Nouvelles perspectives en sciences sociales
Revue internationale de systémique complexe et d'études relationnelles

Diversité sociale et sémantique : représentation socio-sémantique d'un corpus scientifique, le cas du corpus *ACL Anthology*

Social Diversity and Semantics: Socio-Semantic Representation of a Scientific Corpus, the Case of the *ACL Anthology Corpus*

Elisa Omodei, Yufan Guo, Jean-Philippe Cointet and Thierry Poibeau

Volume 11, Number 1, November 2015

Sur le thème de l'analyse de données textuelles informatisée

URI: https://id.erudit.org/iderudit/1035935ar
DOI: https://doi.org/10.7202/1035935ar

See table of contents

Publisher(s)

Prise de parole

ISSN

1712-8307 (print)
1918-7475 (digital)

Explore this journal

Cite this article

Omodei, E., Guo, Y., Cointet, J.-P. & Poibeau, T. (2015). Diversité sociale et sémantique : représentation socio-sémantique d'un corpus scientifique, le cas du corpus *ACL Anthology*. *Nouvelles perspectives en sciences sociales*, *11*(1), 145–179. https://doi.org/10.7202/1035935ar

Article abstract

We propose a new method to extract multiword expressions from scientific papers. Our approach is made of two major steps: a first list of candidates is extracted based on a score using frequency and specificity information. This list is then filtered based on the status of the term in the abstract of the scientific papers under investigation. These abstracts are annotated using a text zoning analyser. The terms are then classified in different categories according to the text zoning analysis: we make a difference between terms appearing in the method section of the abstract *vs* terms appearing in other zones. This method is applied to the ACL Anthology collection, containing the papers published by the ACL between 1980 and 2008. We show that the technique we use allows us to model interesting facts concerning the evolution of the domain and of the methods used in computational linguistics.

Diversité sociale et sémantique : représentation socio-sémantique d'un corpus scientifique, le cas du corpus *ACL Anthology*

Elisa Omodei[1]
Department of Mathematics and Computer Science
Rovira i Virgili University, Tarragone

Yufan Guo[2]
IBM Research, Almaden, San Jose

Jean-Philippe Cointet
Institut des Systèmes Complexes de Paris-Île de France (ISC-PIF)

Thierry Poibeau
Laboratoire Lattice, UMR 89094, CNRS,
ENS et Université Paris 3 Sorbonne Nouvelle

1. L'analyse automatique de la littérature scientifique

L'analyse des masses de données (en anglais « *big data* ») est un thème de recherche porteur aujourd'hui. Les masses de données permettent en effet de mettre au jour des phénomènes difficilement observables sans méthodes automatiques, et la numérisation de tous les secteurs de la société permet aujourd'hui

1 Ce travail a été effectué alors que le premier auteur travaillait conjointement à l'Institut des Systèmes Complexes de Paris-Ile de France (ISC-PIF) et au laboratoire LATTICE.

2 Ce travail a été effectué alors que l'auteur était au département de Computer Science and Engineering de l'Université de Washington.

d'avoir accès à des données en grandes quantités pour un grand nombre de domaines.

La science est un des domaines qui produisent ainsi de nombreuses données informatisées (littérature scientifique, mais aussi données brutes sous forme de textes, d'images, de chiffres, etc.) et la numérisation des données passées permet aujourd'hui d'avoir accès, pour plusieurs domaines très variées, à des collections d'articles de recherche s'étendant sur plusieurs dizaines d'années. Le monde de la linguistique informatique n'est pas en reste et l'*ACL Anthology*[3] met aujourd'hui à la disposition des chercheurs plus de 24 500 articles au format PDF. Les plus anciens articles datent de 1965 (première édition de COLING), mais ce n'est qu'à partir des années 1980 qu'on commence à avoir des données relativement conséquentes, le volume allant grandissant chaque année depuis lors (il y a donc une très grande disparité dans les volumes de données disponibles suivant les années considérées). Il existe des bases de données similaires pour la biologie et le domaine biomédical (par exemple *Medline*), les systèmes complexes ou la physique (par exemple *APS data set* de l'*American Physical Society*) pour citer quelques bases ayant fait l'objet d'enquêtes diverses.

Ces données ont fait l'objet de nombreux travaux en fonction des propriétés des bases de données bibliographiques privilégiées. Ces bases sont ainsi souvent utilisées pour en extraire des réseaux de collaboration que l'on construit en liant les auteurs selon des liens de co-publication avec l'objectif de comprendre sa topologie[4] ou les processus de morphogenèse sous-jacents[5]. La structure du réseau de références est au cœur du projet « scientométrique » et a généré de très nombreux développements depuis les premiers

3 ACL, soit *Association for Computational Linguistics*.

4 Mark Girvan et Mark E. J. Newman, « Community Structure in Social and Biological Networks », *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, nº 12, 2002, p. 7821-7826.

5 Roger Guimera *et al.*, « Team Assembly Mechanisms Determine Collaboration Network Structure and Team Performance », *Science*, vol. 308, nº 5722, 29 avril 2005, p. 697–702.

travaux sur les réseaux d'inter-citation[6] et de co-citation[7]. Encore bien d'autres dimensions d'analyse sont susceptibles d'être mobilisées pour dresser des cartes de domaines scientifiques: données géographiques associées aux publications, institutions de rattachement des auteurs, ou encore projet structurant le travail de *recherche*. Mais c'est bien le contenu textuel (qu'il provienne des titres, résumés ou mots-clés utilisés par les auteurs pour étiqueter leurs articles) qui a suscité, avec l'analyse des références, le plus grand nombre de travaux depuis les travaux séminaux de Michel Callon[8]. Dans ce travail, nous portons une attention particulière aux dynamiques cognitives, et donc à l'analyse du contenu textuel. Cette analyse est néanmoins couplée aux trajectoires individuelles des chercheurs dans cet espace conceptuel, ce qui nous permet d'interroger, de façon empirique et à grande échelle, les dynamiques d'innovation dans le champ de la linguistique computationnelle.

L'*ACL Anthology* a connu un intérêt particulier en 2012 pour les 50 ans de l'*Association for Computational Linguistics*. Un atelier s'intitulant « *Rediscovering 50 Years of Discoveries* » a été organisé cette année-là[9] : il s'agissait pour l'association de jeter un regard sur l'évolution du domaine depuis 50 ans. Au-delà de ces circonstances particulières, cet événement a été l'occasion d'analyser les données accumulées depuis 50 ans (mais pour les raisons données plus haut, la plupart des études portent sur les articles produits depuis 1980) avec les outils modernes issus à la fois du

6 Eugene Garfield, « Citation Analysis as a Tool in Journal Evaluation », *Science*, vol. 178, n° 4060, 3 novembre 1972, p. 471-479.

7 Henry G. Small, « Co-Citation in the Scientific Literature : A New Measure of the Relationship between Two Documents », *Journal of American Society for Information Science*, vol. 24, n° 4, 1973, p. 265-269.

8 Michel Callon, Jean-Pierre Courtial et Françoise Laville, « Co-Word Analysis as a Tool for Describing the Network of Interaction between Basic and Technological Research : The Case of Polymer Chemistry », *Scientometrics*, vol. 22, n° 1, 1991, p. 155-205; Michel Callon, John Law et Arie Rip, *Mapping the Dynamics of Science and Technology: Sociology of Science in the Real World*, Basingstoke, McMillan, 1986.

9 Rafael E. Banchs (dir.), *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, Jeju Island, Corée, Association for Computational Linguistics, 2012.

traitement des langues et des systèmes complexes, afin d'analyser l'évolution du domaine. L'analyse de ce type de données passe en général par l'extraction d'informations pertinentes (auteurs, mots-clés utilisés, etc.) puis par leur mise en relation : on obtient alors des graphes et les algorithmes développés pour l'analyse des réseaux sociaux peuvent être sollicités. Les relations évoluent au cours du temps : c'est alors à l'algorithmique des graphes évolutifs qu'il faut faire appel. Ces techniques sont mises en œuvre pour répondre à des questions liées à l'histoire des sciences ou, du moins, à l'histoire des différents domaines scientifiques pour lesquels on dispose d'archives conséquentes : on voit donc ici une alliance possible entre le traitement automatique des langues et les systèmes complexes, pour permettre de voir sous un jour nouveau de grandes masses de données qui sont difficiles à analyser sans outils. Les outils permettent de mettre au jour des faits chiffrés et quantifiés, et de vérifier ainsi certaines hypothèses sur l'histoire et l'évolution du domaine, mais aussi sur les techniques utilisées, la mobilité des chercheurs entre différentes thématiques, etc.

L'article « *Towards a Computational History of the ACL: 1980-2008*[10] » est de ce point de vue très riche. Les auteurs essaient de déterminer les grands domaines de recherche au sein du TAL[11] depuis une trentaine d'années. Ils montrent aussi des résultats moins prévisibles, comme l'effet de concentration de la recherche dû aux sources de financement américaines : quand une agence américaine sponsorise des recherches sur un thème donné, celui-ci devient dominant et fédérateur; à l'inverse, pendant les époques avec moins de financement et sans campagne d'évaluation sur un thème privilégié, la communauté est plus dispersée. Ces résultats peuvent sembler logiques mais il est malgré tout remarquable de pouvoir les observer directement, suite à une

10 Ashton Anderson, Dan Jurafsky et Daniel A. McFarland, « Towards a Computational History of the ACL : 1980-2008 », dans P*roceedings of the ACL, op. cit., Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, Jeju, Corée, Association for Computational Linguistics, 2012, p. 13–21.

