Notes de lecture

Rubrique préparée par Denis Maurel

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

François RASTIER, La Mesure et le Grain. Sémantique de corpus, *Honoré Champion*, 2011, 272 pages, ISBN 978-2-7453-2230-2.

Lu par **Amal GUHA**

Laboratoire MoDyCo - UMR 7114, Université Paris Ouest Nanterre La Défense

Cet ouvrage de François Rastier présente une vue articulée de différents domaines de la linguistique – y compris le traitement automatique du langage – auxquels l'auteur a contribué. Par ses travaux antérieurs, la linguistique de corpus est l'angle d'attaque permettant des propositions, aussi bien épistémologiques qu'applicatives. François Rastier présente des cas concrets de méthodes applicatives, mais discute aussi de critères précis pour les études textométriques. À rebours de l'approche « logico-grammaticale » dominante dans le domaine du TAL, il propose une approche de la sémantique s'appuyant sur les corpus, en donnant sa place à la philologie, et à nombre de variables liées au document dans ce qu'il a de matériel. L'auteur fait l'inventaire des limites et des préjugés largement partagés du domaine, notamment en discutant en détail des ontologies, et du Web sémantique. La problématique des corpus lui permet de proposer une approche philologique-herméneutique, qui tient compte des problématiques des textes et des documents.

Contenu de l'ouvrage

Ce livre comporte neuf chapitres, une bibliographie, un index des auteurs et la référence à un glossaire de sémantique interprétative en ligne.

Le premier chapitre résume les enjeux épistémologiques de la linguistique de corpus. Les corpus permettent d'aborder la langue par la parole. Un texte est écrit non seulement « dans une langue », mais dans un genre.

Le chapitre 2 propose un modèle du texte comme unité minimale, sinon élémentaire. Il présente et discute les modèles existants, hiérarchiques (notamment la TEI, *Text Encoding Initiative*), et réticulaires (non hiérarchiques). S'intéresser à la philologie permettrait de clarifier l'usage actuel confus de la notion de « métadonnées ». Les systèmes informatiques peuvent traiter plus facilement des critères relevant du document, qui font partie des « nouveaux observables » produits par la linguistique de corpus.

Le chapitre 3 présente une étude empirique de classification de textes. Une classification automatique non supervisée a confirmé une classification préalable des textes en genres.

Le chapitre 4 (doté de trois annexes) aborde l'expression des représentations collectives à travers la langue, la doxa.

Le chapitre 5 propose de reconsidérer le problème de la polysémie d'acceptions, car les acceptions n'entrent pas en concurrence au sein des mêmes genres. La seconde partie du chapitre étudie l'innovation sémantique dont témoignent les emplois, et propose de la nommer *néosémie*.

Le chapitre 6 présente une étude, relevant de la linguistique comparée, visant à caractériser de façon contrastive les discours (scientifique, littéraire, philosophique) et les genres.

Le chapitre 7 (doté d'une annexe) présente les réflexions élaborées en marge du projet européen Pincip.net, mené par Mathieu Valette. Il visait à la détection automatique de sites racistes, en intégrant tous les paliers de la complexité textuelle, au-delà des systèmes ordinaires fondés sur des mots-clés.

Le chapitre 8 présente et décortique le Web sémantique. Il plaide pour une « sémantique du Web » qui échapperait aux postulats métaphysiques d'une problématique de représentation des connaissances, et la ferait évoluer vers la *production* de connaissances.

Le chapitre 9 présente la sémantique de corpus comme une « sous-discipline auxiliaire des sciences de la culture ». Il discute d'abord en détail des conséquences pour les études littéraires des avancées permises par la linguistique de corpus. Le chapitre – et le livre – concluent sur un plaidoyer pour l'engagement scientifique, non au sens étroit de l'engagement politique du scientifique, mais au sens de la lutte contre l'ignorance et de la satisfaction des besoins sociaux.

Commentaire

La problématique du corpus permet à l'auteur de confronter à chaque étape des pans distincts présentés par certains de ses ouvrages antérieurs, dont les problématiques se recouvraient peu, depuis ses premiers travaux sur l'idéologie, jusqu'à *Sémantique interprétative* (1987) et surtout *Arts et sciences du texte* (2001), en passant par *Sémantique et recherches cognitives* (1991). Nous retenons deux thématiques transversales aux domaines abordés, qui ont accompagné notre lecture : la question de l'unité pertinente et l'approche philologique.

