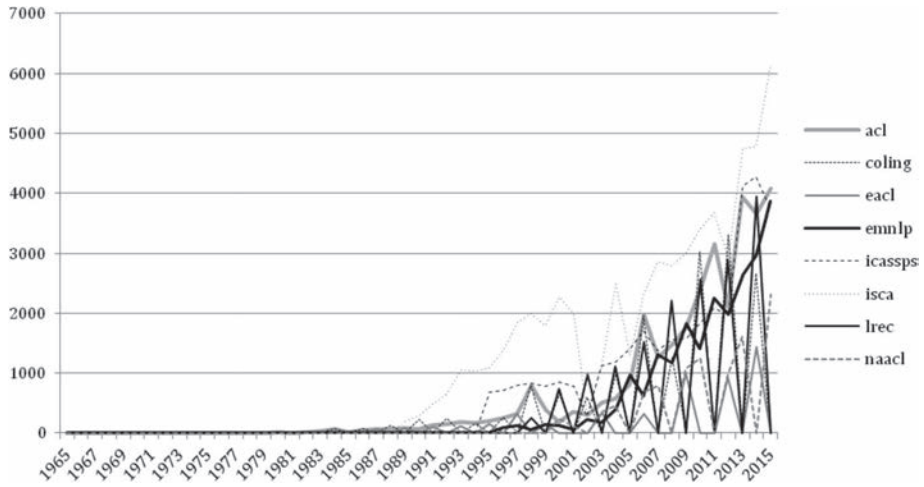

Figure 15. Évolution du nombre de références dans les articles au fil du temps pour 8 conférences importantes

par cette association en 2005 d'accroître le nombre de page de chaque article publié lors des conférences annuelles de 4 à 5, à la condition que la page supplémentaire ne contienne que des références, pour encourager les auteurs à mieux se citer entre eux. La courbe en dents de scie des conférences LREC, EACL et NAACL est due au fait qu'elles sont biennales.

De la même façon, il est difficile d'analyser l'évolution du nombre d'articles cités au cours du temps comme on le voit en figure 15, du fait des fréquences variables de ces conférences. Pour résoudre ce problème, on peut considérer la somme des articles cités jusqu'à une année donnée. On voit alors (figure 16) que les articles d'ISCA cités croissent fortement. Il en est de même avec les articles d'ACL, avec un retard à présent comblé. ICASSP arrive en troisième position, puis suit un groupe avec COLING et EMNLP, suivi de LREC et NAACL et enfin EACL.

6.3. Citations des auteurs

Nous avons ensuite étudié les graphes de citation d'auteurs (figure 17) et comparé le pourcentage d'auteurs se trouvant dans la plus grande composante fortement connexe à l'intérieur de chacune des 34 publications. On voit que les auteurs d'un ensemble de publications d'origine nord-américaine (*Computational Linguistics*, EMNLP, CONLL, HLT, NAACL, ACL, TACL) ont une forte habitude de se citer les uns les autres.

Le tableau 7 donne le classement des 10 auteurs les plus cités, avec le nombre de citations. On peut mettre en regard le nombre de publications de ces auteurs, ainsi

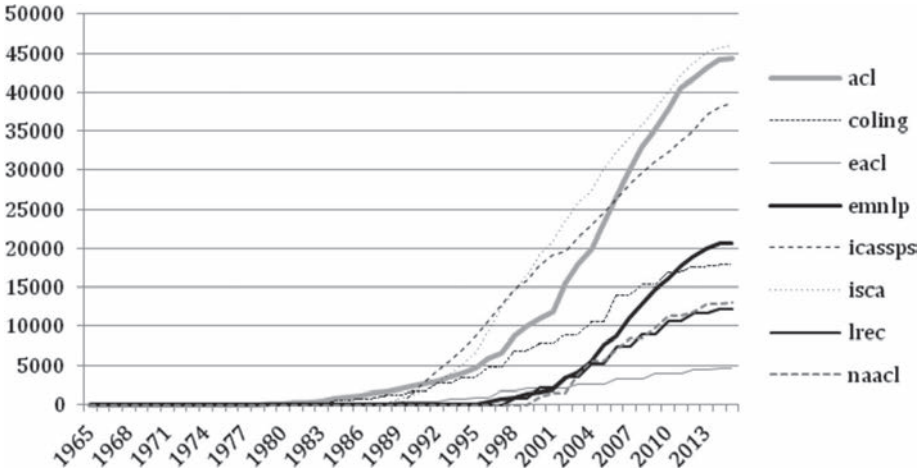

Figure 16. Évolution du nombre d'articles ayant été cités au fil du temps pour 8 conférences importantes

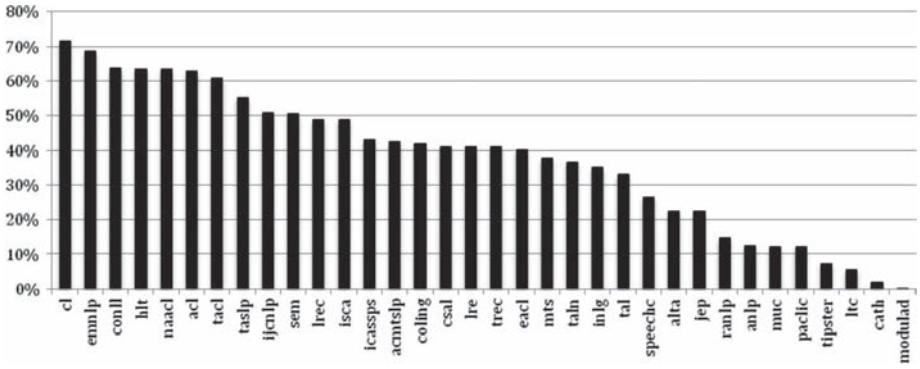

Figure 17. Pourcentage d'auteurs dans la plus grande composante fortement connexe

que le ratio entre le nombre de citations et le nombre d'articles signés. On voit que ce ratio est très variable, certains auteurs ayant relativement peu publié, mais étant très cités. On a également calculé le pourcentage d'auto-citations (citations dans un article rédigé par le même auteur) qui montre également des pratiques variables.

6.4. Citations des articles

La figure 18 donne le nombre moyen d'articles (degré moyen) de chaque publication citée dans l'ensemble des 34 publications. On voit que les articles publiés dans *Computational Linguistics* sont de loin les plus cités, avec plus de 20 citations en

Tableau 7. Liste des 10 auteurs les plus cités

Nom	Nombre de citations	Nombre d'articles écrits par l'auteur	Ratio nombre de citations / nombre d'articles écrits par l'auteur	Pourcentage d'auto-citations (%)
Hermann Ney	5 200	343	15	18
Franz Josef Och	4 098	42	98	2
Christopher D Manning	3 972	116	34	5
Philipp Koehn	3 121	39	80	2
Dan Klein	3 080	99	31	8
Michael John Collins	3 077	53	58	4
Andreas Stolcke	3 053	130	23	7
Mark J F Gales	2 540	195	13	19
Salim Roukos	2 505	67	37	2
Chin-Hui P Lee	2 450	218	11	18

moyenne par article. Suivent NAACL, ACL et EMNLP, puis HLT et CONLL. Les articles des revues dans le domaine du traitement de la parole (CSAL, TASLP, *Speech Communication*) et surtout des conférences sont moins cités, en ligne avec les usages des communautés scientifiques correspondantes, et il est évident que les articles sont moins cités s'ils sont publiés dans une langue autre que l'anglais (voir TAL, TALN, JEP et Modulad).

