Présence et représentation des femmes dans le traitement automatique des langues en France

Karën Fort, Aurélie Névéol

Présence et représentation des femmes dans le traitement automatique des langues en France

Karën Fort*, Aurélie Névéol**

*Univ. Paris-Sorbonne/STIH, 28, rue Serpente, 75006 Paris, France
karen.fort@paris-sorbonne.fr,
http://karenfort.org/
**LIMSI-CNRS, rue John von Neumann, 91405 Orsay, France
aurelie.neveol@limsi.fr
https://perso.limsi.fr/neveol/

Résumé. Nous présentons ici les résultats d'une étude menée dans un premier temps dans le cadre du blog « Éthique et TAL » sur la place des femmes dans le domaine du traitement automatique des langues (TAL) en France [1]. Cette étude montre que bien que les femmes représentent près de la moitié (47 %) des adhérent·e·s à l'association savante du TAL, l'ATALA, elles restent, malgré les efforts réalisés, sous-représentées dans les positions les plus visibles et prestigieuses. Le recueil des données est par ailleurs rendu complexe du fait de l'absence de statistiques sur le sujet. Nous proposons donc des solutions pratiques et des outils pour rendre visible ces déséquilibres, afin de pouvoir les traiter.

1 Introduction

Le traitement automatique des langues (TAL) est une discipline hybride dans laquelle évoluent à la fois des informaticien·ne·s et des linguistes. Sans surprise, une comparaison rapide entre la composition des enseignant·e·s-chercheur·e·s dans les sections CNU 27 (informatique) et 7 (sciences du langage) montre une opposition franche entre ces domaines, puisqu'on trouve 69 % de Maîtresses de conférences en section 7, pour 26 % en section 27, et respectivement 50 % et 19 % de Professeures [2]. Il nous a donc paru intéressant d'analyser comment cette mixité disciplinaire se reflète dans la place des femmes dans les instances qui la représentent.

En 2016, les organisateurs de la plus prestigieuse des conférences du domaine, la rencontre annuelle de l'ACL (Association for Computational Linguistics), ont demandé, *via* leur blog, de proposer des responsables de domaines (*area chairs*) pour le comité de lecture. Ils ont ajouté quelques statistiques quant aux propositions reçues dans un billet [3], où ils appellent à plus de diversité. Entre autres déséquilibres (notamment une sur-représentation des Américains (56 %) et des Européens (32 %)), 78 % des proposé·e·s (par eux-mêmes ou par des tiers) sont des hommes. Ils en profitent pour citer le rapport sur les procédures de nomination à ACL présenté

1. Voir : http://www.ethique-et-tal.org/2016/11/18/la-question-quon-ne-posait-pas/.
2. Voir : http://cache.media.enseignementsup-recherche.gouv.fr/file/statistiques/18/6/Section_27_768186.pdf.
3. https://acl2017.wordpress.com/2016/11/11/last-call-for-area-chairs-a-call-for-diversity/

lors d'ACL 2016. Ce rapport a été commandité suite à des remarques sur le manque de diversité dans les instances de l'association. Il détaille sept recommandations pour améliorer la situation, en particulier concernant les procédures de nomination des membres de différentes structures liées à l'ACL (par exemple, les *area chairs* de la conférence). Deux de ces recommandations (5 et 6, p. 3) visent à sensibiliser les membres des instances et plus largement de la communauté aux questions de diversité.

Le problème a donc été reconnu, analysé et des solutions sont proposées. Le rapport recommande également un suivi de la situation (*via* des statistiques sur le sujet) sur le long terme. En effet, pour que les choses changent, encore faut-il que le problème soit identifié et qu'un suivi régulier soit assuré (Desrosières et Didier, 2014).

En France, ce travail a été réalisé en partie pour la communauté « extraction et gestion des connaissances » à l'occasion du défi EGC 2016 par Cabanac et al. (2016). Cependant, à notre connaissance, il n'existe encore aucun équivalent pour la communauté TAL et nous ne disposons pas de statistiques facilement accessibles. Nous avons donc arpenté les sites Web des conférences TALN, de l'ATALA, de la revue TAL, nous avons demandé de l'aide, sur les réseaux sociaux et ailleurs, pour retrouver des informations désormais ensevelies dans les plis de la mémoire numérique.