11 TAL : Traitement Automatique des Langues.

modélisation du domaine : il n'était pas du tout évident que les campagnes d'évaluation américaines aient un effet aussi visible sur un corpus aussi vaste que l'*ACL Anthology*. Ce résultat montre également bien le poids de la recherche américaine dans ce corpus sur la période 1980-1990.

L'étude d'Ashton Anderson et de ses collaborateurs présente des résultats et une méthode d'analyse importants dont nous nous inspirons largement dans cet article. Nous souhaitons pour notre part pouvoir catégoriser automatiquement les mots-clés suivant le type d'information qu'ils véhiculent. Nous proposons donc de combiner l'analyse des termes avec la reconnaissance automatique de la structure argumentative des textes analysés (ce que les anglo-saxons appellent « *argumentative zoning* » ou « *text zoning* »[12]), ce qui permet de typer les termes en fonction du type de phrase dans lequel ils apparaissent.

L'article est organisé comme suit. Nous présentons dans un premier temps la technique d'extraction de termes utilisée. Nous poursuivons avec la présentation de la technique permettant le marquage de la fonction argumentative des phrases (*text zoning*) mise en œuvre pour catégoriser les termes repérés à l'étape précédente. Nous présentons enfin différents résultats de l'application de cette technique au corpus *ACL Anthology* afin d'en faire ressortir certains faits remarquables. Nous concluons cet article par un résumé et quelques perspectives.

2. Annotation de la structure argumentative

La reconnaissance et l'annotation de la structure discursive des articles scientifiques sont devenues des enjeux importants pour la communauté du traitement des langues. Ces techniques peuvent en effet permettre de savoir si une section d'un article scientifique donné concerne, par exemple, le protocole expérimental employé, les données d'expériences ou la discussion et la comparaison avec les travaux antérieurs. Ce type d'analyse donne des résultats de plus en plus précis et commence à intéresser les

12 Simone Teufel, *Argumentative Zoning : Information Extraction from Scientific Articles*, thèse de doctorat, University of Edinburgh, 1999.

grandes maisons d'éditions scientifiques, dans la mesure où on peut ainsi enrichir les bases de connaissances existantes et proposer de nouveaux parcours de lecture.

2.1. État de l'art

Les premiers travaux d'importance dans le domaine sont certainement ceux de Simone Teufel qui a proposé de catégoriser les phrases d'articles scientifiques sur le traitement automatique des langues suivant sept étiquettes différentes : BKG (arrière-plan scientifique), OTH (description neutre de travaux antérieurs), OWN (description neutre du travail de l'auteur), AIM (objectifs de l'article), TXT (annonce de l'organisation de l'article), CTR (comparaison avec des travaux antérieurs) et BAS (description des travaux antérieurs sur lesquels s'appuie l'article)[13].

La tâche est appelée « *rhetorical zoning* » ou « *argumentative zoning* » par l'auteur, dans la mesure où le balisage doit permettre d'identifier la fonction rhétorique ou argumentative de chaque phrase du texte.

Le travail initial de Teufel[14] est fondé sur l'annotation manuelle de 200 articles représentatifs du domaine, issus des conférences de l'*ACL* et de la revue *Computational Linguistics*. Un classifieur est ensuite mis au point par apprentissage automatique, c'est-à-dire que des règles sont inférées automatiquement à partir du corpus annoté manuellement (ce qu'on appelle le « corpus d'apprentissage »). Le classifieur exploitant cet ensemble de règles est capable d'annoter automatiquement de nouveaux textes sans qu'il soit nécessaire d'avoir recours à d'autres traitements manuels. Teufel rapporte que le système automatique donne le bon résultat dans 70 % des cas (le taux d'accord est de 88 % quand ce sont des humains qui procèdent à une annotation similaire; autrement dit, une part de la tâche est subjective et un taux de 100 % est impossible à obtenir : de ce point de vue, 70 % peut être vu comme un bon résultat permettant une mise en œuvre opérationnelle malgré le taux d'erreur). Le classifieur repose sur un

13 *Ibid.*
14 *Ibid.*

modèle bayésien naïf car les méthodes plus sophistiquées testées par l'auteur ne semblent pas permettre d'obtenir de meilleurs résultats.

Teufel montre dans une publication ultérieure[15] comment cette technique peut être utilisée pour générer des résumés automatiques de qualité. Les techniques de résumé traditionnelles sont fondées sur la sélection de phrases en fonction de leur intérêt informatif supposé, essentiellement sur la base des noms et des verbes qui la composent (les mots les plus centraux, souvent appelés centroïdes[16], ce qui pose un problème pour générer des textes tenant compte de la variété du texte de départ). Le repérage de la structure argumentative répond partiellement à ce problème dans la mesure où il est dès lors possible de générer des résumés reflétant les différentes zones repérées ou, au contraire, privilégiant une zone donnée suivant les besoins informationnels du lecteur.

Teufel a enfin montré comment le marquage argumentatif peut être couplé avec les références scientifiques. Les articles scientifiques sont en effet fondées sur des citations des travaux antérieurs mais ces citations peuvent avoir différentes finalités : simple mention de travaux antérieurs donnant l'arrière-plan de la recherche en cours, travaux précis auxquels s'oppose la publication en cours, référence à des travaux utilisant le même protocole expérimental, etc. Coupler repérage de référence et balisage argumentatif permet de typer les citations, toujours dans le but de faciliter la lecture en fonction des besoins informationnels du lecteur[17].

Les travaux de Teufel ont depuis donné lieu à différents types de travaux, d'une part pour affiner la méthode d'annotation,

---

15 Simone Teufel et Mark Moens, « Summarizing Scientific Articles – Experiments with Relevance and Rhetorical Status », *Computational Linguistics*, vol. 4, n° 28, 2002, p. 409-445.

16 Dragomir Radev *et al.*, « Centroid-Based Summarization of Multiple Documents », *Journal on Information Processing Management*, vol. 40, n° 6, 2004, p. 919-938.

17 Simone Teufel, Advaith Siddharthan et Dan Tidhar, « Automatic Classification of Citation Function », *Proceedings of Empirical Methods in Natural language Processing* (EMNLP), Association for Computational Linguistics, 2006, p. 103-110.

d'autre part pour vérifier son applicabilité à différents domaines scientifiques. Pour le premier point, les recherches ont porté sur les traits pertinents pour la classification, l'évaluation de différents algorithmes pour la tâche et surtout la diminution de la quantité de textes à annoter pour obtenir un système fonctionnel. Pour le second, ce sont surtout les domaines de la biomédecine et de la biologie qui ont montré le plus d'intérêt pour ce type de techniques, du fait de la quantité d'articles disponible dans ce domaine et de la nécessité d'accéder de manière transversale à cette littérature (les biologistes peuvent par exemple avoir besoin d'accéder à tous les protocoles expérimentaux pour un problème donné)[18].

Les travaux de Yufan Guo et de ses collaborateurs reprennent l'analyse de la structure argumentative en complétant les travaux initiaux de Teufel sur un certain nombre de points : recours à une vaste liste de critères pour déterminer la classification des phrases, évaluation de plusieurs algorithmes d'apprentissage et diminution de la quantité de données annotées à fournir au système pour l'entraînement. Ces travaux proposent en particulier d'avoir recours à l'apprentissage actif (*active learning*) pour entraîner leur système. On sait en effet que l'« apprentissage actif » permet de réduire la quantité de données annotées en utilisant parallèlement une grande masse de données non annotées : l'analyseur peut par exemple s'assurer qu'une configuration linguistique particulière repérée dans le corpus d'apprentissage (c'est-à-dire le petit jeu de données annoté manuellement) est bien productive, en vérifiant qu'elle est largement attestée dans les données non annotées, ce qui permet de limiter le travail manuel tout en garantissant la représentativité des règles inférés[19].

---

18 Yoko Mizuta *et al.*, « Zone Analysis in Biology Articles as a Basis for Information Extraction », *International Journal of Medical Informatics*, vol. 75, n° 6, , 2006, p. 468-487; Imad Tbahriti *et al.*, « Using Argumentation to Retrieve Articles with Similar Citations : An Inquiry into Improving Related Articles Search in the Medline Digital Library », *International Journal of Medical Informatics*, vol. 75, n° 6, 2006, p. 488-495.

19 Yufan Guo, Anna Korhonen et Thierry Poibeau, « A Weakly-Supervised Approach to Argumentative Zoning of Scientific Documents », *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Edimbourg, 2011,

Cette méthode est bien indiquée dans notre cas dans la mesure où les corpus à analyser en traitement des langues (et particulièrement l'*ACL Anthology*) sont souvent d'assez grande taille quoiqu'ils ne sont évidemment pas annotés. Les traits utilisés pour l'apprentissage sont de trois types : i) positionnels (localisation de la phrase au sein de l'article), ii) lexicaux (mots, classes de mots, bigrammes, etc. sont pris en considération) et iii) syntaxiques (les différentes relations syntaxiques, ainsi que les classes de noms en position sujet et les classes de noms en position objet sont pris en considération). L'analyse est donc considérablement plus riche que celle de Teufel mais nécessite en contrepartie un analyseur syntaxique.