Il est impressionnant de constater (tableau 8) que 42 % des articles ne sont jamais cités et que 40 % des auteurs ne sont jamais cités. Après analyse, il apparaît que certains de ces auteurs appartiennent à une autre communauté scientifique, dans laquelle eux et leurs articles sont cités, mais ont très peu publié dans NLP4NL.

6.5. H-Index

Une publication a un H-Index de N si N est le plus grand nombre d'articles qui sont parus et qui sont cités au moins N fois. Le calcul du H-Index pour les 34 publications (figure 19) montre que c'est la conférence ACL qui a le plus grand H-Index, avec 75 articles cités 75 fois ou plus. Elle est suivie de TASLP (66), *Computational Linguistics* (58), HLT (56), EMNLP (55), ICASSP-S (54) et ISCA (51).

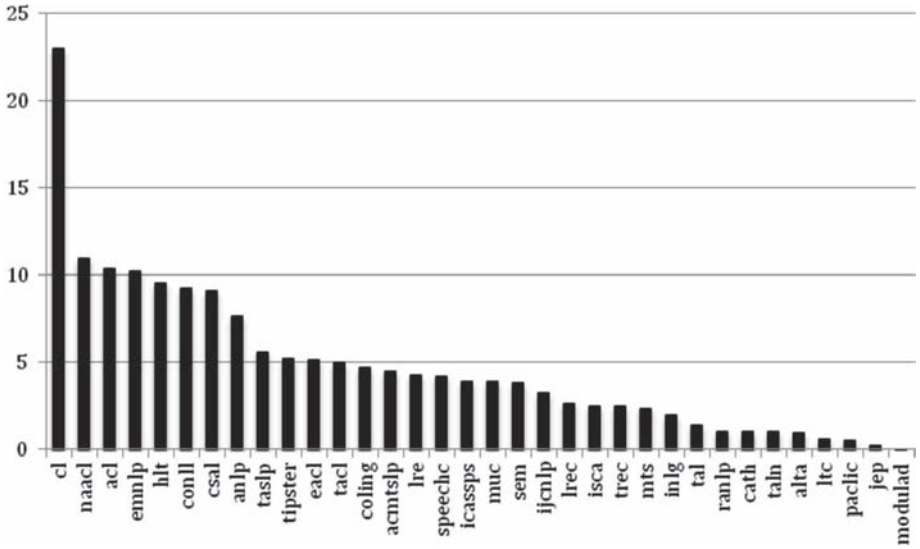

Figure 18. Degré moyen d'articles cités pour les 34 publications

Tableau 8. Articles et auteurs jamais cités

	Nombre	Pourcentage (%)
Articles jamais cités	27 183	42
Auteurs jamais cités	19 740	40

Cette analyse porte sur toute la durée de 50 années couverte par NLP4NLP, mais ne prend en compte que les publications contenues dans le corpus. On peut comparer ces résultats avec le H-Index de Google Scholar[13] daté de mars 2016 où la totalité de la littérature scientifique est explorée mais uniquement sur la période des cinq dernières années (tableau 9). La conférence ACL y apparaît aussi en première position avec un H-Index de 65 et un H5-median (moyenne du nombre de citations de ces 65 articles) de 99, suivie par EMNLP (56), IEEE ICASSP (54), IEEE TASLP (51) et NAACL (48), et on note la forte progression de LREC (38).

13. http://scholar.google.com/citations?view_op=top_venues&hl=en&vq=eng_computationallinguistics.

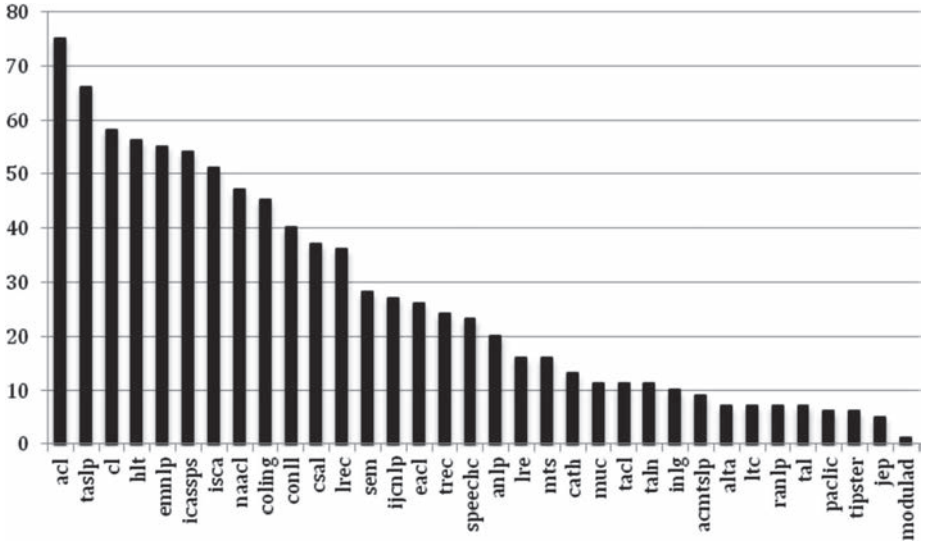

Figure 19. H-Index général pour les 34 publications

7. Utilisation des ressources linguistiques

Nous avons mené une analyse de la mention des ressources linguistiques dans les articles du corpus. Les ressources linguistiques sont les briques qu'utilisent les chercheurs pour mener leurs recherches et développer leurs systèmes. Dans une acception large, elles regroupent les données (corpus, lexiques, dictionnaires, bases terminologiques…), les outils (analyseurs morpho-syntaxiques, analyseurs prosodiques, logiciels d'annotations…), les éléments d'évaluation des systèmes (métriques, logiciels, corpus d'apprentissage, d'essais et de test) et les méta-ressources (bonnes pratiques, formalismes, normes, standards…). Nous avons pris en compte les ressources recensées dans la LRE Map (Calzolari *et al*., 2012). Cette base de données a été mise en place dans le cadre du projet Européen FlaReNet et est constituée par les auteurs des articles des différentes conférences du domaine, qui sont invités lors de la soumission de leurs propositions d'article à remplir un questionnaire donnant les principales caractéristiques des ressources linguistiques produites ou utilisées dans le cadre de leurs travaux de recherche rapportés dans leur article. La base LRE Map que nous avons utilisée est constituée des informations recueillies dans 10 conférences de 2010 à 2012. Le nombre de ressources collectées est de 4 396. Après nettoyage de ces entrées (corrections des noms de ressources, élimination des doublons, regroupement des différentes versions d'une même famille de ressources…), nous avons obtenu 1 301 ressources différentes dont la mention dans les articles du corpus NLP4NLP a été étudiée.

