Notes de lecture

Rubrique préparée par Denis Maurel

Université François Rabelais Tours, LI (Laboratoire d'informatique)

Agnès TUTIN, Francis GROSSMANN. L'écrit scientifique, du lexique au discours : autour de Scientext. *Presses universitaires de Rennes*. 2013. 232 pages. ISBN 978-2-7535-2846-8.

Lu par **Nadine LUCAS**

GREYC - CNRS UMR 6072 – Université de Caen Basse-Normandie

L'étude de l'écrit scientifique vise en particulier à mieux cerner ce qui fait la spécificité du langage scientifique et à comparer les rhétoriques disciplinaires, étudiées à travers un corpus contemporain sur les sciences humaines et sociales mais aussi les sciences expérimentales (médecine) et les sciences appliquées. Au centre des préoccupations, il y a la manière dont sont mises en scène, dans le lexique et les discours, les procédures de découverte et de validation des connaissances. L'ouvrage s'appuie sur un ensemble de recherches effectuées à partir du corpus Scientext, corpus en français et en anglais développé par plusieurs équipes de chercheurs sous la responsabilité du LIDILEM à Grenoble 3. Deux problématiques sont privilégiées, le positionnement *sociolinguistique de l'auteur scientifique par rapport à ses devanciers et à ses pairs, et celle du* raisonnement*, tel qu'il apparaît dans l'argumentation ou les chaînes causales.*

Contenu et structure

Cet ouvrage collectif porte sur les retombées d'un projet de l'ANR, Scientext, achevé fin 2009. Il concerne la constitution, et l'exploitation *via* une interface informatique, d'un corpus scientifique, composé d'articles et de thèses en français, soit 219 textes consultables (4,8 millions de mots). Le sous-corpus des didacticiens (300 textes, 1,1 million de mots) est beaucoup plus restreint et non partagé, de même que le corpus exploratoire sur les évaluations de communications de colloques (502 documents, 34 857 mots). Le corpus en anglais issu de BMC est beaucoup plus conséquent : 3 381 articles (13,8 millions de mots) dans le domaine médical et biologique.

Une introduction et une bibliographie, suivie des résumés des contributions, encadrent trois parties d'inégale longueur. L'ouvrage compte dix contributions assez homogènes, huit sont issues du LIDILEM, maître d'œuvre du projet. Sur les neuf contributions centrales, la majorité aborde l'aspect lexical et explicite de l'écrit académique, en recensant dans des sous-corpus les occurrences de termes emblématiques : le verbe *voir*, le verbe *causer*, les collocations telles que *On constate* ou *résultats intéressants.* Les études sur le français dominent, deux seulement portent sur l'anglais. La constitution d'un dictionnaire dynamique est

traitée par G. Williams et C. Millon à partir d'un gros corpus de médecine en anglais. Quatre contributions ont des problématiques un peu différentes. L'une, en fin de première partie, traite de l'ingénierie du projet, deux contributions, réunies dans la deuxième partie, traitent d'applications pédagogiques, enfin une seule contribution en troisième partie traite de la titraille des articles du corpus.

L'objectif était de permettre l'outillage de l'étude de phénomènes linguistiques sur corpus par des linguistes, grâce notamment à ScienQuest, une interface simple d'interrogation. Cet objectif est atteint. Les conclusions des études de corpus reflètent l'utilité des concordanciers et la satisfaction des linguistes de pouvoir recenser les occurrences recherchées. Sur le plan de l'analyse linguistique, les conclusions montrent cependant les limites du dispositif. Elles reflètent une certaine déception, souvent explicite, par rapport aux hypothèses de constance d'une phraséologie scientifique et de représentativité des expressions recherchées. Les mots candidats sont peu fréquents et inégalement répartis entre les disciplines.

Commentaire

Voici donc un projet qui vise « à outiller la linguistique ». Sans doute faut-il se rappeler qu'il ne s'agit pas ici d'un ouvrage destiné à un public de TAL. L'exploration des outils prend du temps et ce sont donc les premiers pas d'utilisateurs presque tous novices qui sont présentés ici. Sans entrer dans le détail, nous relevons les tendances générales et les exceptions qui confirment la règle.

L'article d'A. Falaise sur le corpus et les outils est curieusement situé tout à la fin de la première partie, et donne donc tardivement la perspective des moyens mis en œuvre à l'intention des linguistes. Avec le logiciel ScienQuest, il leur propose un outil « intuitif » pour explorer leur corpus, mais cela se solde souvent par une réduction drastique des choix possibles à l'aide de menus. On notera aussi que le site de Scientext n'est pas mis en avant : on trouve son adresse[1] p. 16 de l'introduction et dans les références.

Il est vrai que Scientext ne dissocie pas clairement le corpus, le modèle et les outils. À ce titre, il ne semble pas pouvoir se prêter à d'autres problématiques que celles retenues par les premiers partenaires. Dans ScienQuest, le mode de consultation dit « sémantique » en décevra plus d'un, c'est en effet la vue par défaut, donnant les résultats de grammaires déjà construites en fonction des questions traitées par l'équipe. Le mode libre permet de construire une recherche à l'aide d'un assistant, enfin le mode avancé permet de faire des requêtes classiques avec opérateurs et distance entre termes.

Si la base Scientext représente un corpus relativement varié pour des linguistes, elle souffre de limitations d'accès, du fait des droits d'auteur : de fait, la moitié du corpus brut collecté en français n'est consultable qu'en intranet, et le corpus analysé par Syntex ne l'est pas non plus. On est encore loin de la diversité des corpus et des études menées par l'école scandinave ou des outils statistiques de l'école anglaise.

1. http://scientext.mrsh.alpes.fr.