2.2. Application de l'analyse argumentative au corpus de l'ACL

La méthode développée par Guo et ses collègues semble particulièrement bien adaptée à notre problème. Nous souhaitons en effet catégoriser les termes repérés à l'étape précédente afin, notamment, d'identifier les méthodes mentionnées dans le corpus *ACL Anthology* et pouvoir ainsi analyser, par exemple, leur évolution dans le temps. Les termes apparaissant dans des phrases se rapportant au protocole expérimental employé sont donc susceptibles de particulièrement nous intéresser. Il faut noter à ce propos qu'il n'y a pas de frontière étanche entre thèmes et méthodes de recherche dans la mesure où le traitement automatique des langues s'appuie sur ses propres résultats pour concevoir des systèmes en couches empilées : ainsi, un analyseur sémantique reposera fréquemment sur un analyseur syntaxique employé comme outil (et apparaissant donc dans la section méthodologique de l'article).

L'annotation ne porte que sur les résumés des articles. Nous faisons en effet l'hypothèse que les résumés contiennent assez d'information et sont assez redondants pour observer l'évolution du domaine. À l'inverse, aborder l'étude en utilisant le texte

p. 273-283; Yufan Guo, Roi Reichart et Anna Korhonen , « Improved Information Structure Analysis of Scientific Documents through Discourse and Lexical Constraints », *Proceedings of Human Language Technologies : Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*, 2013, p. 928-937.

complet des articles entraînerait probablement du bruit et complexifierait inutilement les traitements.

Le jeu d'annotation initialement adopté comporte six catégories différentes et une catégorie AUTRE pour les phrases ne pouvant pas être catégorisées par les étiquettes définies (soit 7 étiquettes en tout). Ces étiquettes sont les suivantes :

- OBJECTIF : décrit les objectifs de l'article;
- MÉTHODE : méthodes employées par l'article;
- RÉSULTATS : résultats obtenus;
- CONCLUSION : conclusion de l'article;
- ARRIÈRE-PLAN : contexte scientifique;
- TRAVAUX LIÈS : positionnement par rapport à des travaux directement liés à ceux présentés;
- AUTRES TRAVAUX : positionnement par rapport à d'autres travaux.

Ces catégories sont reprises de divers travaux précédents[20]. Il nous a semblé important de reprendre un jeu de catégories existantes dans la mesure où ces catégories, avec de légères variations, se sont globalement imposées depuis les premiers travaux de Teufel. Certaines catégories sont malgré tout peu présentes dans les résumés de l'*ACL Anthology*, et finalement quatre catégories transparaissent principalement : les catégories OBJECTIF, ARRIÈRE-PLAN, RÉSULTATS et MÉTHODE. Il est rare de trouver des comparaisons avec d'autres travaux dans les résumés de l'ACL Anthology (alors qu'on en trouve fréquemment dans les résumés en biologie par exemple).

Une centaine de résumés d'article issus de l'*ACL Anthology* ont ensuite été annotés manuellement avec ces catégories (environ 500 phrases, les résumés de l'*ACL Anthology* étant souvent très courts dans la mesure où il s'agit en grande majorité de résumés d'articles de conférence). Les articles annotés ont été choisis aléatoirement, en s'assurant, toutefois, qu'ils couvrent différentes périodes et qu'ils contiennent des termes variés. L'annotation a

20 Notamment les travaux de Mizuta (*op. cit.*) et de Guo (*op. cit.*).

été faite en suivant le guide d'annotation mis au point par Guo, notamment en ce qui concerne les phrases complexes, se rapportant potentiellement à plus d'une catégorie définie (un jeu de préférences est défini pour résoudre ces cas difficiles).

L'algorithme décrit par Guo et ses collaborateurs[21] est ensuite repris et adapté à notre cas de figure. L'analyse se fonde en particulier sur les traits positionnels, lexicaux et syntaxiques comme cela a été expliqué dans la section précédente. En revanche aucune information spécifique au domaine n'est ajoutée, ce qui rend le processus plus simple à modéliser et à reproduire. Pour l'analyse syntaxique, l'analyseur *C&C* est utilisé[22] et pour la classification, nous avons recours à l'implémentation des SVM linéaires de *Weka*[23]. Comme résultat, pour chaque phrase du corpus, l'algorithme associe une étiquette choisie parmi les étiquettes possibles.

2.3. Résultats et interprétation

Pour valider les résultats obtenus, un ensemble de résumés est choisi aléatoirement. Les quatre catégories principales sont bien représentées mais inégalement réparties : 18,05 % des phrases sont catégorisées comme ARRIÈRE-PLAN, 14,35 % comme OBJECTIF, 14,81 % comme RÉSULTAT et 52,77 % comme MÉTHODE. On voit bien, à la lecture de ces chiffres, l'importance de la dimension méthodologique dans le domaine.

On observe ensuite, pour chaque catégorie possible, le pourcentage de phrases correspondant effectivement à cette étiquette, ce qui permet de mesurer les performances du système en termes

---

21 Yufan Guo, Anna Korhonen et Thierry Poibeau, *op. cit.*

22 James Curran, Stephen Clark et Johan Bos, «Linguistically Motivated Large-Scale NLP with C&C and Boxer», *Proceedings of the 45th Meeting of the Association for Computation Linguistics (ACL)*, 2007, p. 33-36. C&C est un analyseur syntaxique à large couverture ayant obtenu de bonnes performances pour l'anglais.

23 Weka est une « boîte à outils » intégrant différents algorithmes d'apprentissage très utilisée en informatique; l'algorithme SVM, pour « *Support Vector Machines* », est un des algorithmes de classification les plus efficaces pour le traitement automatique des langues.

de précision. Les résultats obtenus sont présentés dans le tableau suivant.

Tableau 1

Résultat de l'analyse argumentative (en précision)

Catégorie	Précision
Objectif	83,87 %
Arrière-plan	81,25 %
Méthode	71,05 %
Résultats	82,05 %

Ces résultats sont conformes à l'état de l'art (si on les compare avec ceux de Guo et de ses collègues[24] par exemple). On voit que les résultats sont globalement satisfaisants, particulièrement en regard du peu de phrases annotées pour l'entraînement (c'est-à-dire que le « corpus d'entraînement » annoté manuellement est très petit). La richesse des traits pris en compte et la stratégie d'apprentissage actif permettent en outre d'avoir des résultats portables d'un domaine à l'autre sans tâche d'annotation lourde. Les résultats sont légèrement moins bons pour la catégorie MÉTHODE car celle-ci est sans doute plus diversifiée que les autres et donc moins facile à cerner.

L'exemple 1 est un texte annoté suite à l'analyse du système (il s'agit d'un article de de Gary Geunbae Lee *et al.*[25]), choisi au hasard parmi ceux qui présentent une bonne diversité dans les catégories utilisées). La catégorisation s'effectue au niveau des

24 Yufan Guo, Anna Korhonen et Thierry Poibeau, *op. cit.*

25 Gary Geunbae Lee, Jeongwon Cha et Jong-Hyeok Lee, « Syllable-Pattern-Based Unknown-Morpheme Segmentation and Estimation for Hybrid Part-of-speech Tagging of Korean », *Computational Linguistics*, vol. 28, n° 1, 2002, p. 53-70.

phrases, ce qui n'est pas sans poser de problèmes : par exemple, dans ce résumé, le fait qu'une méthode hybride est utilisée est indiqué dans une phrase étiquetée OBJECTIF par le système. Les phrases marquées MÉTHODE contiennent toutefois des mots-clés précieux, comme « *lexical pattern* » ou « *tri-gram estimation* », ce qui peut permettre d'inférer le fait qu'il s'agit d'un système hybride. On aperçoit au passage des problèmes de numérisation, qui sont typiques du corpus étudié : l'*ACL Anthology* comprend des textes convertis automatiquement à partir de fichiers PDF de conférences passées, ce qui entraîne parfois des problèmes de qualité.

Exemple 1

Un résumé annoté avec l'analyseur de la structure argumentative. Les catégories ajoutées au texte sont indiquées en gras.

Most of errors in Korean morphological analysis and POS (Part-of-Speech) tagging are caused by unknown morphemes . **ARRIERE-PLAN**
This paper presents a generalized unknown morpheme handling method with POSTAG (POStech TAGger) which is a statistical/rule based hybrid POS tagging system . **OBJECTIF**
The generalized unknown morpheme guessing is based on a combination of a morpheme pattern dictionary which encodes general lexical patterns of Korean morphemes with a posteriori syllable tri-gram estimation. **METHODE**
The syllable tri-grams help to calculate lexical probabilities of the unknown morphemes and are utilized to search the best tagging result. **METHODE**
In our scheme, we can guess the POS's of unknown morphemes regardless of their numbers and positions in an eojeol, which was not possible before in Korean tagging systems. **RESULTATS**
In a series of experiments using three different domain corpora , we can achieve 97% tagging accuracy regardless of many unknown morphemes in test corpora . **RESULTATS**

3. Application : contribution à l'étude de l'évolution du traitement automatique des langues d'après l'*ACL Anthology*

Comme nous l'avons dit dans l'introduction, nous nous situons dans la lignée de travaux liés aux 50 ans de l'*ACL*[26]. L'*ACL Anthology* est utilisé ici comme un cas d'étude typique : le corpus s'étendant sur une période de temps conséquente (plus de 30 ans si on retient les articles depuis 1980), il peut être intéressant d'en étudier les grandes évolutions.