Le tableau 10 donne le classement des ressources linguistiques en fonction du nombre d'articles où elles sont mentionnées (que l'on qualifie de « présence »). Il donne également pour chaque ressource son type (corpus, lexique, outil…), le nombre de

Tableau 9. Classement 2016 des 20 publications ayant le plus fort H-Index selon Google Scholar sur 5 années (2011-2015)

Rang	Publication	H-5 Index	H-5 Median
1	Meeting of the Association for Computational Linguistics (ACL)	65	99
2	Conference on Empirical Methods in Natural Language Processing (EMNLP)	56	81
3	IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)	54	73
4	IEEE Transactions on Audio, Speech and Language Processing (TASLP)	51	78
5	North American Chapter of the Association for Computational Linguistics (NAACL)	48	71
6	International Conference on Spoken Language Processing (INTERSPEECH)	39	70
7	International Conference on Language Resources and Evaluation (LREC)	38	64
8	International Conference on Computational Linguistics (COLING)	38	59
9	arXiv Computation and Language (cs.CL)	37	70
10	Computer Speech and Language (CSL)	32	51
11	Speech Communication (SpeCom)	32	49
12	Computational Linguistics (CL)	31	40
13	Conference on Computational Natural Language Learning (CONLL)	24	36
14	Language Resources and Evaluation (LRE)	23	42
15	International Workshop on Semantic Evaluation (SEMEVAL)	23	41
16	Conference of the European Chapter of the Association for Computational Linguistics (EACL)	21	34
17	International Joint Conference on Natural Language Processing (IJCNLP)	20	27

Tableau 9 – (suite)

Rang	Publication	H-5 Index	H-5 Median
18	IEEE Spoken Language Technology Workshop (SLT)	18	28
19	Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)	18	27
20	Workshop on Statistical Machine Translation	18	24

mentions de la ressource (« occurrences »), les premiers auteurs qui l'ont mentionnée, ainsi que la première publication, la première année et la dernière année où elle a été mentionnée. On voit que « WordNet » apparaît en première position, suivi de « Timit », « Wikipedia », du « Penn Treebank » et du logiciel d'analyse de parole « Praat ».

Nous avons étudié l'évolution du nombre de ressources citées (présence) en rapport du nombre d'articles publiés au fil des ans (figure 20). Il apparaît que les courbes correspondantes se croisent en 2005, date à partir de laquelle plus d'une ressource est donc mentionnée en moyenne dans chaque article. On peut y trouver l'illustration du passage de la suprématie des méthodes à base de connaissances (*Knowledge-based*) à celle des méthodes guidées par les données (*Data-driven*).

On peut également étudier la propagation d'une ressource dans le corpus d'articles. La figure 21 donne la propagation de la ressource « WordNet », apparue initialement dans la conférence HLT en 1991, puis se diffusant les années suivantes d'abord dans les conférences de linguistique computationnelle, puis également dans celles du traitement de la parole.

On peut également attribuer un facteur d'impact aux ressources en fonction du nombre d'articles qui les citent, suivant les présences telles qu'elles apparaissent dans le tableau 10. Le tableau 11 donne ces Facteurs d'impact pour les ressources de type « données » et de type « outils ».

8. Thèmes de recherche

8.1. Fréquence et présence des termes

La modélisation des thèmes de recherche d'un domaine scientifique est un challenge en traitement du langage naturel (voir par exemple Paul et Roxana (2009), Hall *et al.* (2008)). Notre approche est d'extraire les termes scientifiques dans le corpus et de calculer leur fréquence globale et l'évolution de cette fréquence dans le temps pour refléter l'importance et l'évolution des thèmes de recherche qu'ils représentent. Nous

Tableau 10. Mention des ressources linguistiques de la LRE Map dans les articles de NLP4NLP

Ressource	Type	# prés.	# occur.	Premiers auteurs mentionnant la RL	Premiers corpus mentionnant la RL	Première année	Dernière année	Rang
WordNet	Lexique (texte)	4 203	29 079	Daniel A Teibel, George A Miller	hlt	1991	2015	1
Timit	Corpus (parole)	3 005	11 853	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	Corpus (texte)	2 824	20 110	Ana Licuanan, J H Xu, Ralph M Weischedel	trec	2003	2015	3
Penn Treebank	Corpus (texte)	1 993	6 982	Beatrice Santorini, David M Magerman, Eric Brill, Mitchell P Marcus	hlt	1990	2015	4

Tableau 10 – (suite)

Ressource	Type	# prés.	# occur.	Premiers auteurs mentionnant la RL	Premiers corpus mentionnant la RL	Première année	Dernière année	Rang
Praat	Outil (parole)	1 245	2 544	Carlos Gussenhoven, Toni C M Rietveld	isca	1997	2015	5
SRI Language Modeling Toolkit	Outil (texte)	1 029	1 520	Dilek Z Hakkani-Tür, Gökhan Tür, Kemal Oflazer	coling	2000	2015	6
Weka	Outil (logiciel)	957	1 609	Douglas A Jones, Gregory M Rusk	coling	2000	2015	7
Europarl	Corpus (texte)	855	3 119	Daniel Marcu, Franz Josef Och, Grzegorz Kondrak, Kevin Knight, Philipp Koehn	acl, eacl, hlt, naacl	2003	2015	8
FrameNet	Lexique (texte)	824	5 554	Beryl T Sue Atkins, Charles J Fillmore, Collin F Baker, John B Lowe, Susanne Gahl	acl, coling, lrec	1998	2015	9
GIZA++	Outil (logiciel)	758	1 582	David Yarowsky, Grace Ngai, Richard Wicentowski	hlt	2001	2015	10