Par ailleurs, la majorité des linguistes se font encore une piètre idée de l'informatique, jugeant par exemple que détecter des expressions discontinues est « très difficile ». En conséquence, ils ont très peu recours à la consultation en mode libre ou avancé, plus souple que les concordanciers. Les résultats de l'analyseur syntaxique Syntex de Bourigault restent sous-utilisés. Il y a donc matière à apprentissage et à progression dans ce domaine.

Bien des auteurs se sont montrés méfiants : ils se sont contentés d'ajouter des décomptes d'occurrences à leur description du français académique contemporain, sans étendre leur sous-corpus. La majorité des études portent sur des collections d'une trentaine de textes, principalement sur les sciences humaines. À l'opposé, la contribution de M.-P. Jacques fait exception en traitant de l'ensemble du corpus français disponible, à travers leurs titres et sous-titres. L'étude de Williams et Millon se réfère à une vaste collection d'articles en anglais issue du corpus médical BMC, dans le but de générer un dictionnaire.

D'un point de vue méthodologique, la majorité des études sont centrées sur l'usage du lexique à travers quelques disciplines, mesuré par la fréquence brute d'occurrences. Il est assez étonnant pour des linguistes que l'approche reste attachée au métalangage explicite, par exemple « nous nous appuyons sur Untel » pour étudier les emprunts de méthode. Mais le logocentrisme ne devrait pas heurter un public de talistes, également attaché à la valeur littérale des indices, le « sémantisme » des mots.

Du moins, l'ouvrage aura le mérite de montrer répétitivement que la valeur lexicale ne suffit pas à capter le raisonnement causal, ou le positionnement sociolinguistique, les deux problématiques qui sous-tendent les investigations. Un constat important affleure : la représentativité des termes relève du prototype sémantique et ne correspond nullement à une grande fréquence d'emploi, non plus qu'à une distribution uniforme et transdisciplinaire des termes considérés.

La bibliographie est bien développée sur l'approche linguistique logocentrique et l'approche sociolinguistique. En revanche, elle ne traite pas des aspects stylistique et pédagogique de la rhétorique scientifique, ni de la relation entre l'illustration et le texte. Sans surprise, elle n'établit aucun lien non plus entre l'étude des textes scientifiques et les travaux de fouille de textes associant linguistes et informaticiens.

Pour les informaticiens linguistes, malheureusement, la base de textes Scientext est peu exploitable. Elle est trop restreinte, à l'heure des *big data*, à la fois en nombre de textes et de disciplines représentées. L'impossibilité d'accès direct au corpus français, brut ou annoté est rédhibitoire pour l'exploration occasionnelle ; le seul corpus téléchargeable est celui de BMC en anglais et il était déjà disponible.

La plate-forme Scientext peut cependant aider les talistes lexicologues à explorer semi-manuellement le corpus proposé, par exemple pour détecter des patrons syntaxiques associés à des formules d'usage, la « phraséologie ». Il est également possible d'avoir accès sur demande au corpus brut.

Jesse TSENG. Prépositions et postpositions : approches typologiques et formelles. *Hermès-Lavoisier*. 2013. 253 pages. ISBN 978-2-7462-4518-1.

Lu par **Marie-Hélène LAY**

Laboratoire FoReLL (EA3616) – Université de Poitiers

Jesse Tseng propose ici un ouvrage consacré à l'étude des prépositions, présentées comme une catégorie lexicale « majeure » pour laquelle la question de la faiblesse grammaticale se pose tout particulièrement. De taille restreinte, cette catégorie présente des éléments caractérisés par la multitude de leurs emplois (due à la polysémie et à des phénomènes de grammaticalisation) et par le fait qu'ils entrent dans des constructions « contraintes » (syntaxiquement déterminées ou lexicalement figées). Souvent considérée comme inerte sur le plan morphologique (absence de productivité flexionnelle et dérivationnelle), cette catégorie peut présenter, par ailleurs, des comportements morpho-phonologiques complexes (contraction). Les six contributions de ce volume sont des contributions longues (de 25 à 40 pages environ) abordant la question des prépositions faibles pour six langues différentes. Les faits étudiés sont variés, comme le sont les perspectives descriptives et analytiques abordées.

Le chapitre 1 offre une présentation détaillée de l'inventaire des diverses formes susceptibles d'être employées comme des prépositions en tswana. Elles constituent une classe hétérogène du point de vue phonologique, morphologique et syntaxique comme du point de vue de leurs emplois, ce qui conduit l'auteur à les analyser en comparaison avec d'autres connecteurs et joncteurs. La contribution traite des éléments situés à la marge gauche des constituants nominaux qui informent sur les relations syntaxiques et sémantiques que ce constituant entretient avec le reste de la phrase. Ils sont très largement analysés comme des préfixes du nom par la tradition, bien qu'un petit nombre d'entre eux (ce qui est une propriété des langues subsahariennes) puisse fonctionner comme des éléments autonomes indécomposables. Leur identification en tant que « mot » repose sur la propagation postlexicale du ton haut et sur la possibilité ou l'impossibilité de l'apparition d'un abaissement du registre haut, le *downstep* étant la manifestation de la démarcation lexicale. *Le*, *ka*, *ke*, d'une part, *-a*, *go -ng* d'autre part et enfin *ko*, *fa*, *mo* sont successivement étudiés dans leurs emplois comme prépositions, connecteurs et joncteurs liés au constituant ou pas.