26 Ashton Anderson, Dan Jurafsky et Daniel A. McFarland, *op. cit.*

3.1. Repérage des termes spécifiques

Il nous a semblé particulièrement intéressant de nous pencher sur l'évolution des méthodes employées en traitement automatique des langues. Pour cela, il est nécessaire d'identifier les termes particulièrement présents dans les phrases étiquetées MÉTHODE.

L'extraction terminologique vise à identifier de façon automatique les termes pertinents dans un corpus en utilisant des méthodes de traitement automatique des langues en vue de proposer une modélisation conceptuelle d'un domaine. L'approche classique pour extraire des termes d'un corpus peut être décomposée en deux parties. Dans une première phase des outils d'analyse linguistique sont utilisés pour construire une liste de candidats possibles qui sont filtrés dans une seconde phase.

La construction des termes candidats consiste classiquement[27] à appliquer au texte un étiqueteur morphosyntaxique (« *POS-tagging* ») puis à utiliser les informations grammaticales associées à chaque mot pour effectuer une analyse syntaxique de surface (appelée « *chunking* » en anglais) qui permette d'identifier les groupes nominaux dans le texte, les multi-termes extraits (parfois également appelés n-grammes) constituant ainsi des candidats grammaticalement valides. Dans une deuxième phase, les termes sont filtrés, soit en faisant appel à des ressources extérieures, soit en fonction de scores associés tels que leur fréquence ou leur spécificité[28].

Dans cette étude, nous avons utilisé une stratégie minimale d'extraction terminologique. Nous nous sommes attachés aux contenus présents dans les résumés du corpus. Pratiquement, nous avons utilisé le module *NLTK* (*Natural Language ToolKit*) de traitement automatique des langues : une fois une liste de

27 Didier Bourigault et Christian Jacquemin, « Term Extraction + Term Clustering: An Integrated Platform for Computer-Aided Terminology », *Proceedings of the Ninth Conference on European 78 Chapter of the Association for Computational Linguistics*, EACL, 1999, pages 15–22.

28 Katarina Frantzi, Sophia Ananiadou et Hideki Mima, « Automatic Recognition of Multi-Word Terms: The C-value/NC-value method », *International Journal on Digital Libraries*, Springer-Verlag, vol. 3, n° 2, 2000, p. 115-130.

multi-termes obtenue, nous avons simplement sélectionné l'ensemble des termes apparaissant dans au moins 10 articles différents.

Nous calculons ensuite la spécificité de chaque terme par rapport aux catégories définies pour l'analyse discursive. La spécificité est calculée grâce au test de Kolmogorov-Smirnov, qui quantifie une distance entre les fonctions de répartition empiriques de deux échantillons :

$$D = \max_{x} \left| S_{N_1}(x) - S_{N_2}(x) \right|$$

où $S_{N_1}(x)$ et $S_{N_2}(x)$ sont les fonctions de répartition empiriques des deux échantillons (ce qui correspond, dans notre cas, au nombre d'occurrences du terme dans chaque zone identifiée par le marquage argumentatif et au nombre total d'occurrences (en considérant tous les termes) dans chaque zone)[29]. Une valeur importante de *D* pour un terme donné signifie donc que le terme est très spécifique d'une zone. À l'inverse, une valeur faible indique que le terme est éparpillé dans toutes les zones et est donc peu spécifique.

La liste est ensuite triée par mesure de spécificité et les cent premiers termes sont ensuite catégorisés par un expert du domaine. On obtient ainsi le tableau 2 : celui-ci ne contient pas tous les éléments *a priori* pertinents (c'est-à-dire toutes les méthodes utilisées en traitement automatique des langues) mais il contient les termes les plus spécifiques ainsi qu'ils sont calculés par la méthode précédente. Il ne faut donc pas s'étonner de trouver une liste incomplète par rapport à l'ensemble des méthodes utilisées dans le domaine.

29 William H. Press *et al.*, *Numerical Recipes: The Art of Scientific Computing*, 3e édition, New York, Cambridge University Press, 2007.

Tableau 2

Classement des termes les plus spécifiques trouvés dans les phrases étiquetées METHODE

Méthodes		
Catégorie	Méthode	N-grammes
Machine learning	Bayesian methods	Baesyan
	Vector Space model	space model, vector space, cosine
	Genetic algorithms	genetic algorithms
	HMM	hidden markov models, markov model
	CRF	conditional random fields
	SVM	support vector machines
	MaxEnt	maximum entropy model, maximum entropy approach, maximum entropy
	Clustering	clustering algorithm, clustering method, word clusters, classification problem
Speech & Mach. Trans.	Language models	large-vocabulary, n-gram language model, Viterbi
	Parallel Corpora	parallel corpus, bilingual corpus, phrase pairs, source and target languages, sentence pairs, word pairs, source sentence
	Alignment	phrase alignment, alignment algorithm, alignment models, ibm model, phrase translation, translation candidates, sentence alignment
NLP Methods	POS tagging	part-of-speech tagger, part-of-speech tags
	Morphology	Morphology two-level morphology, morphological analyzer, morphological rules
	FST	finite-state transducers, regular expressions, state automata, rule-based approach
	Syntax	syntactic categories, syntactic patterns, extraction patterns
	Dependency parsing	dependency parser, dependency graphs, prague dependency, dependency treebank, derivation trees, parse trees
	Parsing	grammar rules, parser output, parsing process, parsed sentences, transfer rules
	Semantics	logical forms, inference rules, generative lexicon, lexical rules, lexico-syntactic, predicate argument
Applications	IE and IR	entity recognition, answer candidates, temporal information, web search, query expansion, google, user queries, keywords, query terms, term recognition
	Discourse	generation component, dialogue acts, centering theory, lexical chains, resolution algorithm, generation process, discourse model, lexical choice
	Segmentation	machine transliteration, phonological rules, segmentation algorithm, word boundaries
Words and Resource	Lexical knowledge bases	lexical knowledge base, semantic network, machine readable dictionaries, eurowordnet, lexical entries, dictionary entries, lexical units, representation structures, lookup
	Word similarity	word associations, mutual information, semantic relationships, word similarity, semantic similarity, semeval-2007, word co-occurrence, synonymy
	Corpora	brown corpus, dialogue corpus, annotation scheme, tagged corpus
Evaluation	Evaluation	score, gold standard, evaluation measures, estimation method
Calculation & complexity	Software	tool development, polynomial time, software tools, series of experiments, system architecture, runtime, programming

3.2. Évolution des méthodes dans le temps

L'analyse automatisée du corpus permet avant tout de tracer l'évolution des différentes tendances dans le temps. Pendant la période considérée, les méthodes utilisées ont beaucoup changé, le principal fait marquant étant peut-être le recours massif à l'apprentissage depuis la fin des années 1990. Cette tendance est marquée par un recours quasi systématique, dans les articles actuels, à des expérimentations donnant lieu à des résultats chiffrés.

Pour confirmer de façon quantitative cette hypothèse, nous nous intéressons à l'évolution dans le temps de la proportion de phrases étiquetées RÉSULTAT. Sur la figure 1, nous pouvons ainsi observer que la courbe correspondante croît de façon quasi linéaire du début des années 1980 jusqu'à la fin des années 2000. La figure 2 montre un résultat relativement similaire pour l'évaluation, qui s'est généralisée à la même période. Seulement, comme le marquage discursif n'inclut pas de catégorie spécifique pour l'évaluation, il est nécessaire de repérer ce type d'information au sein des phrases marquées MÉTHODE (la façon d'évaluer étant généralement détaillée avec les méthodes, comme en témoignent les termes trouvés (voir le tableau 2).