La question de l'unité pertinente est sous-jacente à la plupart des problématiques abordées : la place prise par le palier du lexique se situe dans l'héritage, issu à la fois d'une tradition « logico-grammaticale » de la langue, et d'une tradition documentaliste plus récente. Il convient de ne pas nier cette place, et d'employer les ressources élaborées à cette échelle. De plus, l'échelle des segments entre deux espaces est plus commode pour les traitements informatisés. François Rastier illustre

ceci par nombre de références à la lexicométrie lorsqu'il appuie empiriquement ses thèses. Toutefois, il propose le texte comme unité fondamentale (si ce n'est « minimale ») de la linguistique : c'est l'échelle qui rend mieux compte de la position de pivot entre les pratiques et le système de la langue. François Rastier discute, en de nombreux endroits, des entreprises qui s'appuient sur des ontologies hors contexte : elles sont vouées à l'échec, puisqu'elles commencent par « décontextualiser », c'est-à-dire par priver la parole du contexte qui permet son interprétation. La question de la polysémie apparaît alors comme un artefact de ce découpage inadéquat (lequel ne permet d'ailleurs pas un traitement pertinent du phénomène de la néosémie). Il discute spécifiquement des ontologies et du Web sémantique aux chapitres 2 et 8.

L'appel à recourir à la tradition herméneutique devrait être naturellement entendu par le TAL, qui s'est jusqu'ici surtout appuyé sur une linguistique chomskyenne, détachée des textes. Les corpus seront peut-être l'occasion d'une prise en compte raisonnée des documents qui portent les textes : le TAL, compris comme technologie, non comme science, peut renoncer à la prétention à expliquer tous les faits de langage, pour s'attacher à la réalité philologique.

Il est à remarquer que l'informatique peut tirer parti de la mise au premier plan des pratiques : en effet, il s'agit, pour toute application spécifique, de répondre à un besoin particulier ; il n'est pas indispensable alors de se priver de tous les éléments de contextualisation, de traitements spécifiques au genre et aux documents, desquels on peut tirer parti pour répondre à ce besoin.

François Rastier illustre par cet ouvrage la vitalité du courant « néo-Saussurien » (qu'il a contribué à fonder), en mettant l'accent sur une démarche différentielle et contrastive, en sémantique en particulier. Il situe la place de la linguistique au sein des sciences de la culture : science à la position intermédiaire, qui non seulement cherche à décrire des lois générales, mais permet aussi de rendre compte de la singularité des textes.

Les vues de François Rastier à propos des attendus épistémologiques qui sous-tendent la plupart des travaux de TAL sont souvent d'une concision radicale. Cependant, il prend le soin de les argumenter. Il expose dans cet ouvrage sa compréhension de la science linguistique, comme une partie des sciences de la culture. Le remembrement suggéré s'articule, par l'usage des corpus, autour des textes, interfaces entre la langue vue comme système, et les performances. De façon analogue, ce livre est une invitation pour le TAL à inclure dans ses outils des éléments qui relèvent de la pratique. La prise en compte des genres textuels impliqués, et de la singularité de chaque besoin applicatif permet, pour une application donnée, de limiter l'ambition théorique à la généralité héritée de la tradition du TAL. Elle permet de tirer parti, pragmatiquement, de « nouveaux observables ».

Gérard LIGOZAT, Raisonnement qualitatif sur le temps et l'espace, *Hermès-Lavoisier*, 2011, 560 pages, ISBN 978-2-7462-3117-7.

Lu par **Philippe MULLER**

Université Paul Sabatier, Toulouse – IRIT

Ce livre présente un ensemble de formalismes de représentation d'informations temporelles et spatiales, et les moyens de raisonner sur ces informations. L'accent est mis non pas sur des données numériques mais sur des cadres où l'information est essentiellement « qualitative », c'est-à-dire ici relationnelle : savoir qu'un événement a lieu avant un autre, pendant un autre, ou bien que des régions de l'espace sont dans des relations d'inclusion, de contact, à gauche l'une de l'autre, etc. L'intérêt dans une perspective de traitement du langage naturel est de fournir une sémantique formelle précise à un ensemble de concepts importants dans la langue, et de donner des outils utilisables quand il s'agit de faire des inférences sur ces données. Il s'agit ici de considérer des représentations à l'expressivité bien calibrée, pour lesquelles des procédures opératoires sont réalisables, d'où l'insistance sur des problèmes de satisfaction de contraintes binaires. On en voit notamment l'importance dans les efforts de normalisation de l'annotation d'informations temporelles, avec la norme ISO-TimeML, et d'informations spatiales et spatio-temporelles avec la norme en cours d'élaboration ISO-Space.