2 Représentation des femmes dans le TAL en France

2.1 Association savante

L'association pour le traitement automatique des langues (ATALA), notre association savante, comprend deux instances de direction : le comité permanent (CPERM) et le conseil d'administration (CA).

Le CPERM, dont la composition varie constamment, du fait de la présence en son sein des organisateurs de la conférence TALN ($n-1$, n, $n+1$), comprend actuellement 9 hommes et 7 femmes (soit presque 44 % de femmes). Ce presque équilibre est une réussite remarquable. Il est particulièrement intéressant de noter que la parité est parfaite parmi les membres cooptés (ceux qui ont le mandat le plus long, 4 ans) : 2 hommes et 2 femmes. La situation est beaucoup moins équilibrée au CA, avec 5 femmes pour 15 hommes (25 % de femmes).

Notons également que les présidents des deux instances sont actuellement des hommes. En ce qui concerne la présidence de l'ATALA, cela n'a pas toujours été le cas mais les femmes restent minoritaires dans ce rôle (12,5 % de femmes).

2.2 Revue nationale principale

L'une des très grandes réussites de l'ATALA est sa revue, auto-gérée et en accès libre, la revue TAL. Cette revue ne pourrait pas fonctionner sans son comité de relecture (CR), qui réalise un important travail, afin de publier chaque année trois numéros, dont en général un *varia* (numéro non thématique, dont les rédacteur·trice·s en chef·fe sont membres du CR) et deux numéros spéciaux (avec un rédacteur·trice en chef.fe membre du CR et des co-rédacteur·trice·s en chef·fe invité·e·s). A l'heure actuelle, le CR de la revue comprend 33 membres (et un·e secrétaire), dont 10 femmes (soit un peu plus de 30 % de femmes). Il est à noter que les membres du CR sont cooptés et non élus par la communauté ou le CA de l'ATALA.

Si l'on considère les numéros disponibles en ligne, hors *varia* (dont les rédacteur·trice·s en chef·fe sont des membres du CR), on y trouve 15 femmes et 30 hommes comme rédacteur·trice·s en chef·fe et seuls deux numéros (sur une vingtaine) n'ont que des femmes comme rédactrices en cheffe (à comparer aux 9 qui n'ont que des hommes comme rédacteurs en chef) [4].

2.3 Conférence nationale principale

Une rapide analyse des comités d'organisation des différentes conférences TALN montre que sur les 22 éditions, seules 2 ont été présidées par des femmes seules.

En ce qui concerne les conférencier·ère·s invité·e·s, l'affaire est moins simple, car les données sont parfois difficiles à trouver. Nous avons pu obtenir les noms des invités pour tous les TALN entre 2005 et 2017 (sachant qu'il n'y en a pas eu en 2009 et en 2014). Nous avons identifié 29 intervenants, dont seulement 8 sont des femmes (soit à peu près 28 %).Les données concernant les prix TALN et RECITAL sont disponibles sur le site de l'ATALA pour les éditions 2008 à 2017. Ainsi, parmi les auteur·trice·s des articles primés sur cette période, on compte 6 hommes et 6 femmes (soit 50 % de femmes) pour RECITAL, et 10 femmes et 25 hommes (soit 29 % de femmes) pour TALN. Il est intéressant de noter que sur les 14 articles primés à TALN sur cette période, 7 ont une femme comme première autrice (soit 50 %). Pour continuer dans les prix, le prix de thèse de l'ATALA a lui été attribué quatre fois à un homme (2011, 2012, 2013, 2017) et trois fois à une femme (2014, 2015, 2016) - soit 43 %.