Figure 1

Évolution dans le temps de la proportion de phrases catégorisées par l'outil d'analyse discursive comme étant des phrases concernant des résultats (par rapport au nombre total des phrases contenues dans les articles publiés dans l'année correspondante)

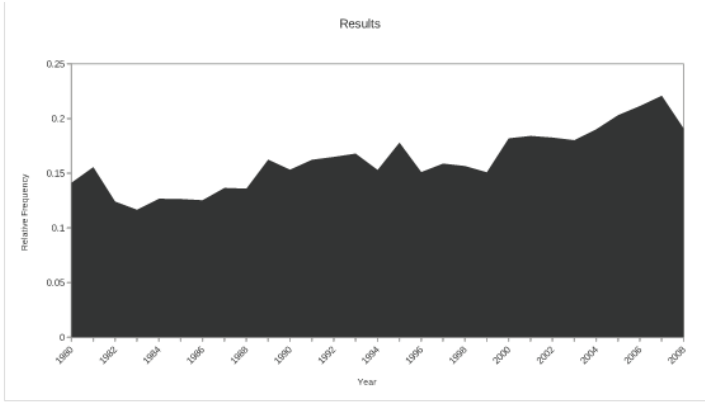
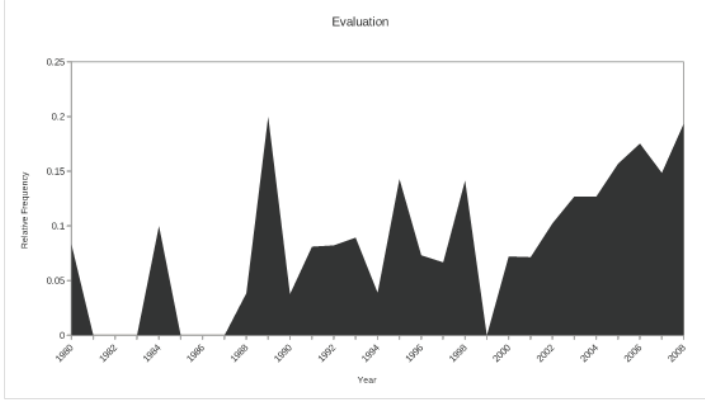


Figure 2

Nombre d'articles du corpus dans lesquels au moins un des termes du groupe Evaluation apparaît dans une zone étiquetée METHODE (score normalisé par le nombre total d'articles publiés dans l'année)



Il est ensuite possible de faire des traitements plus fins pour suivre dans le temps l'évolution des différents groupes de méthodes identifiées. Les résultats apparaissent sur les figures 3a à 3f.

Les méthodes à base de règles et de ressources linguistiques élaborées manuellement se maintiennent ou baissent légèrement, tandis que les méthodes à base d'apprentissage connaissent un succès de plus en plus grand à partir des années 1990. Ceci n'est pas en soi surprenant : on sait que des systèmes à base de règles continuent d'être utilisés tandis que l'apprentissage s'est généralisé. La figure indique toutefois un constat plus équilibré qu'on ne pourrait le penser : les deux types de méthodes coexistent. Les méthodes d'apprentissage sont probablement souvent employées en collaboration avec des méthodes fondées sur l'apprentissage et les deux paradigmes se complètent sans doute plus qu'ils ne s'opposent.

Le détail montre des tendances qu'il faudrait confirmer par une étude plus approfondie. On voit toutefois le succès de l'analyse en dépendance à la fin des années 1980 (probablement grâce au succès des grammaires d'arbres adjoints à cette époque) puis, à nouveau, un certain succès depuis les années 2000 grâce au développement des techniques d'apprentissage et des corpus étiquetés en dépendances (ce qui a par exemple donné lieu à plusieurs tâches partagées *(« shared tasks »)* lors des conférences *CONLL* de 2006 à 2009[30]).

Les méthodes d'apprentissage se succèdent par vagues mais chaque méthode continue par la suite d'être employée, perfectionnée et appliquée à de nouvelles tâches. Les HMM *(Hidden Markov Model)* et les n-grammes connaissent un pic très net dans les années 1990, probablement suite aux expériences initiales de Frederick Jelinek et ses collègues inaugurant l'ère de la traduction automatique statistique[31]. Les SVM (*Support Vector Machines*) et les CRF (*Conditional Random Fields*) ont eu un succès plus récent, comme on le sait.

---

30 *CONLL : Conference on Computational Natural Language Learning.*

31 Peter F. Brown *et al.*, « A Statistical Approach to Machine Translation », *Computational Linguistics*, vol. 16, n° 2, 1990, p. 79-85.

Figure 3a.

Évolution dans le temps de la fréquence relative des différents éléments étudiés : les sous-domaines scientifiques

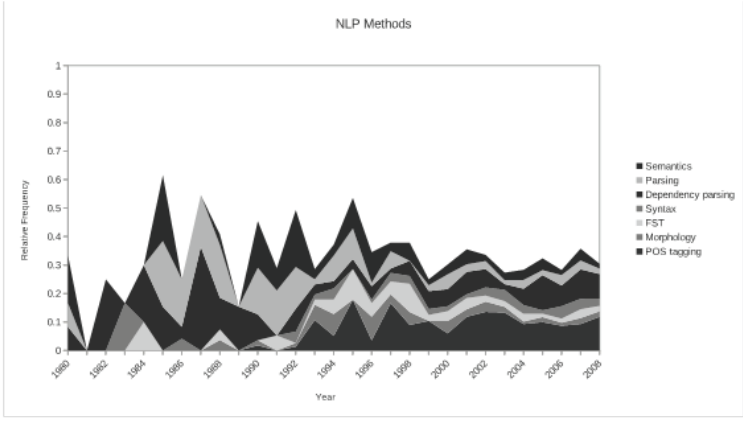


Figure 3b

Évolution dans le temps de la fréquence relative des différents éléments étudiés : les applications

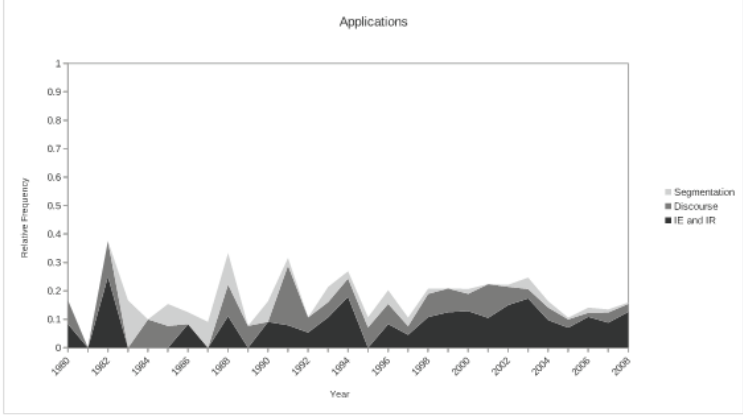


Figure 3c

Évolution dans le temps de la fréquence relative des différents éléments étudiés : les approches par apprentissage

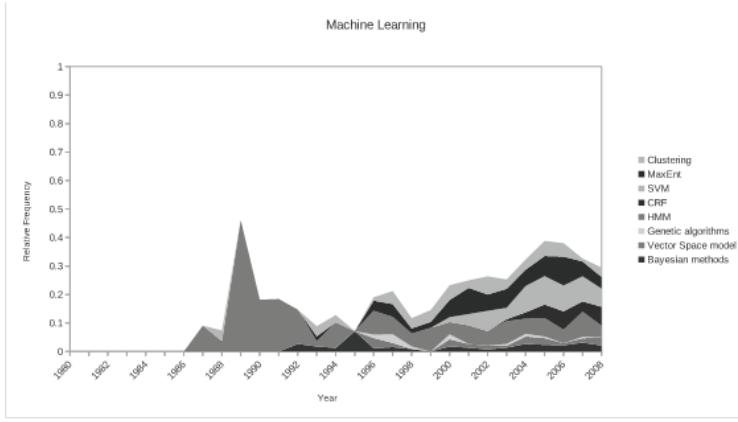


Figure 3d

Évolution dans le temps de la fréquence relative des différents éléments étudiés : les corpus parallèles

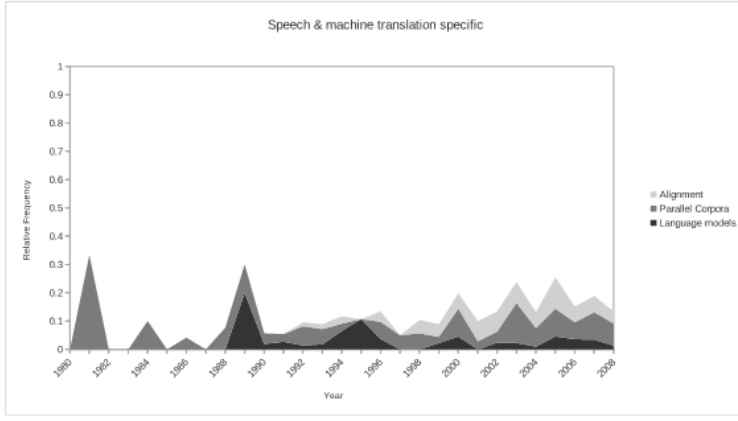


Figure 3e

Évolution dans le temps de la fréquence relative des différents éléments étudiés : les ressources