Le chapitre 2 présente une étude contrastive des contractions préposition et article en français et en allemand. L'analyse proposée ici avance que les contractions sont dues dans les deux langues à des composantes différentes de la grammaire. La contraction en français s'effectue au niveau phono-morphologique et produit des prépositions fléchies appartenant au lexique présyntaxique, elle est obligatoire et exclut de ce fait certaines formes de coordination distante sans reprise de la préposition. Elle est déterminée par le genre et le nombre du mot (masculin singulier ou pluriel) qui suit, mais ne se produit pas s'il y a élision de la voyelle de l'article (*à l'homme*). En revanche en allemand, ces formes peuvent apparaître contractées ou non, dans des contextes comparables mais avec des interprétations distinctes, selon

qu'il s'agit d'une lecture de définitude pragmatique ou de définitude sémantique. Ce phénomène postsyntaxique est déterminé lexicalement et pas phonologiquement. Par ailleurs, aussi bien la préposition que le déterminant gardent une indépendance syntaxique malgré la forme morphologique qui amalgame les deux, ce qui ne serait pas le cas en français, les amalgames n'occupant alors qu'une position syntaxique.

Le chapitre suivant se penche sur le statut morphosyntaxique délicat des prépositions : elles apparaissent sous une forme préfixée, sans être pour autant des marqueurs casuels mais bien plutôt des prépositions présentant une déficience phonologique leur interdisant de se projeter en catégorie autonome. L'analyse de la distribution des six prépositions faibles du kabyle *f* (sur), *s* (avec, instrument), *o* (comitatif), *g* (dans), *i* (datif) et *n* (génitif) amène l'auteur à conclure qu'elles sont la tête du syntagme lorsqu'elles précèdent un nom à l'état d'annexion ; elles font alors partie du domaine phonologique de leur complément nominal et ne sont pas des marqueurs casuels au sens strict : elles n'introduisent pas de complément phrastique régi dans ces contextes. Des considérations sur le gabarit des complémenteurs amènent à la conclusion que les prépositions faibles affixées ne peuvent pas être visibles en syntaxe. Leur affixation dans ce contexte est un processus de composition, et non de concaténation syntaxique, ce qui permet en outre d'expliquer les restrictions portant sur le redoublement des prépositions faibles dans les constructions à long mouvement *wh*.

Le chapitre 4 concerne un aspect particulier des postpositions du coréen afin de rendre compte des empilements possibles de plusieurs éléments portant sur le même constituant. Rompant avec une tradition les envisageant comme des cas, l'auteur les traite ici comme des enclitiques avec un statut syntaxique de tête faible. Ils sont formalisés en HPSG et traités comme relevant de trois sous-catégories : marqueurs syntaxiques cas et relations syntaxiques. *-i*, *-leul- -eun*, *-e* sont étudiés. Elles s'attachent à un item lexical comme des suffixes dont elles ne partagent pourtant pas les comportements. Elles se comportent comme des clitiques qui se combinent avec un syntagme et s'attachent en phonologie au dernier mot. Elles sont analysées comme des têtes faibles qui prennent pour complément le syntagme dominant l'hôte phonologique dont elles héritent les propriétés syntaxiques. Cette analyse permet l'économie d'une distinction entre deux homonymes « postposition casuelle / non-casuelle ». Elle simplifie aussi la description des empilements, explicitant les contraintes sur l'ordre et la cooccurrence.

La contribution du chapitre 5 propose une classification et une analyse formelle de prépositions de certains parlers kurdes. Il s'agit des prépositions qui présentent deux allomorphes à l'exclusion des prépositions composées et des prépositions nominales. La *forme simple* de la préposition se combine avec des compléments de forme pleine (forte) alors que sa *forme absolue* n'accepte que les compléments pronominaux de forme faible. Cette dernière est étudiée plus en détail, afin de rendre plus précisément compte du mode de la réalisation de ces compléments : les morphèmes personnels liés peuvent être divisés en deux classes selon leurs propriétés de placement, clitique et désinence personnelle sur le verbe, apparaissant en distribution complémentaire (en corrélation avec les aspects transitif et/ou intransitif des verbes et le temps employé). Un traitement formel (HPSG) de

l'alternance entre les formes simples et absolues est ensuite proposé. Il repose d'une part sur une classification des propositions intégrant deux dimensions (le caractère nominal ou non de la préposition et la réalisation argumentale – constituant ou affixe), et d'autre part sur les informations contenues dans les entrées lexicales respectives des membres de chaque classe.

Le chapitre 6 présente une étude comparée des deux formes *pe* et *a* qui, outre des emplois prépositionnels comparables pour partie à ceux de *à* et *sur* en français, marquent, sous certaines conditions, l'objet direct d'un verbe transitif : elles projettent alors des groupes nominaux dont la nature est celle de leur complément. Leur apparition peut être corrélée avec le type sémantique de l'objet et avec les propriétés des structures dites à incorporation : ils ont nécessairement une dénotation de type individu ne pouvant donc pas être incorporés sémantiquement et sont exclus des objets directs ayant une dénotation de type propriété. Par ailleurs, un objet direct non incorporé peut être interprété comme topicalisé. Deux formalisations sont alors suggérées, les analysant soit comme des marques de topicalisation, soit comme des marques de cas fort. Les facteurs déterminant la présence ou l'absence de *pe* et *a* sont organisés en une échelle hiérarchique formée de trois paramètres ayant des valeurs graduelles : le caractère animé de l'objet direct, son caractère spécifique et sa topicalisation. Dans une perspective comparative, ce sont ces valeurs qui fondent la différence de l'emploi de *pe* et de *a*, seul l'espagnol étant sensible à la topicalisation.