Les informations sont encore plus difficiles à excaver concernant les comités de chaque conférence, nos données sur le sujet sont relativement éparses, donc moins fiables. Le comité de programme (ou d'organisation) compte de 22 (2014) à 33 % (2005 et 2016) de femmes selon les années et le comité de lecture (ou scientifique) entre 25 et 30 %. Lister les président·e·s de sessions (*chairs*) pour chaque conférence est une gageure, mais en 2014, les femmes étaient 3 (sur 12), en 2016, elles étaient 5 (sur 13) et en 2017 elles étaient 5 (sur 14) soit environ 33 %. En 2017, quatre tables rondes ont été organisées, avec une parité parfaite pour trois d'entre elles (9 hommes et 9 femmes participant au total) et une quatrième table ronde exclusivement masculine (5 hommes) ramenant la participation globale des femmes aux tables rondes à 39 %.

Il est à noter que le choix des président·e·s de session, des relecteur·trice·s (comité de lecture) et des organisateur·trice·s se fait par cooptation. A notre connaissance, les conférencier·ère·s invité·e·s sont choisis par le CPERM à partir d'une liste proposée par les organisateurs de la conférence.

3 Combien de femmes, dans le TAL français ?

Toutes ces données n'ont d'intérêt pour l'analyse que si l'on connaît la proportion de femmes dans le domaine. Or, cette information n'est pas facilement disponible.

3.1 Autrices d'articles à TALN

La part des femmes parmi les auteur·trice·s des articles acceptés dans les conférences TALN a été présentée par Patrick Paroubek lors de l'assemblée générale de l'ATALA en 2014

4. À titre de comparaison, le rapport femmes/hommes dans les comités de rédaction de 77 revues en systèmes d'information était de 15/85 en 2011 (Cabanac, 2012).

à l'occasion des 20 ans de la conférence TALN. Les chiffres montrent une évolution modeste sur deux décennies, avec 24 % de femmes autrices en 1997 (pour 73 % d'hommes et 3 % d'auteurs au prénom mixte ou de genre inconnu) contre 29 % de femmes autrices en 2014 (pour 57 % d'hommes et 13 % d'auteurs au prénom mixte ou de genre inconnu) [5]. Le même travail réalisé par Mariani et al. (2014) sur l'anthologie de la conférence LREC [6] estime à 34 % la part des femmes dans les auteur·trice·s d'articles de notre domaine.

Le problème de ce type de source (outre les prénoms difficiles à classifier) est qu'il pourrait induire des biais en cascade : il n'est en effet pas impossible que les femmes voient leurs articles moins souvent acceptés que ceux des hommes, comme il a été montré par Wenneras et Wold (1997). Mais en l'absence d'autre source d'information, nous étions prêtes à évaluer la part des femmes dans notre domaine, en France, à environ 30 %.

3.2 Membres de l'ATALA

D. Nouvel et P. Paroubek nous ont fourni une information fondamentale : le sexe des adhérent·e·s à l'ATALA. Si l'on considère la totalité des adhérent·e·s de 2003 à 2016, on obtient 640 femmes, 696 hommes et 247 personnes au prénom mixte ou de genre inconnu, soit un taux de 47 % de femmes en excluant les inclassables. Même si tous les inclassables étaient des hommes (943), on aurait donc plus de 40 % de femmes parmi les membres de notre association savante.

Soit les 13 % d'auteurs au prénom mixte ou de genre inconnu de TALN 2014 sont en fait des femmes, soit les femmes publient moins à TALN, soit, pour une raison inconnue, elles s'inscrivent davantage à l'ATALA.

4 Conclusions et propositions

La première conclusion, peu surprenante, est qu'il existe bien un déséquilibre dans le TAL. Il est important de noter qu'il est plus marqué lorsqu'il s'agit de positions plus visibles (conférences invitées, présidences, etc), ce qui correspond à l'observable dans la fonction publique (Drucker-Godard et al., 2017) [7] et en général (effet « plafond de verre »). On pourrait sans doute réduire assez rapidement l'écart en sensibilisant au problème les membres des différentes instances citées ici, en rappelant l'état de l'art sur les biais en défaveur des femmes en sciences (Moss-Racusin et al., 2012; Larivière et al., 2013) et en s'inspirant, pourquoi pas, des recommandations de l'ACL. Encore faudrait-il pour cela des données, car « ce qui n'est pas compté ne compte pas ».