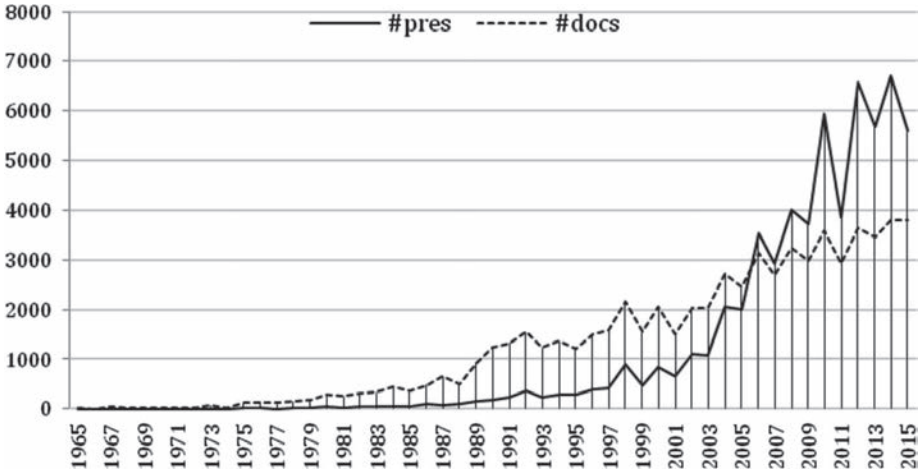

Figure 20. Évolution du nombre de mentions des ressources (#pres) dans les articles (#docs) au fil du temps

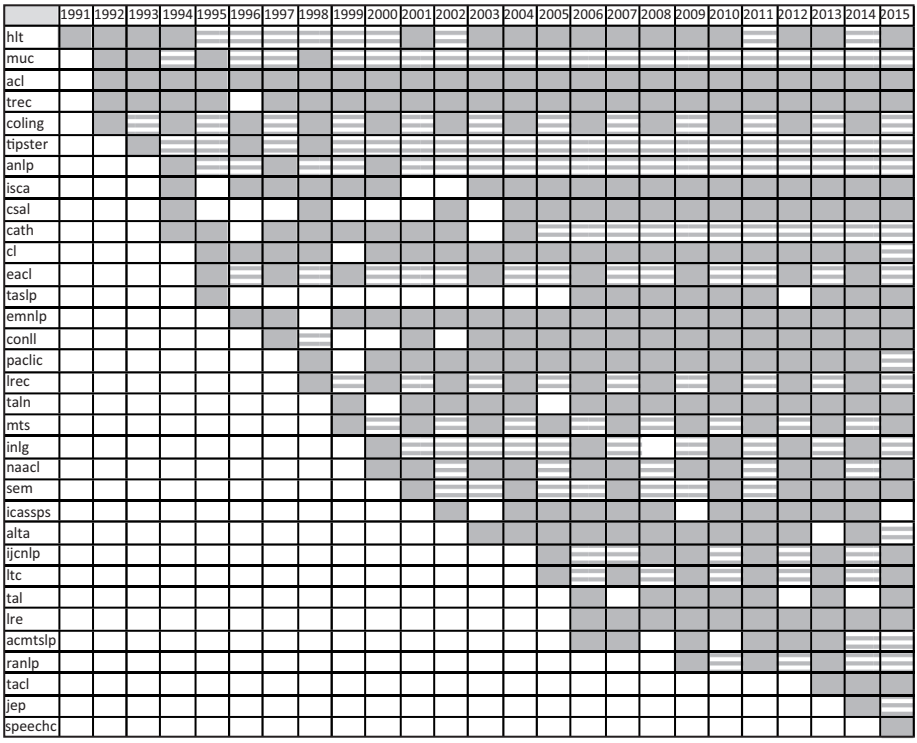

Figure 21. Diffusion de la mention de la ressource « Wordnet » dans les conférences et revues de NLP4NLP. Les cases hachurées correspondent à des années où la conférence n'a pas eu lieu ou la revue pas publiée

Tableau 11. Facteur d'impact des ressources linguistiques (données et outils)

Données	Facteur d'impact (présence)	Outils	Facteur d'impact (présence)
Wordnet	4 203	Praat	1 254
Timit	3 005	SRI Language Modeling Toolkit	1 029
Wikipedia	2 824	Weka	957
Penn Treebank	1 993	GIZA++	758
Europarl	855		
FrameNet	824		

partons donc du corpus NLP4NLP qui contient un total de 269 539 220 mots, la plupart en anglais. Pour extraire de ce corpus les termes scientifiques, il nous a fallu éliminer les éléments du langage courant, ce que nous avons fait en extrayant les groupes nominaux grâce au logiciel d'analyse syntaxique profonde TagParser[14] et en utilisant des corpus de langue anglaise hors du domaine spécialisé étudié ici : British National Corpus (BNC)[15] (BNC, 2007), Open American National Corpus (OANC[16]) (Ide *et al*., 2010), Suzanne corpus release-5[17], archives EuroParl (de 1999 à 2009)[18], et une collection de textes des domaines du sport, de la politique et de l'économie, pour un total de 200 millions de mots, en veillant à ne pas y inclure de textes liés au traitement du langage.

Nous avons exploré les 61 661 documents en langue anglaise et extrait 3 485 408 termes différents (unigrammes, bigrammes and trigrammes) et 23 871 856 occurrences de termes, que nous avons regroupés en tenant compte des variantes dans leur expression.

Le tableau 12 donne le classement des 10 termes les plus fréquents selon leur nombre d'occurrences dans le corpus. Y sont également mentionnés les variantes dans l'expression du terme, la fréquence correspondante et le nombre d'articles où il apparaît, dans l'absolu (présence) et relativement au nombre total d'articles (présence relative).

14. www.tagmatica.com.
15. www.natcorp.ox.ac.uk.
16. www.americannationalcorpus.org.
17. www.grsampson.net/Resources.html.
18. www.statmt.org/europarl.

Tableau 12. Classement des 10 termes les plus fréquents dans le corpus NLP4NLP

Terme	Variantes	Nombre d'occurrences	Fréquence	Nombre de présences	Présence relative
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	134 060	0,00609	14 353	0,22671
SR	ASR, ASRs, Automatic Speech Recognition, SRs, Speech Recognition, automatic speech recognition, speech recognition	128 590	0,00584	20 324	0,32102
LM	LMs, Language Model, Language Models, language model, language models	111 582	0,00507	12 809	0,20232
Annotation	Annotations	111 142	0,00505	11 992	0,18942
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	101 333	0,0046	13 803	0,21802
Classifier	Classifiers	98 092	0,00446	11 513	0,18185
NP	NPs, noun phrase, noun phrases	94 808	0,00431	9 584	0,15138

Tableau 12 – (suite)

Terme	Variantes	Nombre d'occurrences	Fréquence	Nombre de présences	Présence relative
Parser	Parsers	86 901	0,00395	9 636	0,1522
Segmen-tation	Segmentations	76 232	0,00346	10 850	0,17138
SNR	SNRs, Signal Noise Ratio, Signal Noise Ratios, signal noise ratio, signal noise ratios	68 722	0,00312	6 848	0,10817