Ces travaux peuvent sembler ardus à un étudiant de TAL, mais ils permettent d'aborder de façon concrète (par la diversité des langues étudiées) un certain nombre de problèmes centraux pour la linguistique générale : par exemple l'imbrication complexe des « niveaux » de description que l'on tend trop souvent à vouloir dissocier strictement : il y est régulièrement question de problèmes phono-morphologiques, morphosyntaxiques voire phono-syntaxiques.

Philipp CIMIANO, Christina UNGER, John McCRAE. Ontology-Based Interpretation of Natural Language. *Morgan & Claypool publishers*. 2014. 155 pages. ISBN 978-1-6084-5989-6.

Lu par **François LÉVY**

Université Paris 13 – LIPN

Ce livre propose une architecture de TAL modulaire, fondée sur les standards du Web et des données ouvertes. L'analyse syntaxique et la représentation sémantique sont menées en parallèle. Les données pour cela sont compilées à partir d'un lexique disposant d'une large palette de notations linguistiques et de références à une ontologie. Le lexique et l'ontologie utilisent des formats standard du Web, ce qui les rend échangeables et réutilisables.

Le livre publié par nos trois collègues de Bielefeld est issu d'un cours à l'ESSLI et dans leur université. Il réussit en cent cinquante pages un exposé clair et pédagogique de l'architecture de TAL qu'ils proposent. C'est le terme *interpretation*

employé dans le titre qui m'a donné envie d'y regarder de plus près, en ce qu'il évoque une conception non compositionnelle. Effectivement, le lien entre texte et représentation sémantique ne se limite pas à un décalque de l'analyse syntaxique. Il repose sur une conception du lexique à laquelle les auteurs ont consacré beaucoup d'efforts, et qui est intéressante. Je schématise la conception de l'ensemble avant de proposer une analyse plus personnelle de ses perspectives.

Formalismes et calculs

Au départ, il y a un texte et un ensemble formalisé de connaissances *a priori* sur le monde. À l'arrivée, des connaissances formalisées supplémentaires apportées par le texte. Les connaissances sur le monde sont décrites dans une ontologie. Le chapitre 2 rappelle les principes et les axiomes qui sous-tendent OWL DL et OWL2 DL et leur correspondance avec une partie de la logique du premier ordre (LPO). L'interprétation utilise une grammaire d'arbres adjoints lexicalisée (LTAG) pour l'analyse syntaxique. Le chapitre 3 explique le couplage de cette grammaire et du formalisme sémantique des DUDES[2] qui est pour l'essentiel celui de la théorie des représentations discursives (DRT), augmenté pour les besoins du couplage. Très schématiquement, chaque arbre syntaxique élémentaire représente un mot ou une expression avec les places syntaxiques qui seront utilisées pour sa combinaison avec les autres arbres de la phrase. De plus, chaque arbre élémentaire est associé à une DRS qui en constitue la représentation sémantique : le vocabulaire est celui de l'ontologie (dans une formulation en LPO). Les *paires sélectives* qui augmentent la DRS associent chacune une place syntaxique de l'arbre et une variable de l'univers de la DRS ; de plus, la *variable principale* est celle qui est marquée pour une possible unification. Grâce à ce couplage, les opérations de combinaison des DUDES sont définies en parallèle avec celles des arbres syntaxiques (substitution et adjonction). Le pouvoir d'expression de cette construction n'est pas analysé, mais il me semble plus fort qu'un typage générique. Pour rester dans le domaine du football qui illustre le livre, le sujet de « gagner » est une équipe (c'est le typage générique de l'image de *gagne(m, eq)*) ; de plus, elle est gagnante d'un certain match qui est la variable principale de la DUDES de « gagner », ce qui introduit une relation avec ce match et augmentera la contrainte lors des unifications suivantes. Le texte ne précise toutefois pas comment l'inférence dans l'ontologie est utilisée dans les DUDES.

Reste, et ce n'est pas un mince problème, à rassembler les connaissances linguistiques utilisées pour l'interprétation, autrement dit les arbres élémentaires et leurs DUDES associés. Cette partie du travail est séparée en deux modules : la constitution d'un lexique syntactico-sémantique dont le format se veut indépendant du choix syntaxique, et la génération d'une grammaire dans le formalisme choisi. Le premier est assuré par le format de lexique Lemon : la version actuelle (2.0) repose sur une ontologie de 29 classes et 54 propriétés et fait aussi appel aux descriptions syntaxiques de l'ontologie Lexinfo, soit 192 classes, 156 propriétés et 271 individus permettant de noter une grande variété de propriétés syntaxiques[3]. On a là un format

2. DUDES signifie *Dependancy-based Underspecified DRS.*

3. Chiffres calculés à partir des ontologies publiques indiquées par le livre, juillet 2014.

de lexique bien plus détaillé que Skos. Le livre donne un ensemble d'exemples avec une réelle finesse linguistique : noms de classes et noms relationnels, verbes d'état, d'événement, résultatifs, adjectifs intersectifs, scalaires, modifieurs de classe, relationnels. Chacune des entrées lexicales fournies décrit des variantes morphologiques, un comportement syntaxique que je lis comme un cadre de sous-catégorisation, et un sens qui est un fragment de l'ontologie de domaine lié aux positions libres du cadre.

La génération de la grammaire à partir d'une de ces entrées produit plusieurs arbres selon les variantes morphologiques et les emplois : pluriel des noms, relations prépositionnelles, passif des verbes, comparatif et superlatif des adjectifs par exemple. D'après le schéma d'implémentation qui conclut le chapitre 5, l'automatisation repose sur des patrons adaptés à chaque cadre et aux arguments présents. J'ai compté quatre-vingt-quatre cadres dans Lexinfo – il n'était donc pas possible de décrire le dispositif en quelques pages.