Gérard Ligozat a contribué de façon majeure à ce domaine de recherche, à la fois sur l'étude des formalismes temporels et spatiaux, ainsi qu'à la généralisation des calculs sur des systèmes de contraintes relationnelles. Ce livre est donc une somme, qui couvre de façon très complète les modèles utilisables et leurs propriétés formelles et opératoires. Il comprend aussi un survol très rapide des applications de ces formalismes, notamment dans le cadre du traitement automatique des langues.

Les trois premiers chapitres se concentrent sur le temps, à travers l'étude des relations entre intervalles de temps, et plus particulièrement les relations dites « de Allen » (chapitre 1), les raisonnements praticables sur ces données (chapitre 2), et une extension du formalisme aux intervalles « généralisés », c'est-à-dire qui expriment des relations entre événements répétés (chapitre 3). Les chapitres 4 et 5 portent sur d'autres formalismes qualitatifs appliqués à des relations spatiales (topologique, d'orientation, de distance entre entités spatiales). Les chapitres 6 et 7 montrent les pistes existantes pour ajouter des données plus quantitatives aux informations symboliques, pour prendre en compte les durées, par exemple, ou bien l'incertitude de l'information représentée. Le chapitre 8 porte sur les formalismes qualitatifs dans le domaine conceptuel, où la géométrie concerne les liens entre concepts, une approche popularisée par Peter Gärdenfors. Les chapitres 9 à 13 reprennent plus en détail certaines propriétés formelles des théories présentées, qui permettent de mieux caractériser leurs modèles logiques, les équivalences structurelles entre théories (en théorie des catégories ou en logique modale), et la

complexité des raisonnements que l'on peut faire, en général cette fois-ci, sur les contraintes qualitatives. Le livre termine sur un inventaire des applications pertinentes, et sur les perspectives actuelles de la recherche en raisonnement qualitatif, notamment l'épineux problème de la combinaison de formalismes différents, souhaitable pour atteindre des descriptions complémentaires (par exemple de relations spatiales et temporelles dans le cas du mouvement), mais difficile formellement et pratiquement.

On ne peut qu'être très impressionné par l'expertise qui a été rassemblée ici, et le degré de précision et de rigueur dans la description des théories et de leurs propriétés, ainsi que de l'effort constant de généraliser les travaux et de faire des ponts formels entre les différents formalismes. On a incontestablement ici un livre de référence pour le domaine, et il est dommage qu'il ne soit écrit qu'en français (pour l'instant ?). Les nombreux exemples et figures qui illustrent les problèmes abordés facilitent grandement la compréhension des aspects techniques.

Maintenant, si l'on se place du point de vue du traitement des langues, il faut bien admettre que le lien fait entre les formalismes qualitatifs et les problématiques de TAL est réduit à la portion congrue. Je mentionne plus haut les efforts de normalisation des annotations, car elle souligne l'importance des données temporelles et spatiales dans des tâches de TAL, comme l'extraction d'information, les systèmes de question-réponse, le résumé, ou bien le dialogue homme-machine. Il est d'autant plus dommage de passer si vite sur ces applications, que l'on voit arriver depuis quelques années des travaux qui soulignent l'importance d'avoir un modèle formel qui permet de raisonner sur des données, notamment temporelles, pour contrôler la cohérence d'un ensemble d'informations. Cet aspect est crucial dans le cas d'annotations manuelles, mais également pour guider des systèmes automatiques, voire améliorer leurs résultats. Les résultats sont encore loin d'être stables, et les outils décrits dans ce livre restent encore assez diversement connus dans la communauté TAL. Pour cette raison, les deux premiers chapitres, sur le raisonnement temporel, qui atteignent déjà une centaine de pages, sont une lecture indispensable à qui s'intéresse à l'expression du temps dans une perspective computationnelle. Les traitements de l'expression de l'espace connaissant un développement important, on peut considérer que les parties qui traitent de raisonnement spatial auront la même importance.