8.2. Évolution des thèmes

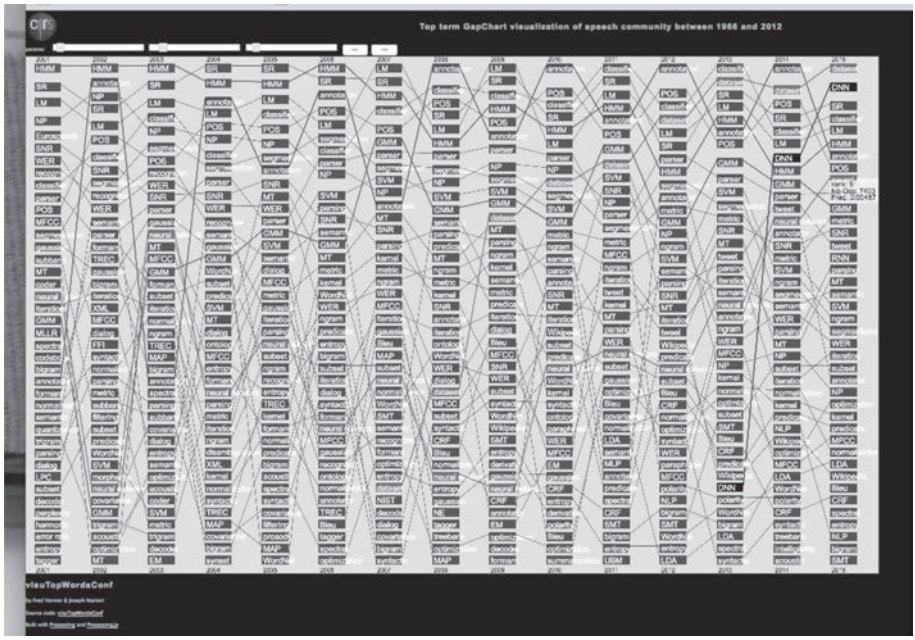

Figure 22. Évolution de la fréquence des termes au fil du temps pour la conférence ISCA-Interspeech de 2001 à 2015

Nous avons ensuite étudié l'évolution de ces thèmes au court du temps. Un logiciel de visualisation[19] a été réalisé pour donner le classement des termes au fil des ans en pouvant jouer sur différents paramètres : période étudiée, nombre de thèmes classés, sélection des thèmes étudiés, classement suivant la fréquence ou la présence (Perin *et al.*, 2016). La figure 22 donne cette évolution dans la conférence ISCA-Interspeech pour les termes « HMM » (Hidden Markov Models), « GMM » (Gaussian Mixtures Models), « Annotation », « Neural Networks », « DNN » (Deep Neural Network) et « Dataset ». On y trouve que la popularité des HMM, après être restée au tout premier rang et avoir été rejointe par les GMM, se tasse un peu ces dernières années. L'évolution en dents de scie de « Annotation » est liée au caractère biennal de la conférence LREC, où ce terme est fréquemment employé puisque la conférence porte sur les ressources linguistiques. Après avoir été à la mode, les réseaux neuronaux ont été en déclin puis sont revenus sur le devant de la scène avec les *Deep Neural Networks* (DNN) et les datasets qui les alimentent.

19. Gapchart : http://vernier.frederic.free.fr/Infovis/rankVis4/.

8.3. Introduction des termes par les auteurs et les publications : une mesure d'innovation ?

Nous avons également étudié qui a introduit un terme scientifique dans le champ de la recherche, quand et dans quelles publications. Nous avons considéré les 61 661 articles rédigés en anglais et les 42 278 auteurs qui ont utilisé les 3 485 408 termes contenus dans ces articles. Le tableau 13 donne la liste des 10 termes classés suivant leur présence en 2015, dernière année que nous avons prise en considération, qui peut refléter leur « succès » actuel, avec pour chacun d'entre eux la mention de la première année où il a été mentionné, les auteurs qui l'ont mentionné pour la première fois et la publication où il a été mentionné, ainsi que le nombre d'occurrences et de présences en 2015. On voit ainsi que « dataset » a été introduit volontairement en 1966 par Laurence Urdang dans *Computer and the Humanities*, et que cette année-là le terme n'était mentionné que dans un seul article, alors qu'il apparaît dans 1 474 articles en 2015 ! À partir de ces informations, nous étudions actuellement la façon de calculer une mesure d'innovation attachée à un auteur ou à une publication.

8.4. Prédiction de thèmes de recherche

Nous avons également exploré s'il était possible de prédire l'évolution des termes de recherche pour les années à venir en fonction du passé. Pour cela, nous avons utilisé la suite de logiciels d'apprentissage automatique Weka[20] (Witten *et al*., 2011). Nous avons appliqué chacun des logiciels aux séries chronologiques de termes classés suivant leur fréquence et retenu celui qui donnait les meilleurs résultats de prédiction avec son paramétrage (durée de l'historique considéré, en particulier) en les vérifiant a posteriori. Nous l'avons alors appliqué aux données de l'ensemble des éléments du corpus NLP4NLP. Le tableau 14 donne le classement des 10 termes les plus fréquents en 2013 et 2014 avec leur fréquence, les thèmes prédits pour 2015 en fonction des classements passés et le classement réellement observé en 2015. On voit que la prédiction est correcte pour le premier terme (« dataset »). Le terme prédit ensuite était « annotation » qui n'apparaît qu'en neuvième place, LREC n'ayant pas eu lieu en 2015. Il est suivi de « POS » qui se trouve effectivement apparaître un rang derrière à la quatrième place.

Puisque nous avons l'information sur les observations réelles des classements annuels, il nous est possible de mesurer la fiabilité des prédictions en mesurant la distance entre les fréquences prédites pour les termes et les fréquences observées. La figure 23 donne cette distance pour des prédictions sur les années 2011 à 2015 à partir des séries chronologiques allant jusqu'à 2010. On voit que cette distance croît nettement en 2013, soit 3 ans après que la prédiction ait été effectuée et on peut donc penser qu'on ne peut raisonnablement prédire ce qui va se passer dans un domaine de recherche au-delà d'un horizon de 2 ans (voire moins en cas de découverte majeure).

20. www.cs.waikato.ac.nz/ml/weka.

Tableau 13. Liste des 10 termes les plus populaires en 2015 classés suivant leur présence dans les articles

Terme	Année d'apparition du terme	Auteurs ayant introduit le terme	Article(s)	Nombre d'occurrences du terme la dernière année	Nombre d'articles mentionnant le terme (présence) la dernière année
Dataset	1966	Laurence Urdang	cath1966-3	14 060	1 474
Classifier	1967	Aravind K Joshi, Danuta Hiz	C67-1007	8 202	999
Linear	1967	Marian W Cobin	cath1967-5	2 065	918
Optimization	1967	Ellis B Page	C67-1032	3 317	901
Normalization	1967	Bruce A Beatie	cath1967-16	2 974	774
HMM	1982	Cory S Myers, Stephen E Levinson	taslp1982-86	7 575	688
Acoustic	1967	David Shillan	C67-1018	1 906	646
Spectral	1971	Arne Zettersten	cath1971-12	2 486	608
Filtering	1973	Eugenio Morreale, Massimo Mennucci	C73-2024	1 719	604
Toolkit	1980	C Raymond Perrault	J80-3003	1 155	603

Tableau 14. Prédiction de thèmes de recherche avec le logiciel 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

Il est également possible d'effectuer la mesure de la distance entre prédiction et observation chaque année. Cela donne une mesure de la « surprise », comme étant la différence entre ce à quoi on s'attendait et ce qui est effectivement réalisé, les années où cette « surprise » est la plus grande pouvant correspondre à des ruptures épistémologiques. La figure 24 donne l'évolution de cette distance de 2011 à 2015. On voit que 2012 semble avoir été une année de changements importants, qui de fait correspond au regain d'intérêt pour les méthodes neuromimétiques et à l'émergence de l'apprentissage profond (*Deep Learning*).