Trois thèmes plus pointus concluent l'ouvrage : la désambiguïsation, le temps, le *Question Answering*. Une ontologie des intervalles temporels permet la modélisation du temps verbal (présent, passé, futur) et des expressions temporelles comme *aujourd'hui*, *hier*, *la semaine prochaine*. Je reviendrai sur la désambiguïsation dans l'analyse. La traduction de questions en requêtes SPARQL est une application sur des textes d'une phrase focalisés sur le contenu de la base de données de football.

Analyse

Le dispositif formel est cohérent et complet. Les données sont modularisées. De plus, elles sont toutes en RDF, connaissances du monde et données lexicales. Il est donc facile de les publier, de les analyser ou de les réutiliser. Il est aussi possible d'augmenter l'ontologie Lexinfo, d'ajouter un champ au lexique ou de modifier le moteur d'utilisation sans toucher au reste.

Les auteurs signalent eux-mêmes que l'expérimentation en grandeur nature est encore en projet. J'ai relevé quelques problèmes du passage à l'échelle. En premier lieu, l'inventaire des arbres syntaxiques élémentaires est contraint par la sémantique : « *The whole prepositional phrase corresponds to one atomic élément on the semantic side, so it should also be one atomic élément on the syntactic side* » (p. 40). Cela conduit à la multiplication des arbres syntaxiques élémentaires.

Le livre souligne aussi, à juste titre, la nécessité d'obtenir des représentations comparables pour des formulations différentes (p. 7-8) : « *an early opening goal by X* » et « *X opened the scoring shortly after kick off* » évoquent tous deux un match, un but, son auteur X et sa date. Si les fragments d'ontologie figurant dans les DUDES sont homologues pour des formulations différentes, c'est au prix d'une spécialisation poussée : les représentations sémantiques de *win* comme de *loose* évoquent un match. Cependant, le Webster indique que l'on peut aussi gagner un tournoi, une guerre, une récompense, sa vie, sans compter des sens plus éloignés. Ce thème est repris dans le chapitre sur la résolution des ambiguïtés : un élément peut avoir un choix de représentations sémantiques, celles-ci sont filtrées par la compatibilité des remplisseurs de leurs places syntaxiques trouvés dans le texte. On

gagne en pouvoir d'expression, mais assez peu en largeur de domaine, car chaque représentation est spécialisée.

Le choix de la variable principale des composants incluant un verbe semble aussi une difficulté : les exemples incluent des formes progressives, des passifs, des relatives, mais le degré de généralité de chaque exemple est difficile à apprécier. La marque du temps verbal s'unifie à la variable principale du verbe ; pour l'exemple de « gagner », il s'agit du match, mais je n'ai pas vu la solution pour « respecter », dont la représentation n'est pas ainsi réifiée et qui n'a pas de variable principale indiquée.

Conclusion

Le premier mérite de ce livre est d'allier résolument syntaxe et sémantique sans supposer une relation d'ordre entre elles. La modularisation très claire et la normalisation des données, le recours aux standards du Web sont très intéressants. Le tout repose sur l'idée, à mon sens pertinente, que la sémantique lexicale dépend du domaine et que la modularité permet l'adaptation. Il reste, et c'est je crois une condition inhérente au projet d'apprendre le lexique dont les auteurs font état, à calibrer le type de domaine pour lequel cette architecture passe à l'échelle.

Pierre-André BUVET. La Dimension lexicale de la détermination en français. *Honoré Champion*. 2013. 473 pages. ISBN 978-2-7453-2604-1.

Lu par **Catherine SCHNEDECKER**

LiLPa – Université de Strasbourg

L'ouvrage propose une typologie des déterminants du français fondée sur un inventaire de formes plus vaste que ce qui est traditionnellement préconisé, compte tenu d'une démarche globale appréhendant les phénomènes aux plans syntaxique, sémantique et même morphologique. Dans cette optique, le lexique occupe une place cruciale et il opère doublement, en tant que matériau constitutif de certaines sous-catégories de déterminants mais aussi en tant que tête lexicale du SN imposant divers niveaux de contraintes (syntaxiques et sémantiques) sur le déterminant. S'ensuit une nouvelle cartographie de la détermination qui oblige à repenser les frontières intercatégorielles.

L'ouvrage de Pierre-André Buvet résulte d'une HDR soutenue en 2009 à l'université de Paris 13. Pour autant, cet ouvrage ne se réduit pas aux aspects de la détermination, déjà nombreux, abordés par l'auteur. Il constitue au contraire d'abord une synthèse sur la détermination, comme le montre sa structure : en effet, l'ouvrage se subdivise en deux parties, plus une importante annexe qui fournit trois listes des principaux prédéterminants et antédéterminants du français, de l'ensemble des déterminants nominaux et de nombreuses séquences déterminatives figées.

La première partie de l'ouvrage comprend trois chapitres consistant en la définition et les propriétés de la détermination dans le cadre théorique auquel souscrit l'auteur. Le chapitre 2 inventorie les formes dites de la détermination simple *vs* complexe,

dont la caractérisation s'appuie sur d'abondantes batteries de tests. Le chapitre 3 aborde, comme l'indique son titre, l'ensemble des modifications du GN, dont une catégorie de modifieurs propositionnels (les complétives, infinitives et participiales), souvent occultés par les grammaires. Quant à la seconde partie, elle se consacre principalement aux déterminants défini et démonstratif, jusque dans leurs emplois anaphoriques. Est abordée également la modification des SN définis. Le chapitre 6 porte sur la détermination possessive, généralement peu abordée comme cela est rappelé à juste titre.

Il serait long et fastidieux d'entrer dans le détail de cette synthèse foisonnante par le nombre de déterminants étudiés ainsi que par l'abondance des tests, des exemples et des tableaux présentés tout au long d'une démarche extrêmement méthodique.