La deuxième conclusion de cette étude est en effet que malgré des efforts récents, nous manquons de données publiées, en particulier en ce qui concerne les conférences TALN (présidences de sessions, responsabilités intermédiaires). En outre, l'information publiée ici concernant les membres de l'ATALA ne l'est nulle part ailleurs. Enfin, nous n'avons pas accès aux articles refusés à TALN ou à la revue TAL et ne pouvons donc évaluer le taux de refus pour

5. Il est à noter que la conférence est passée entre temps à une relecture en double aveugle.

6. Les soumissions à LREC ne sont pas anonymes.

7. Voir `http://www.cnrs.fr/mpdf/spip.php?article205` pour le CNRS et `http://cache.media.enseignementsup-recherche.gouv.fr/file/Charte_egalite_femmes_hommes/90/6/Chiffres_parite_couv_vdef_239906.pdf` pour l'enseignement supérieur.

les femmes et les hommes Il faut noter que les articles sont maintenant proposés sous forme anonyme, il n'y a donc a priori pas de biais de genre. Névéol et al. (2017) ont cependant montré que cet anonymat est tout relatif dans une communauté aussi réduite que la nôtre alors qu'il est plus robuste dans un contexte international.

Afin d'affiner certains résultats, comme le nombre de femmes dans le TAL, nous pourrions suggérer de compter les hommes et les femmes par *crowdsourcing* (participation volontaire) dans les amphithéâtres des conférences auxquelles nous participons. L'application `IT Counts` permet de réaliser ce type de comptage facilement et de le partager [8]. Nous pourrions également entrer en contact avec les responsables des plateformes d'organisation de conférences, de type `ScienceConf` ou `Easychair` pour leur demander de sensibiliser à ces questions les organisateurs de conférences, par le biais d'un message sur la page d'accueil, par exemple. Ce type de décision pourrait être prise lors de la journée PRISME-G, qui outre un lieu de réflexion deviendrait alors un lieu d'action.

Références

Cabanac, G. (2012). Shaping the landscape of research in information systems from the perspective of editorial boards : A scientometric study of 77 leading journals. *Journal of the American Society for Information Science and Technology 63*(5), 977–996.

Cabanac, G., G. Hubert, H. D. Tran, C. Favre, et C. Labbé (2016). Un regard lexico-scientométrique sur le défi EGC 2016. In *16eme Conférence Internationale Francophone sur l'Extraction et la Gestion de Connaissance (EGC 2016)*, Reims, France, pp. 419–424.

Desrosières, A. et E. Didier (2014). *Prouver et gouverner : Une analyse politique des statistiques publiques*. Sciences humaines. La Découverte.

Drucker-Godard, C., T. Fouque, M. Gollety, et A. Le Flanchec (2017). Enseignant-chercheur au féminin : la place des femmes dans les universités. *Recherches en Sciences de Gestion* (118), 125–145.

Larivière, V., C. Ni, Y. Gingras, B. Cronin, et C. Sugimoto (2013). Bibliometrics : Global gender disparities in science. *Nature 504*, 211–3.

Mariani, J., P. Paroubek, G. Francopoulo, et O. Hamon (2014). Rediscovering 15 years of discoveries in language resources and evaluation : The LREC anthology analysis. In *Actes de International Conference on Language Resources and Evaluation*, Reykjavik, Islande.

Moss-Racusin, C. A., J. F. Dovidio, V. L. Brescoll, M. J. Graham, et J. Handelsman (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences 109*(41), 16474–16479.

Névéol, A., K. Fort, et R. Hwa (2017). Report on EMNLP Reviewer Survey. Technical report, Association for Computational Linguistics.

Wenneras, C. et A. Wold (1997). Nepotism and sexism in peer-review. *Nature 387*, 341–343.

8. Voir : `http://itcounts-app.org/#/home`.

Les femmes dans le TAL

Summary

We present here the results of a study that we first led for the "Éthique et TAL" blog on the subject of women in the natural language processing domain (NLP) in France. This study shows that although women represent nearly half of the members (47%) of the French association for computational linguistics, ATALA, and although some efforts were made, they remain under-represented in the most visible and prestigious positions. Besides, the absence of statistics on the subject makes it difficult to investigate. We therefore propose some practical solutions and tools to make the imbalance more visible and solve the issue.