On peut également effectuer la mesure de cette distance pour un terme particulier, permettant de voir comment ce terme évolue en deçà ou au delà de ce qu'on prévoyait. La figure 25 donne cette évolution pour le terme « *Deep Neural Network* » (DNN). On

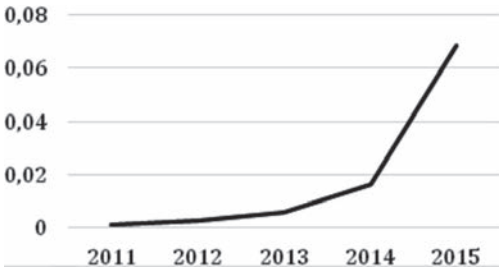

Figure 23. Fiabilité des prédictions : erreur des prédictions effectuées à partir de 2011 selon l'horizon temporel

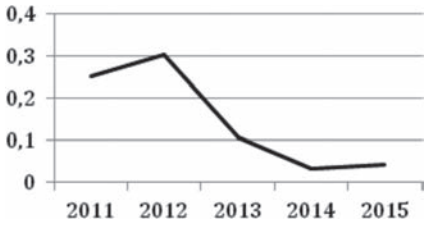

Figure 24. Évolution de la distance entre prédiction et observation au fil du temps comme une mesure de la « surprise » pouvant correspondre à une rupture épistémologique

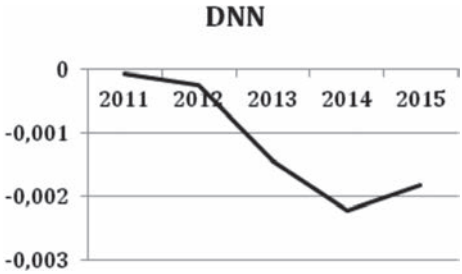

Figure 25. Mesure de l'appréhension de l'émergence d'un thème de recherche : les Deep Neural Networks

voit que jusqu'en 2014, on ne prévoyait pas que cette approche connaîtrait un tel succès, et qu'à partir de 2014, la surprise est moins grande, cette approche entrant dans la panoplie habituelle des outils de traitement automatique de la langue.

Le tableau suivant donne les prédictions pour les cinq années à venir à compter de 2016 : sans surprise on s'attend à ce que les réseaux de neurones plus ou moins profonds et plus ou moins récurrents continuent d'occuper l'attention des chercheurs (tableau 15).

Tableau 15. Prédictions pour les cinq années à venir (2016-2020)

Observé 2014	Observé 2015	Prédiction 2016	Prédiction 2017	Prédiction 2018	Prédiction 2019	Prédiction 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

Tableau 16. Définitions de l'(auto-)réutilisation et de l'(auto-)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

9. Réutilisation et plagiat

Nous avons enfin étudié la réutilisation et le plagiat d'articles au sein du corpus. Pour cela, nous avons comparé un à un les 65,003 articles contenus dans NLP4NLP, après analyse syntaxique profonde effectuée à partir du logiciel TagParser (Francopoulo, 2007) pour réduire l'importance des variantes stylistiques et ne pas tenir compte des tournures de langage générales, et leurs 48,894 auteurs. La comparaison entre un article et les articles antérieurs ou parus la même année s'effectue ensuite en comparant des fenêtres de sept entités lexicales à l'aide de la distance de Jaccard et, après expérimentations, en retenant comme semblables les couples d'articles dont la ressemblance est supérieure à 4 %. On considère alors quatre cas possibles définis ainsi : si les deux articles ont au moins un auteur en commun et que la source est citée, on parlera « d'auto-réutilisation », sinon « d'auto-plagiat ». Si les deux articles n'ont pas d'auteurs en commun et que la source est citée, on parlera de « réutilisation », sinon de « plagiat » (tableau 16). Les résultats montrent que le nombre d'auto-réutilisations et d'auto-plagiats est très important, avec 18 % des articles ainsi catalogués. Ce nombre est trop important pour que l'on puisse mener une vérification manuelle. 205 articles ont le même titre et 130 articles ont le même titre et exactement la même liste d'auteurs ! Le tableau 17 donne le nombre de couples d'articles identifiés comme auto-réutilisés ou auto-plagiés pour chaque paire de publications. On voit que le flux d'articles est particulièrement important entre les conférences IEEE-ICASSP et ISCA-Interspeech, ainsi qu'entre les conférences et les revues d'un même domaine, comme entre IEEE-ICASSP ou ISCA-Interspeech et TASLP, CSAL ou Speech Com, ce qui semble très normal. Par contre, le nombre de réutilisations et de plagiats est très faible et ne concerne à peine que 0,3 % des articles, comme le montre le tableau 18 qui donne le nombre d'articles identifiés comme réutilisés ou plagiés. Une vérification manuelle possible ici en raison du faible nombre de cas détectés montre que la quasi-totalité des plagiats détectés n'en sont même pas vraiment (mauvaise orthographe du nom des auteurs cités ou du titre de l'article cité, référence correcte à un autre article antérieur…).

Nous avons alors étudié le laps de temps séparant une première publication d'une réutilisation (figure 26). Il apparaît que 38 % des réutilisations se font la même année, 71 % l'année suivante, 83 % dans les deux ans et 93 % dans les 3 ans.