Nous présenterons simplement quelques-uns des points forts et originaux de ce travail, qui complètent une littérature déjà très fournie sur le sujet.

Un premier point fort de l'ouvrage a trait à l'ouverture théorique ainsi qu'à la multiplicité des travaux et des approches qui y sont exploités ou cités, allant de la sémantique formelle à la théorie des opérations énonciatives en passant par la grammaire générative.

Corollairement – et c'est un deuxième point fort – la détermination n'est pas seulement traitée sous un angle syntaxique : la sémantique, et même la morphologie, avec les questions de figement et de composition, sont également partie prenante de la démarche de l'auteur.

Ceci expliquant cela, la « portée » de la détermination, si l'on peut dire, s'en trouve considérablement modifiée. En effet, P.-A. Buvet démontre qu'elle ne concerne pas exclusivement la classe de ce qu'on a pu traditionnellement dénommer « articles », mais qu'elle englobe aussi des adverbiaux (*trop de*), des noms (*tonne, flopée, armée*) ou des séquences verbales (*je ne sais quel, n'importe quel*). Qu'à côté des déterminants standard existent des séquences déterminatives complexes, par exemple figées (*un méchant* dans *un méchant rhume*), et que les formes sémantiques de la détermination sont plus variées qu'il n'est dit généralement dans les grammaires, qu'elles soient intensives ou aspectuelles (*un début de, un_naissant*). De là aussi vient que la place des déterminants au sein d'un SN est autrement plus variable qu'il n'est généralement dit. Par ailleurs, ce que met en évidence cet ouvrage est l'interaction étroite entre la détermination et les propriétés des N actualisé : « *Les propriétés des noms déterminés sont le plus souvent des facteurs qui sont minimisés, voire totalement négligés dans les nombreux travaux qui leur sont consacrés* ».

Cela n'est pas nouveau dans la mesure où le facteur nominal ou lexical intervient, comme on sait, dans les oppositions traditionnelles entre massif et comptable ou concret et abstrait. Mais celles qui se manifestent ici engagent des sous-classes de N plus fines (les noms de vêtements, de sports, de maladies, d'affects et les noms humains, notamment) et, avec elles, un faisceau de contraintes extrêmement « sophistiquées ».

Bref, telle qu'elle ressort de cet ouvrage, la détermination apparaît de nature à renouveler profondément les descriptions lexicographiques, qui évoquent, par exemple, les emplois figurés de *tas* (*un tas d'ennuis*) ou *torrent* (*un torrent de larmes*) sans aller jusqu'à la détermination ou le contenu des grammaires, notamment scolaires. Dans le cadre méthodologique ambiant, elle pose aussi de sérieux problèmes pour les annotations dites de haut niveau ou l'identification automatique. Mais, plus largement encore, ainsi reconfigurées, les frontières de la détermination obligent à reconsidérer, peut-être même à redessiner, celles des catégories (adjectives, nominales, etc.) qui y sont impliquées. C'est dire si le sujet intéresse non seulement les linguistes de tous bords, mais aussi les didacticiens, les informaticiens ou encore les traducteurs.

Laurent GOSSELIN, Yann MATHET, Patrice ENJALBERT, Gérard BECHER. Aspects de l'itération. L'expression de la répétition en français : analyse linguistique et formalisation. *Peter Lang*. 2013. 372 pages. ISBN 978-3-0343-1415-2.

Lu par **Natalia Grabar**

UMR 8163 Savoirs, Textes, Langage STL – Université de Lille *3*

L'ouvrage regroupe trois contributions dédiées à la formalisation de l'itération temporelle. Il s'agit d'un ouvrage interdisciplinaire, où les auteurs, les chercheurs en linguistique (sémantique formelle), informatique et logique, se proposent de modéliser la notion de l'itération temporelle. Comme la réflexion a été faite dans le cadre d'un projet, les chercheurs se focalisent sur des notions identiques ou proches et il existe également des renvois entre les contributions. Le travail est effectué avec des exemples langagiers, réels ou jouets. En revanche, le lien avec le TAL et le traitement automatique de corpus est absent excepté la mention d'une thèse d'informatique non présentée dans l'ouvrage.

L'itération temporelle est la répétition dans le temps d'un même procès, en sachant que la *répétition dans le temps* sous-entend que les intervalles de procès correspondants ne coïncident pas (il y a au moins une succession des bornes initiales de ces intervalles) et qu'un *même procès* sous-entend, quant à lui, qu'un événement unique corresponde à chaque occurrence de ce procès. La modélisation de cet objet a été donc soumis à trois disciplines. En plus de la modélisation, chaque contribution propose également une visualisation des procès et de leurs itérations. Les modélisations proposées par chaque discipline impliquée se fondent sur les travaux existants tout en proposant des développements nouveaux.