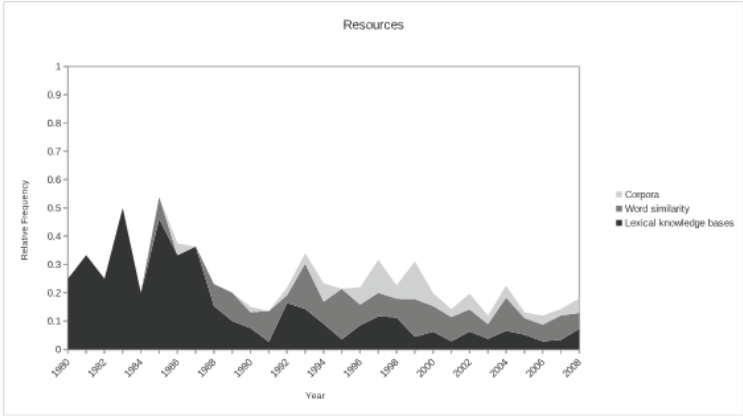
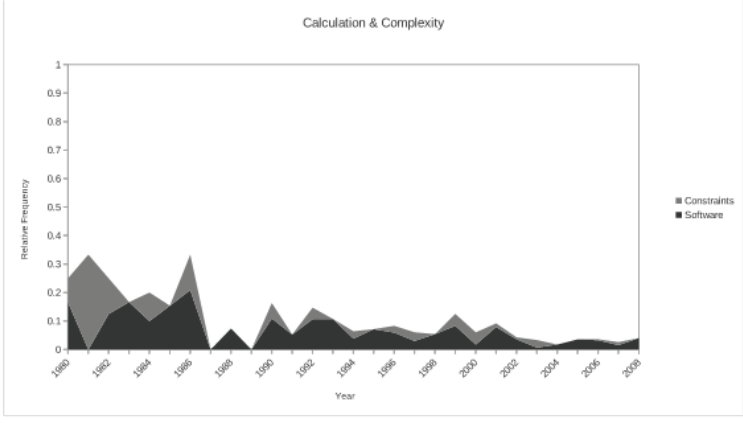


Figure 3f

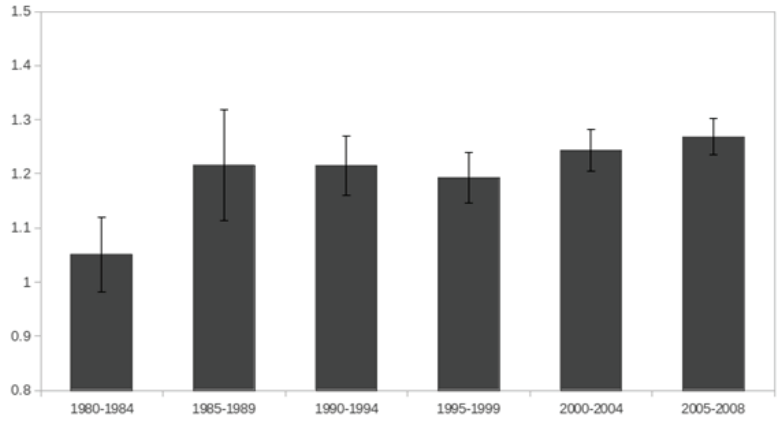
Évolution dans le temps de la fréquence relative des différents éléments étudiés : contraintes et aspects logiciels



Nous nous sommes aussi intéressés à la distribution des méthodes entre les articles et entre les auteurs. La figure 4 montre le nombre moyen de termes apparaissant dans la section MÉTHODE du résumé des articles au cours du temps. On peut observer que le nombre d'éléments méthodologiques augmente, surtout dans les années 1980, montrant peut-être un accroissement de la complexité des systèmes développés.

Figure 4

Évolution du nombre de méthodes par article dans le temps



3.3. La dynamique des auteurs dans l'espace des méthodes

En un sens, on peut dire que ce qui a été observé dans la partie précédente n'est pas entièrement nouveau : nos résultats viennent en quelque sorte plutôt confirmer des faits déjà connus.

La méthode proposée peut toutefois permettre d'aller plus loin : on peut essayer d'observer les dynamiques à l'œuvre dans l'évolution du domaine. Comment les nouvelles méthodes d'analyse sont-elles introduites dans le domaine? Sont-elles plutôt amenées par des chercheurs débutant dans le domaine ou sont-ce plutôt des chercheurs confirmés du domaine qui inventent ou vont chercher dans des domaines connexes de nouvelles méthodes? Les spécialistes du TAL sont-ils en général spécialistes d'une

méthode ou d'un domaine étroit de spécialité ou ont-ils plutôt une expertise large et diversifiée?

Il s'agit évidemment de questions complexes et chaque individu a une trajectoire particulière. Les méthodes automatiques peuvent toutefois donner des indicateurs, surtout dans la durée. Comme nous l'avons déjà vu, Anderson et ses collègues[32] montrent ainsi que les conférences d'évaluation ont eu un impact sur le domaine, en limitant la diversité des recherches à certaines périodes clés, ce qui ne veut pas dire qu'il n'y avait pas à ces époques aussi des recherches originales en dehors de ces campagnes. Il s'agit donc d'essayer de mettre au jour certaines tendances spécifiques d'un domaine scientifique, qui pourraient par exemple amener à des comparaisons avec d'autres domaines scientifiques. Les outils fournissent avant tout des hypothèses : ils poussent le chercheur à aller voir plus loin mais il ne s'agit évidemment que d'outils d'aide à l'analyse. Nous ne prétendons pas donner une vue exacte et absolument objective du domaine.

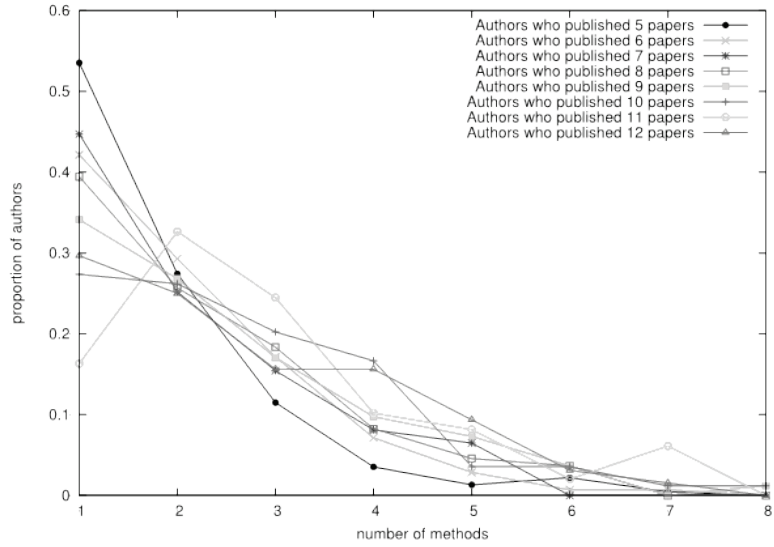
Pour mener à bien notre enquête, nous ne prenons en compte que les auteurs qui ont produit au moins cinq articles dans l'*ACL Anthology*, afin de ne prendre en compte que les auteurs ayant contribué au domaine pendant un temps assez long pour la pertinence de l'étude.

La figure 5 montre le nombre d'auteurs spécialistes d'une ou plusieurs méthodes données. On constate que la plupart des auteurs font référence à une seule méthode. Logiquement, les courbes sont décroissantes : il y a finalement peu d'auteurs utilisant une très large gamme de méthodes différentes. Ces résultats mériteraient évidemment d'être confirmés par une étude de plus grande ampleur prenant en compte une plus grande diversité de termes regroupés par famille. Il nous semble malgré tout que cette expérimentation montre des tendances intéressantes pour ce corpus.

32 Ashton Anderson, Dan Jurafsky, Daniel A. McFarland, *op. cit.*

Figure 5

Proportion d'auteurs experts d'un nombre de méthodes donné, pour différentes catégories de chercheurs.



Nous nous concentrons ensuite sur les « pionniers », que nous définissons comme étant les premiers auteurs ayant publié un article où le terme référant à une méthode donnée apparaît (par exemple, les premiers articles où le terme « *support vector machine* », ou SVM, apparaît). Parmi l'ensemble des articles mentionnant une méthode, seuls les articles correspondant aux 16 premiers centiles (autrement, les 16 % d'articles publiés en premier) sont considérés comme pionniers : cette valeur a été choisie en se fondant sur les travaux d'Everett M. Rogers sur la diffusion des innovations[33], qui montrent l'importance du rôle joué par les innovateurs (qui constituent le premier 2,5 %) et des adopteurs précoces (qui constituent les 13,5 % suivants). Ces deux populations ensemble peuvent être considérées comme formant l'ensemble des pionniers.

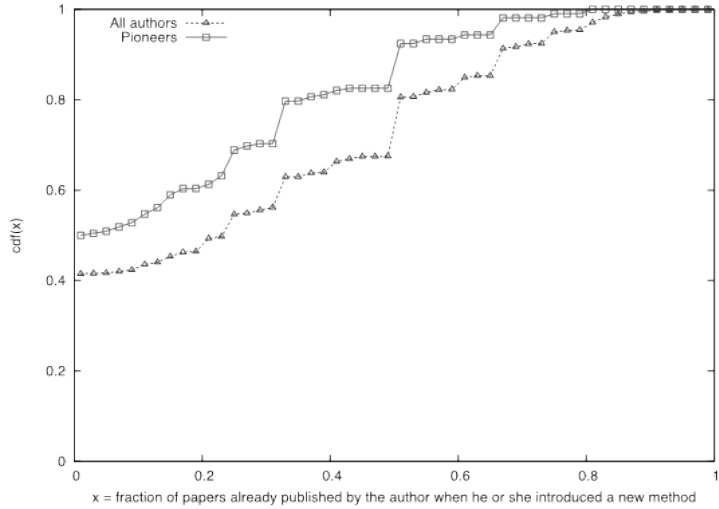
33 Everett M. Rogers, *Diffusion of Innovations*, New York, Simon and Schuster, 1962.