Considérons à présent le délai entre une première publication en conférence et une réutilisation dans un article de revue (figure 27). On observe un même délai retardé

Tableau 17. Matrice d'auto-réutilisation et d'auto-plagiat (en grisé : les 7 sources les plus copiant ou copiées)

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	13	1	0	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	

Tableau 18. Matrice de réutilisation et de plagiat (en grisé : les 7 sources les plus copiant ou copiées)

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	0	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	0	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	0	1	0	2	0	1	0	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	1	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	0	1	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	0	3	3	0	muc
naacl	1	0	0	0	0	0	0	0	1	0	1	1	1	0	0	1	0	0	0	0	0	0	0	1	2	0	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	0	0	0	0	1	0	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	0	0	2	0	2	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	

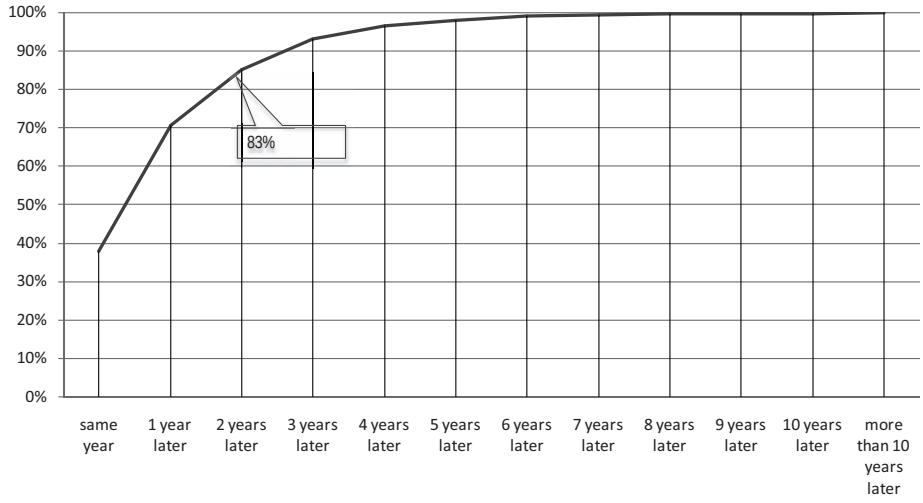

Figure 26. Délai entre une publication et sa réutilisation (en %)

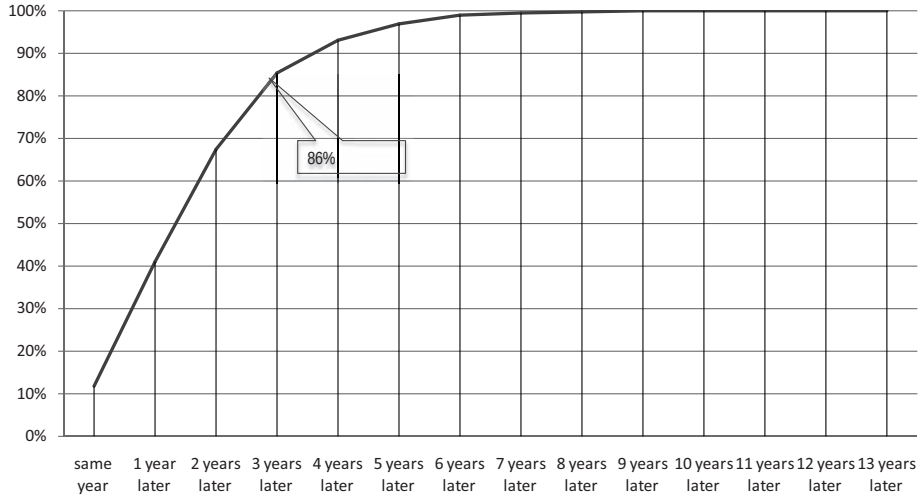

Figure 27. Délai entre publication en conférence et réutilisation en revue (en %)

d'une année : 12% des articles réutilisés le sont la même année, 41 % dans l'année suivante, 68 % dans les 2 ans et 86 % dans les 3 ans.

10. Conclusions et perspectives

Nous avons présenté ici un rapide panorama des principaux résultats obtenus dans l'analyse d'un vaste corpus, NLP4NLP, couvrant une grande partie des publications

relatives au traitement du langage naturel sur une période longue et récente des 50 dernières années, où des avancées notables ont été effectuées, grâce à des efforts de recherche continus et constants rendus possibles par la mise en place d'une infrastructure rassemblant programmes incitatifs de recherche, disponibilité de ressources linguistiques et organisations régulières de campagnes d'évaluation.

Nous avons été confrontés dans cette analyse à l'absence de standards dans l'identification d'entités comme les noms d'auteurs, leur genre, leur affiliation, les titres d'articles ou même de publications, les noms des agences de financement. Établir des standards faciliterait hautement le travail, mais nécessite une coordination internationale afin que les identifiants soient uniques et persistants.

Il nous reste à présent à terminer nos travaux relatifs à la détermination d'une mesure d'innovation que l'on pourrait appliquer aux auteurs et aux publications. Nous souhaiterions pouvoir également automatiser plus avant l'extraction des termes ou des noms d'auteurs en réduisant les taux d'erreurs, analyser la polarité des citations et mieux détecter les signaux faibles indiquant l'approche d'un nouveau paradigme scientifique et pouvant provenir de sources éloignées du domaine que nous traitons.

Remerciements

Les auteurs remercient les collègues de l'ACL, Ken Church, Sanjeev Khudanpur, Amjbad Abu Jbara, Dragomir Radev et Simone Teufel, qui les ont aidés dans la phase de démarrage, Isabel Trancoso, qui a fourni l'analyse des archives d'ISCA sur l'utilisation de l'évaluation et des corpus, Wolfgang Hess, qui a produit et diffusé 14 GOctets d'archives ISCA, Emmanuelle Foxonet qui a transmis une liste d'auteurs avec une information sur leur prénom et leur genre, Florian Boudin, qui a rendu utilisable une anthologie de TALN, Helen van der Stelt et Jolanda Voogd (Springer) qui ont transmis les données de la revue LRE, et Douglas O'Shaughnessy, Denise Hurley, Rebecca Wollman et Casey Schwartz (IEEE) qui ont fourni celles d'IEEE ICASSP et TASLP. Ils remercient également Khalid Choukri, Alexandre Sicard et Nicoletta Calzolari, qui ont fourni les informations sur les conférences LREC, Nicoletta Calzolari, Riccardo del Gratta, Khalid Choukri Irene Russo, Francesco Rubino et Claudia Soria pour la mise en place et la diffusion de la LRE Map, Victoria Arranz, Ioanna Giannopoulou, Johann Gorlier, Jérémy Leixa, Valérie Mapelli et Hélène Mazo, qui ont aidé à récupérer les métadonnées de LREC 1998, et les organisateurs, les relecteurs et les auteurs des conférences et revues des 50 dernières années sans lesquels nous n'aurions jamais pu mener cette analyse !

Excuses

Cette analyse a été menée sur des données textuelles qui couvrent une période de 50 années, et incluent pour les années les plus anciennes des contenus qui ont été scannés et numérisés. Elle utilise des outils de traitement automatique des contenus des articles scientifiques qui peuvent faire des erreurs. Il convient donc de considérer les résultats comme pouvant comporter une certaine marge d'erreur. Les auteurs

souhaitent donc s'excuser à l'avance des erreurs que le lecteur pourra détecter, et seront heureux de les corriger dans les versions ultérieures de cette analyse.

Bibliographie

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

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

Bavelas A. (1950). Communication patterns in task oriented groups. *Revue of the Acoustical Society of America*, vol. 22, p. 271-282.