Dans la première contribution dédiée au modèle linguistique, il est indiqué que l'on distingue traditionnellement deux aspects de l'itération : l'aspect lexical (marqué par les lexèmes verbaux, qui permettent de construire le procès) et l'aspect grammatical (exprimé par les conjugaisons de ces verbes, qui permettent d'exprimer la façon de voir ce procès du point de vue aspectuel). Ici, l'auteur propose d'abandonner cette dichotomie du fait qu'elle ne représente pas correctement la situation en français et n'est pas généralisable à d'autres langues. L'auteur propose

plutôt une opposition sémantique en distinguant l'aspect conceptuel (fondé sur le processus sémantico-cognitif de construction des procès par catégorisation) et la visée aspectuelle (opérant sur la présentation des procès préalablement construits par l'intermédiaire des intervalles). Ceci permet d'avoir un point de vue plus global et complet des procès et de leurs itérations. Parmi d'autres notions linguistiques analysées se trouvent, par exemple, les types de procès (état, activité, accomplissement et achèvement), les phases de procès (préparatoire, initiale, médiane, finale et résultante), les visées aspectuelles de procès (aoristique, inaccomplie, accomplie et prospective), la coupure modale (faisant la distinction entre les procès réalisés et irrévocables et les procès futurs et possibles), les types de compléments circonstanciels temporels (de durée, comme *pendant deux heures, en une semaine*, et de localisation temporelle, comme *en 2014, un jour, lorsque je lisais l'ouvrage*), la portée des circonstanciels, etc. Un autre point important de ce modèle est que l'itération des procès est envisagée de manière compositionnelle, où le calcul de la sémantique globale d'un procès résulte de la sémantique de ses composants. Cependant, l'auteur propose que la compositionnalité soit non pas atomique, mais holiste, où la sémantique des procès est vue dans leur globalité et respecte autant la sémantique des marqueurs individuels que les relations entre ces marqueurs et les connaissances encyclopédiques et pragmatiques. De la même manière, ce modèle permet d'agglomérer les procès et de résoudre les conflits, comme dans le cas de trois marqueurs impliqués dans la phrase « *Depuis deux mois, il mangeait en 10 minutes* » *:* imparfait *mangeait* inaccompli, *en 10 minutes* aoriste et *depuis deux mois* compatible uniquement avec les aspects accomplis et inaccomplis. Le modèle qui s'inscrit dans les travaux en informatique est fondé sur les principes de la programmation objet et des espaces mentaux. Ainsi, les itérations de procès sont caractérisées, par exemple, par les objets (qui correspondent aux procès) et les classes d'objets, les relations entre les classes (associative, d'héritage), les itérateurs déclenchés par les marqueurs linguistiques, les relations générales entre les procès (temporelle, causale et méronimique), les relations temporelles entre les procès (concomitance et succession), la récursivité comme le mécanisme de l'itération, et la sélection pour introduire des contraintes (conditions, exceptions).

Le troisième modèle de l'itération repose sur les principes algébriques et logiques, avec une attention particulière portée à la quantification et la pluralité nominale. Les postulats logiques sont posés et permettent de prendre en charge les notions comme le temps, l'intervalle, les restrictions, les ensembles, les composants ou les relations (hiérarchie, inclusion, mesure et topologie). Comme le travail est effectué avec des données langagières, une description spécifique est proposée pour les différents types de quantifieurs : les déterminants (*e.g., les, tous les, chaque, un, un certain, la plupart des, presque tous les, certains, quelques*), les expressions de quantification explicite (*e.g., trois jours par mois, trois fois par jour, deux mois sur douze*), et les compléments (les heures, les intervalles).

Les chercheurs s'attaquent à des notions liées à l'itération de procès assez complexes à décrire et à représenter. Les notions abordées peuvent être proches (agglomérat, modèle, ensemble et série de procès) ou identiques (quantification, relations entre les éléments itérés, etc.).

Cet ouvrage peut être intéressant pour les chercheurs en TAL travaillant sur la temporalité. Bien que le lien avec les travaux autour de la détection des expressions temporelles, de la norme TimeML et de la construction de lignes de vie et de lignes temporelles ne soit pas établi, les modèles proposés dans l'ouvrage peuvent aider dans le calcul et la représentation des lignes temporelles et des relations entre les procès. De même, les auteurs donnent quelques pistes pour la résolution de conflits entre les représentations sémantiques des expressions temporelles.

Kevin BRETONNEL COHEN, Dina DEMNER-FUSHMAN. Biomedical Natural Language Processing. *John Benjamins publishing company*. 2013. 160 pages. ISBN 978-9-0272-4998-2.

Lu par **Thierry HAMON**

Université Paris-Nord - LIMSI – UPR 3251 – Orsay

Cet ouvrage présente l'état de l'art du traitement automatique des textes biomédicaux en anglais, c'est-à-dire de la littérature scientifique produite par les biologistes et les médecins, ainsi que des textes cliniques décrivant le parcours de santé des patients hospitalisés. Ce livre est destiné aux chercheurs en TAL qui désirent s'intéresser à l'analyse de ce type de documents textuels et passe en revue les différents contextes applicatifs dans lesquels le TAL peut être mis en œuvre. L'ouvrage est composé de onze chapitres assez courts dédiés à la résolution d'un problème applicatif particulier (extraction d'information, recherche d'information, etc.). À l'exception des deux premiers, chaque chapitre est structuré de manière similaire : la thématique abordée est justifiée du point de vue biomédical ; les problématiques et les difficultés liées à la thématique, au domaine et aux types de textes sont mises en avant ; certains travaux ou outils dédiés à cette thématique sont décrits avec plus ou moins de détails.

Le premier chapitre est une courte introduction rappelant les notions de TAL pertinentes pour le traitement de corpus biomédicaux. C'est aussi l'occasion pour les auteurs d'introduire les types de données textuelles disponibles dans le domaine et d'évoquer les difficultés liées à l'analyse de ces textes et les solutions généralement mises en œuvre.

Dans le chapitre 2, les auteurs font un historique du TAL dans le domaine biomédical. Ils présentent ainsi les outils marquant les évolutions majeures du domaine, mais aussi les ressources, les corpus et les portails disponibles (ces portails étant souvent des sources pour la constitution de corpus textuels). Le chapitre se termine par une description des problèmes éthiques et légaux liés aux données cliniques.