Nous essayons de déterminer à quel moment de leur carrière les chercheurs utilisent des méthodes novatrices. Pratiquement, nous examinons à quelle étape de leur carrière les auteurs que nous avons considérés comme « pionniers » ont publié les articles ayant permis de les classer ainsi (par exemple, si un auteur est un des premiers à avoir utilisé les SVM, l'a-t-il fait lors de ses premières publications ou plus tard au cours de sa carrière?). Le résultat est visible sur la figure 6, où on compare la fraction d'articles publiés par les « pionniers » avant d'introduire une nouvelle méthode (par rapport à leur production totale), et le même type de données pour les autres chercheurs (c'est-à-dire la fraction d'articles publiés avant de commencer à utiliser une méthode nouvelle pour eux mais pas pour le domaine). Nous observons que 50 % des « pionniers » n'avaient jamais publié dans le domaine avant d'introduire la nouvelle méthode en question (contre 40 % seulement en ce qui concerne les autres chercheurs). Ces valeurs montrent que les nouvelles méthodes semblent émaner assez largement de nouveaux venus, probablement de chercheurs ayant déjà éprouvé la méthode sur un autre domaine (de fait, l'équipe de Jelinek, qui a joué un rôle essentiel dans l'essor des chaînes de Markov cachées à partir des années 1990[34], avait surtout été active en reconnaissance de la parole jusque-là et n'avait quasiment pas publié d'articles faisant partie du corpus *ACL*, même s'il s'agissait bien évidemment de chercheurs confirmés).

34 Peter F. Brown *et al.*, « A Statistical Approach to Machine Translation », *Computational Linguistics*, vol. 16, n° 2, juin 1990, p. 79-85.

Figure 6

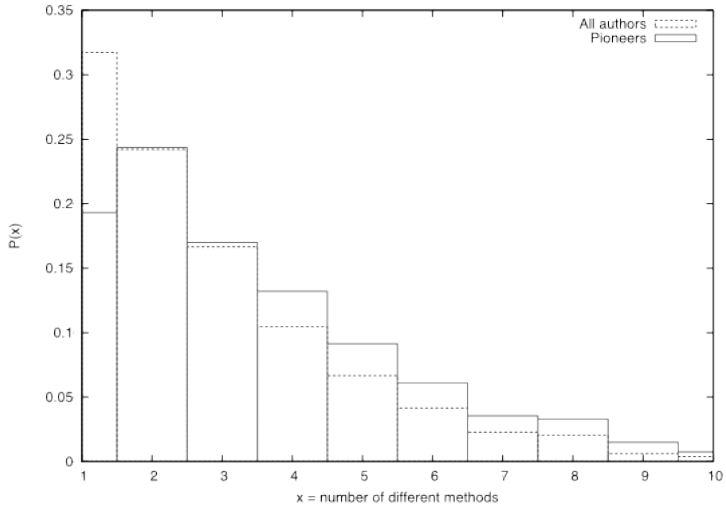
Fonction de répartition de la proportion d'articles que les « pionniers » avaient déjà publiés au moment où ils ont publié leur premier article sur une nouvelle méthode, par rapport à la production totale de leur carrière



La figure 6 révèle aussi que 70 % des « pionniers » ont publié moins du tiers de leur production totale au moment où ils utilisent une nouvelle méthode. On observe donc un regroupement partiel entre ces pionniers et les jeunes chercheurs du domaine ou, comme on l'a vu dans le paragraphe précédent, entre ces pionniers et des chercheurs ayant jusque-là publié dans des communautés proches mais néanmoins différentes. Il faudrait donc étudier en parallèle d'autres corpus (en informatique, en linguistique, en sciences cognitives, etc.) pour pouvoir affiner la description, mais la tâche est dès lors difficile.

Figure 7

Proportion de « pionniers » experts d'un nombre donné de méthodes et comparaison avec ce même indicateur pour l'ensemble des auteurs du corpus



On peut ensuite se poser la question de la diversité de méthodes employées par les auteurs du domaine, en particulier par le groupe que nous avons appelé « pionniers ». La figure 7 montre le nombre de méthodes détectées par article pour les pionniers d'une part (trait continu) et pour l'ensemble des auteurs d'autre part (en pointillés). On voit chez les pionniers (en prenant en compte l'intégralité de leur production scientifique dans la collection *ACL Anthology*) une nette sous-représentation de chercheurs utilisant une seule méthode, et une sur-représentation (statistiquement significative) du nombre d'auteurs utilisant quatre méthodes ou plus. Le groupe que nous appelons « pionniers » a donc une tendance marquée à utiliser plus de méthodes (et aussi à aborder davantage de sous-domaines

du traitement automatique des langues) que l'ensemble des auteurs pris globalement.

Finalement, nous essayons de mesurer les flux entre méthodes : un chercheur ayant travaillé sur une méthode donnée a-t-il plus de chances de travailler ensuite sur telle ou telle autre méthode (par exemple, un chercheur ayant utilisé les HMM a-t-il plus de chances de se tourner vers les SVM ou les CRF si les deux méthodes sont populaires en même temps)? Nous mesurons ces flux en analysant les évolutions de méthodes d'une période à l'autre (articles d'un auteur donné ayant utilisé une méthode pendant une période puis une autre méthode au cours de la période suivante par exemple). Les flux sont ensuite normalisés en prenant en compte le nombre total d'auteurs concernés. Les figures 8, 9 et 10 montrent les résultats ainsi obtenus.

Figure 8

Réseau des flots d'auteurs entre méthodes. Pour chaque couple de méthodes, nous avons calculé le flot de l'une à l'autre en comptant le nombre d'auteurs qui ont publié successivement un article employant la première dans les années 1980 puis la deuxième méthode considérée dans les années 1990. Chaque flot est normalisé en fonction du nombre total d'auteurs concernés (tous les flots inférieurs à 10% sont supprimés).

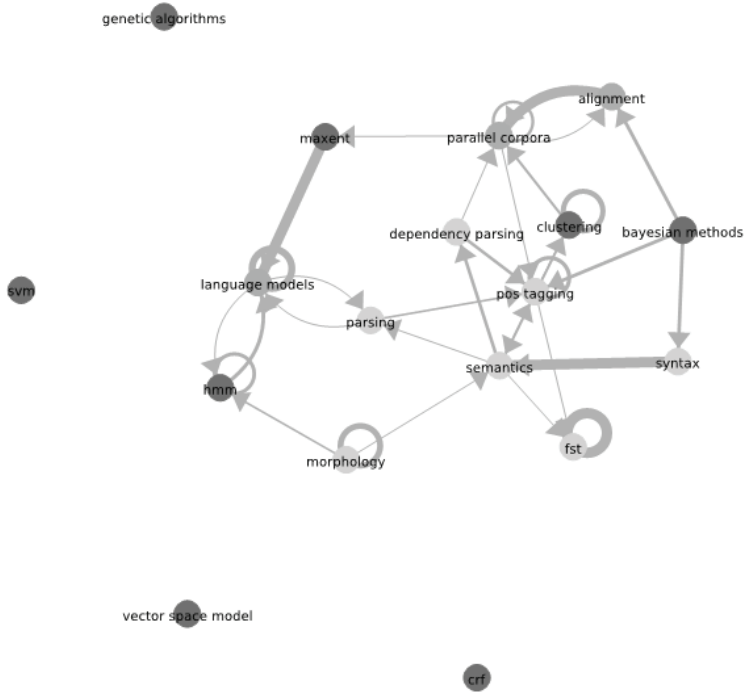


Figure 9

Réseau des flots d'auteurs entre méthodes des années 1990 à la première moitié des années 2000