Calzolari N., Del Gratta R., Francopoulo G., Mariani J., Rubino F., Russo I., Soria C. (2012). The LRE Map. Harmonising Community Descriptions of Resources. In: *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, 23-25 May 2012.

Ding Y., Rousseau R., Wolfram D. (Eds). (2014). *Measuring Scholarly Impact*, Springer. 2014, ISBN: 978-3-319-10376-1.

Dunne C., Shneiderman B., Gove R., Klavans J., Dorr B. (2012). Rapid understanding of scientific paper collections: integrating statistics, text analytics, and visualization. *Revue of the American Society for Information Science and Technology*, vol. 63, n° 12, p. 2351-2369.

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

Francopoulo G., Marcoul F., Causse D., Piparo G. (2013). Global Atlas : Proper Nouns, from Wikipedia to LMF. In: Francopoulo G., ed. *LMF-Lexical Markup Framework*, ISTE/Wiley.

Francopoulo G., Mariani J., Paroubek P. (2015a). NLP4NLP : the cobbler's children won't go unshod. In: *4th International Workshop on Mining Scientific Publications (WOSP2015), Joint Conference on Digital Libraries 2015 (JCDL 2015)*, Knoxville (USA), June 24, 2015.

Francopoulo G., Mariani J., Paroubek P. (2015b). NLP4NLP : Applying NLP to written and spoken scientific NLP corpora. In: *Workshop on Mining Scientific Papers: Computational Linguistics and Bibliometrics, 15th International Society of Scientometrics and Informetrics Conference (ISSI 2015)*, Istanbul (Turkey), June 29, 2015.

Francopoulo G., Mariani J., Paroubek P. (2015c). NLP4NLP : the cobbler's children won't go unshod. *D-Lib*, vol. 21, n° 11/12. www.dlib.org/dlib/november15/francopoulo/11francopoulo.html.

Freeman L.C. (1978). Centrality in social networks, conceptual clarifications. *Social Networks*, vol. 1, n° 1978/79, p. 215-239.

Fu Y., Xu F., Uszkoreit H. (2010). Determining the origin and structure of person names. In: *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, May 2010, Valletta, Malta, European Language Resources Association (ELRA), ISBN: 2-9517408-6-7. p. 3417-3422.

Gollapalli S.D., Li X.-l. (2015). EMNLP *versus* ACL: analyzing NLP research over time. In: *EMNLP 2015*, Lisbon (Portugal), September 17-21, 2015.

Hall D.L.W., Jurafsky D., Manning C. (2008). Studying the History of Ideas Using Topic Models. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'08)*, 363-371.

Ide N., Suderman K., Simms B. (2010). ANC2Go: a web application for customized corpus creation. In: *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.

Jha R., Jbara A.-A., Qazvinian V., Radev D.R. (2016). NLP-driven citation analysis for scientometrics. *Natural Language Engineering*, Available on CJO 2016, doi:10.1017/S1351324915000443.

Joerg B., Höllrigl T., Sicilia M.-A. (2012). 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.

Li H., Councill I., Lee W.C., Giles C.L. (2006). CiteSeerx: an architecture and web service design for an academic document search engine. In: *Proceedings of the 15th Int. Conference on the World Wide Web*.

Mariani J. (1990). La conférence IEEE-ICASSP de 1976 à 1990 : 15 ans de recherches en traitement automatique de la parole. In: *Notes et Documents LIMSI 90-8*, septembre 1990.

Mariani J. (2013). *The ESCA enterprise, ISCA web site-about ISCA-history*. http://www.isca-speech.org/iscaweb/index.php/about-isca/history.

Mariani J., Paroubek P., Francopoulo G., Delaborde M. (2013). Rediscovering 25 Years of Discoveries in Spoken Language Processing : a Preliminary ISCA Archive Analysis. In: *Proceedings of Interspeech 2013*, 26-29 August 2013, Lyon, France.

Mariani J., Paroubek P., Francopoulo G., Hamon O. (2014a). Rediscovering 15 years of discoveries in language resources and evaluation: the LREC anthology analysis. In: *Proceedings of LREC 2014*, 26-31 May 2014, Reykjavik, Iceland.

Mariani J., Cieri C., Francopoulo G., Paroubek P., Delaborde M. (2014b). Facing the Identification Problem in Language-Related Scientific Data Analysis. In: *Proceedings of LREC 2014*, 26-31 May 2014, Reykjavik, Iceland.

Mariani J., Francopoulo G., Paroubek P., Vetulani Z. (2015). Rediscovering 10 to 20 years of discoveries in language & technology. In: *Proceedings of L&TC 2015*, 27-29 November 2015, Poznan, Poland.

Mariani J., Paroubek P., Francopoulo G., Hamon O. (2016). Rediscovering 15 + 2 years of discoveries in language resources and evaluation. *Language Resources and Evaluation Journal*, p. 1-56, ISSN: 1574-0218, doi:10.1007/s10579-016-9352-9.

Mariani J., Francopoulo G., Paroubek P. (2017). Reuse and plagiarism in speech and natural language processing publications. *International Journal on Digital Libraries*. doi:10.1007/s00799-017-0211-0.

Osborne F., Motta E., Mulholland P. (2013). Exploring Scholarly Data with Rexplore, International Semantic Web Conference, Sydney, Australia.

Paul M., Roxana G. (2009). Topic modeling of research fields: an interdisciplinary perspective. In: *Recent Advances in Natural Language Processing (RANLP 2009)*, Borovets, Bulgaria.

Perin C., Boy J., Vernier F. (2016). GapChart: a gap strategy to visualize the temporal evolution of both ranks and scores. *IEEE Computer Graphics and Applications*, Special issue on Sports Data Visualization, September/October 2016.

Radev D.R., Muthukrishnan P., Qazvinian V., Abu-Jbara A. (2013). The ACL anthology network corpus. *Language Resources and Evaluation*, vol. 47, p. 919-944.

Rochat Y. (2009). Closeness centrality extended to unconnected graphs: the harmonic centrality index. In: *Applications of Social Network Analysis (ASNA), 2009*, Zurich, Switzerland.

Tang J., Zhang J., Yao L., Li J., Zhang L., Su Z. (2008). ArnetMiner: extraction and mining of academic social networks. In: *Proceeding of the 14th Int. Conference on Knowledge Discovery and Data Mining*.

The British National Corpus. (2007). *Version 3 (BNC XML Edition)*. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/.

Vogel A., Jurafsky D. (2012). He said, she said: gender in the ACL anthology. In: *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries* (ACL'12). Association for Computational Linguistics, Stroudsburg, PA, USA, p. 33-41.

Witten I.H., Eibe F., Hall M.A. (2011) *Data mining: practical machine learning tools and techniques*. 3rd ed. Morgan Kaufmann, Burlington, USA.