Le chapitre 3 est consacré à la reconnaissance d'entités nommées dans les textes biomédicaux. Après avoir montré l'importance de cette tâche dans le domaine, mais aussi les difficultés à identifier les entités nommées dans les textes (à la fois pour des raisons d'ambiguïté, de polysémie et de métaphore), les auteurs présentent deux systèmes typiques : l'un s'appuyant sur des règles pour reconnaître des noms de gènes, l'autre utilisant des informations statistiques pour apprendre des patrons

d'identification des noms de maladies. Le chapitre se conclut sur la manière d'évaluer les approches proposées et les collections de données disponibles pour une telle évaluation.

Le chapitre 4 s'intéresse à l'extraction de relations sémantiques. Les auteurs se placent d'emblée dans une perspective d'extraction d'information. L'extraction des relations est considérée comme le moyen de remplir des formulaires permettant de décrire une maladie ou un gène avec des informations associées. Ainsi, pour identifier des réseaux génomiques à partir de textes de biologie ou pour décrire cliniquement les patients à travers la fouille de textes cliniques, les méthodes présentées s'appuient aussi bien sur des méthodes par apprentissage que sur des approches à base de règles pour répondre à cette problématique. Comme le chapitre précédent, ce chapitre se termine par une évocation de problématiques liées à l'évaluation.

Le chapitre 5 est consacré à la recherche d'information, principalement dans la littérature biomédicale. Une première partie du chapitre décrit la base PubMed/Medline sur laquelle portent de nombreux travaux. La recherche d'information dans les textes biomédicaux s'appuyant naturellement sur les connaissances du domaine pour réduire les problèmes de polysémie et d'ambiguïté sémantique, le métathésaurus UMLS, qui regroupe plus d'une centaine de ressources terminologiques biomédicales, est considéré comme une ressource primordiale dans le domaine biomédical. Cependant, l'utilisation de l'UMLS n'est pas une solution à ces problèmes : pris dans son ensemble, l'UMLS contient lui-même des termes polysémiques ou ambigus étant donné qu'il a été constitué à partir de nombreuses ressources terminologiques. Des travaux se sont donc intéressés à l'exploitation des informations issues de l'UMLS pour améliorer la recherche d'information (identification des concepts de l'UMLS pertinents pour les requêtes, utilisation de synonymes). Les auteurs abordent aussi cette problématique dans le contexte plus particulier de l'interrogation de bases de connaissances biologiques et de la désambiguïsation des noms de gènes. Enfin, les auteurs présentent quelques travaux liés à la mise en œuvre de systèmes de recherche sur du texte plein, des images, des figures et des légendes de figures.

Dans la continuité du chapitre précédent, le chapitre 6 aborde la problématique de la normalisation de concepts, utile à la fois en extraction et en recherche d'information. Deux aspects sont présentés : la normalisation de noms de gènes et la normalisation de termes issus notamment de textes cliniques. Dans le premier cas, il s'agit d'associer les variantes de noms de gènes que l'on peut trouver dans la littérature scientifique biomédicale, avec la forme normalisée présente dans la base « Entrez Gene » (base recensant les informations liées aux gènes). Après avoir présenté la problématique, les auteurs décrivent les solutions possibles et notamment la mise en œuvre du système GNAT. Dans le second cas, il s'agit de la normalisation de termes issus des textes cliniques, et de l'identification des concepts de l'UMLS. Cela nécessite la reconnaissance de termes associés aux concepts mais aussi de leurs variantes malgré les difficultés liées aux ambiguïtés et à la polysémie. Le système MetaMap développé par la NLM (National Library of Medicine) y est décrit. Les auteurs présentent également ses limites et ses évolutions possibles.

Le chapitre 7 s'inscrit dans la continuité du précédent. Après avoir décrit en détail l'UMLS et Gene Ontology, les auteurs montrent comment le TAL peut être utilisé pour reconnaître dans les textes, les termes issus de ces deux ressources, pour vérifier la qualité des ressources (pour ajouter des relations manquantes par exemple), pour aligner des ontologies ou pour les mettre en relation. Ici aussi, les auteurs illustrent leur propos en décrivant en détail des méthodes dédiées.

Le chapitre 8 s'intéresse au résumé automatique de textes biomédicaux à travers la description d'un système de résumé multidocument destiné à synthétiser la littérature médicale. L'état de l'art de cette thématique se poursuit par la présentation de systèmes dans le domaine de la génomique pour enrichir des bases de données avec des connaissances sur les gènes, c'est-à-dire des descriptions en langue naturelle, des fonctions des gènes ou des interactions de protéines.

Le chapitre 9 est consacré aux systèmes de question-réponse. Après un rappel des principes de base de ces systèmes et des problématiques particulières concernant l'interrogation de la littérature médicale (notamment dans le cadre de la médecine fondée sur les faits – *evidence based medicine*) et celle d'articles de génomique, les auteurs décrivent en détail les différentes étapes de mise en œuvre d'un système dédié à l'interrogation de la base PubMed.

Le chapitre 10 est une réflexion originale sur les méthodes de génie logiciel devant être mises en place dans le contexte du TAL et sur les problématiques d'ingénierie, en particulier pour l'analyse de la qualité des outils de TAL dans le domaine biomédical. La première partie du chapitre est un rappel des techniques classiques de vérification de code, tandis que la seconde partie est consacrée à la mise en œuvre de ces techniques lors du développement d'approches ou d'applications de TAL.

Dans le dernier chapitre, les auteurs décrivent succinctement la problématique de la constitution et l'annotation de corpus dans le domaine biomédical. La majorité du chapitre est consacré à la description des corpus annotés disponibles dans le domaine. On peut regretter qu'une partie du chapitre ne soit pas consacrée à la définition de guide de définition ou d'annotation de corpus.