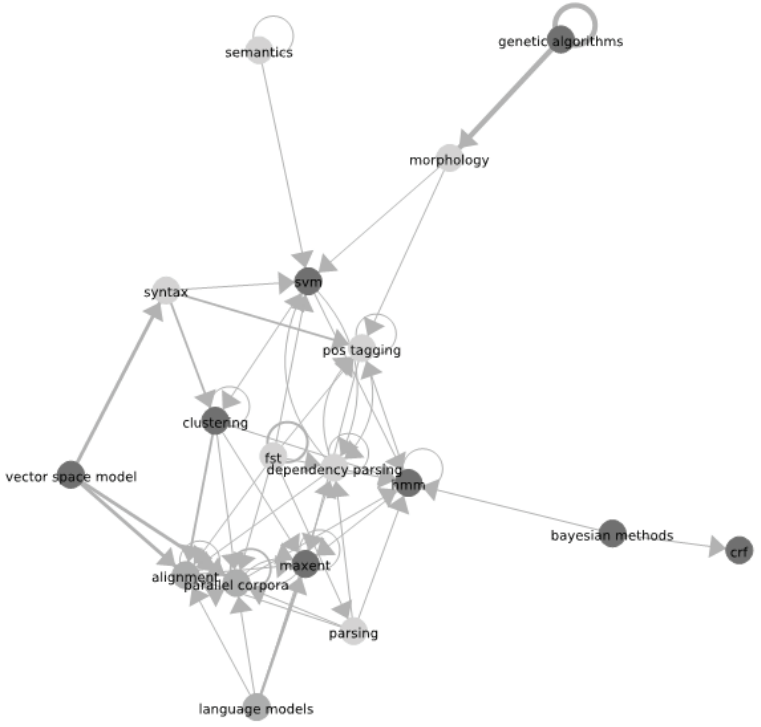
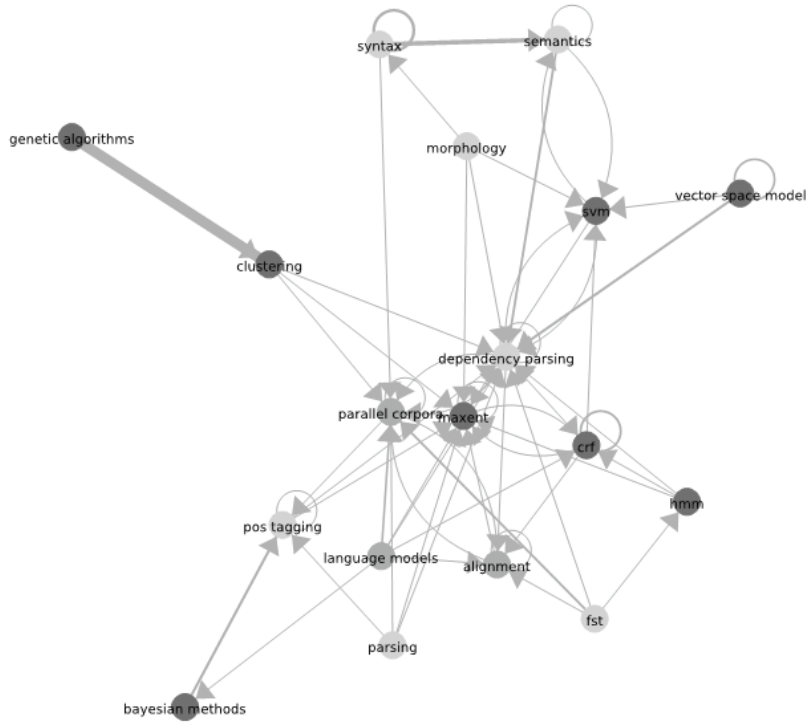


Figure 10

Réseau des flots d'auteurs entre méthodes de la première à la deuxième moitié des années 2000



Nous pouvons observer que le flux d'auteurs depuis les années 1980 jusqu'aux années 1990 concerne principalement les méthodes de TAL, les techniques d'apprentissage automatique n'étant pas encore utilisées, à l'exception des modèles de Markov cachés, qui sont devenus populaires à partir des années 1990 (voir la figure 8). Des années 1990 à la première moitié des années 2000, les méthodes employées concernent davantage l'apprentissage automatique comme, par exemple, les *Support Vector Machines*, devenus très populaires alors (voir la figure 9). De la première à la seconde moitié des années 2000, les chercheurs se

concentrent davantage sur les *Conditional Random Field* (une technique d'apprentissage automatique pour le traitement du langage naturel), et sur un domaine spécifique de la syntaxe : l'analyse en dépendances (*Dependency Parsing* en anglais), qui a fait l'objet de plusieurs campagnes d'évaluation dans les années 2000, en particulier au cours des conférences *CoNLL* de 2006 à 2009 (voir la figure 10). Nous observons aussi que l'analyse morphosyntaxique (*POS tagging*) a toujours occupé un rôle important, ce qui est probablement dû au fait que cette technique est quasi systématiquement utilisée en linguistique informatique comme prétraitement. Enfin, nous remarquons que l'alignement et les corpus parallèles sont devenus majeurs depuis les années 2000, ce qui reflète la popularité de la traduction automatique depuis plus d'une décennie.

4. Conclusion

Nous avons présenté une analyse du corpus *ACL Anthology* visant à en faire ressortir certaines caractéristiques remarquables. Notre analyse se fonde d'une part sur une méthode classique d'extraction de termes, d'autre part sur l'analyse de la structure argumentative des textes considérés afin de catégoriser les termes en fonction de leur contexte et de leur contenu informationnel. Nous avons montré que ce type de technique contribue à affiner la description de la dynamique du domaine.

Il s'agit encore une fois de simples observations. Les outils mettent en avant certains phénomènes qu'il faut ensuite expliquer par un retour aux textes, voire par une enquête de terrain. Cette recherche par nature pluridisciplinaire nous amène maintenant à nous tourner vers des spécialistes d'histoire des sciences pour poursuivre ce travail en collaboration. Les outils et l'infrastructure mise en place sont toutefois d'ores et déjà utilisables et seront appliqués à d'autres corpus, comme le corpus *APS* présenté dans l'introduction.

Bibliographie

Anderson, Ashton., Dan Jurafsky et Daniel A. McFarland, « Towards a Computational History of the ACL : 1980-2008 », *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, Jeju, Corée, Association for Computational Linguistics, 2012, p. 13–21.

Banchs, Rafael E. (dir.), *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, Jeju, Corée, Association for Computational Linguistics, 2012.

Bourigault, Didier et Christian Jacquemin, « Term Extraction + Term Clustering: An Integrated Platform for Computer-aided Terminology », *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics*, EACL, 1999, p. 15–22.

Brown, Peter F. *et al.*, « A Statistical Approach to Machine Translation », *Computational Linguistics*, vol. 16, n° 2, 1990, p. 79-85.

Callon, Michel, Jean-Pierre Courtial et Frédérique Laville, « Co-Word Analysis as a Tool for Describing the Network of Interaction between Basic and Technological Research : The Case of Polymer Chemistry », *Scientometrics*, vol. 22, n° 1, 1991, p. 155-205.

Callon, Michel, John Law et Arie Rip, *Mapping the Dynamics of Science and Technology Sociology of Science in the Real World*, Basingstoke, McMillan, 1986.

Curran, James, Stephen Clark et Johan Bos, « Linguistically Motivated Large-Scale NLP with C&C and Boxer », *Proceedings of the 45th Meeting of the Association for Computation Linguistics (ACL)*, 2007, p. 33-36.

Frantzi, Katarina, Sophia Ananiadou et Hideki Mima, « Automatic Recognition of Multi-Word Terms: The C-value/NC-value Method », *International Journal on Digital Libraries*, Springer-Verlag, vol. 3, n° 2, 2000, p. 115-130.

Garfield, Eugene, « Citation Analysis as a Tool in Journal Evaluation. », *Science*, vol. 178, n° 4060, 3 novembre 1972, p. 471-479.

Girvan, Mark et Mark E.J. Newman, « Community Structure in Social and Biological Networks », *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, n° 12, 2002, p. 7821-7826.

Guimera, Roger *et al.*, « Team Assembly Mechanisms Determine Collaboration Network Structure and Team Performance », *Science*, vol. 308, n° 5722, 29 avril 2005, p. 697–702.

Guo, Yufan, Anna Korhonen et Thierry Poibeau, « A Weakly-Supervised Approach to Argumentative Zoning of Scientific Documents »,

Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Edimbourg, 2011, p. 273-283.

Guo, Yufan, Roi Reichart et Anna Korhonen , « Improved Information Structure Analysis of Scientific Documents through Discourse and Lexical Constraints », *Proceedings of Human Language Technologies : Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*, 2013, p. 928-937.

Lee, Gary Geunbae, Jeongwon Cha et Jong-Hyeok Lee, « Syllable-Pattern-Based Unknown-Morpheme Segmentation and Estimation for Hybrid Part-of-speech Tagging of Korean », *Computational Linguistics*, vol. 28, n° 1, 2002, p. 53-70.

Mizuta, Yoko *et al.*, « Zone Analysis in Biology Articles as a Basis for Information Extraction », *International Journal of Medical Informatics*, vol. 75, n° 6, 2006, p. 468-487.

Press, William H. *et al.*, *Numerical Recipes: The Art of Scientific Computing*, 3e édition, New York, Cambridge University Press, 2007.

Radev, Dragomir *et al.*, « Centroid-Based Summarization of Multiple Documents », *Journal on Information Processing Management*, vol. 40, n° 6, 2004, p. 919-938.

Rogers, Everett M., *Diffusion of Innovations*, New York, Simon and Schuster, 1962.

Small, Henry G., « Co-Citation in the Scientific Literature : A New Measure of the Relationship between Two Documents », *Journal of American Society for Information Science*, vol. 24, n° 4, 1973, p. 265-269.

Tbahriti, Imad *et al.*, « Using Argumentation to Retrieve Articles with Similar Citations : An Inquiry into Improving Related Articles Search in the Medline Digital Library », *International Journal of Medical Informatics*, vol. 75, n° 6,2006, p. 488-495.

Teufel, Simone, *Argumentative Zoning : Information Extraction from Scientific Articles*, thèse de doctorat, University of Edinburgh, 1999.

Teufel, Simone et Mark Moens, « Summarizing Scientific Articles – Experiments with Relevance and Rhetorical Status », *Computational Linguistics*, vol. 4, n° 28, 2002, p. 409-445.

Teufel, Simone, Advaith Siddharthan, Dan Tidhar, « Automatic Classification of Citation Function », *Proceedings of Empirical Methods in Natural language Processing* (EMNLP), Association for Computational Linguistics, 2006, p. 103-110.