# Notes de lecture

Rubrique préparée par Denis Maurel

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

**Claudia LEACOCK, Martin CHODOROW, Michael GAMON, Joel TETREAULT, Automated Grammatical Error Detection for Language Learners, *Morgan & Claypool Publishers*, 2010, 122 pages, ISBN 9781608454709.**

Lu par **Claire GARDENT**

*CNRS/LORIA, Nancy*

*Ce livre présente un survol des travaux portant sur la détection automatique d'erreurs faites par les apprenants d'une langue étrangère. L'ouvrage, assez bref, se concentre sur les erreurs les plus courantes pour les apprenants de l'anglais, à savoir, le choix des prépositions, des articles et des collocations. Il décrit les grands types de corpus et les techniques développées pour détecter ces erreurs. Il introduit également les schémas d'annotation et les méthodes d'évaluation mises en jeu. C'est un ouvrage un peu rapide mais qui donne un bon aperçu des méthodes utilisées et des grandes problématiques du domaine (notamment des difficultés soulevées par la création et l'annotation de corpus utilisables pour l'apprentissage et l'évaluation).*

Le premier chapitre définit la notion d'erreur concernée, la différenciant des erreurs grammaticales ou d'erreurs faites par les locuteurs natifs. Contrairement à ces dernières, les erreurs faites par les apprenants sont en effet pour une bonne part, des erreurs d'usage (comme par exemple, le choix en contexte du bon déterminant ou de la bonne préposition) ou des erreurs résultant soit d'une confusion entre homophones (*e.g.*, « *its* » *vs* « *it's* ») , soit de l'application inappropriée d'une règle morphosyntaxique (*e.g.*, « *He writed* »).

Le deuxième chapitre dresse un historique rapide de la détection d'erreurs et de la complémentarité entre techniques symboliques et approches statistiques. Il explique en particulier, comment l'analyse syntaxique symbolique permet le développement d'approches pour la vérification grammaticale qui fonctionnent relativement bien sur des textes écrits par des locuteurs natifs mais achoppent sur les textes d'apprenants où les erreurs au sein d'une même phrase sont souvent nombreuses. Les approches statistiques, plus robustes, sont plus appropriées et sont couramment utilisées pour détecter les erreurs d'usage. Enfin, parce que la complexité des erreurs à traiter est très variable, les systèmes d'aides aux apprenants

mettent généralement en œuvre une approche hybride combinant un système de règles ou patrons pour les erreurs « hors contexte » dont la détection est simple, des méthodes d'apprentissage automatique pour les erreurs plus complexes telles que les erreurs d'usage, et des méthodes symboliques pour la détection d'erreurs grammaticales « longue distance » tel l'accord sujet-verbe.

Le troisième chapitre revient sur les types d'erreurs faites par les apprenants de l'anglais. Diverses études et analyses de corpus sont passées en revue mettant en avant les différences observées entre les erreurs faites par les locuteurs natifs et celles faites par les apprenants. L'impact de la langue maternelle sur les types d'erreurs faites par les apprenants est brièvement discuté. Enfin, les difficultés posées par l'apprentissage des prépositions, des déterminants et des collocations sont illustrées et explicitées.

Le chapitre 4 porte sur les corpus utilisés pour évaluer et entraîner les systèmes de détection d'erreurs : corpus d'apprenants, corpus de locuteurs natifs et corpus artificiels créés par des méthodes automatiques sur des données de grande taille dans lesquelles les erreurs des apprenants sont simulées à partir de textes bien formés. Les corpus d'apprenants sont listés dans deux sous-sections distinguant les corpus d'apprenants de l'anglais des corpus d'apprenants d'autres langues que l'anglais. L'utilisation de corpus artificiels et de corpus de textes bien formés pour l'évaluation et/ou l'apprentissage de systèmes de détection d'erreurs est ensuite motivée.

Faisant pendant au chapitre 4, qui détaille et justifie les différents types de corpus exploités, le chapitre 5 explicite les problématiques soulevées par l'évaluation des systèmes de détection d'erreurs sur ces différents types de corpus. Après avoir introduit les métriques utilisées, les auteurs recensent les avantages et les inconvénients de deux types d'évaluations possibles : évaluation sur un corpus de textes bien formés *vs* évaluation sur un corpus d'apprenants. Les corpus de textes bien formés présentent l'avantage d'être faciles d'accès et de permettre le développement et la comparaison de systèmes. Cependant, ils entraînent une surestimation de la précision si bien que les résultats acquis sur ces corpus restent de pauvres indicateurs des performances des systèmes de détection d'erreurs en conditions réelles, c'est-à-dire sur des corpus d'apprenants. En revanche, l'évaluation sur un corpus d'apprenants demande un effort important d'annotation et aboutit à une sous-estimation de la précision (seule la solution annotée est considérée comme valide alors que plusieurs solutions sont souvent possibles). Pour pallier ces inconvénients, une solution intermédiaire consiste à utiliser des corpus ciblant l'annotation de certains types d'erreurs (par exemple, choix des prépositions). La discussion de ces corpus est cependant reportée par les auteurs au chapitre 9. Bizarrement, l'évaluation sur le troisième type de corpus, les corpus artificiels, n'est pas abordée.

Les trois chapitres suivants entrent dans le vif du sujet et décrivent la détection et la correction d'erreurs pour les articles et les prépositions (chapitre 6), pour les collocations (chapitre 7) et pour diverses autres types d'erreurs comme la confusion d'homophones, les fautes d'orthographe et les fautes d'accord et de conjugaison (chapitre 8). Pour l'essentiel, les traitements des erreurs de déterminants, prépositions et collocations mettent en jeu des méthodes statistiques (classification, mesures d'association, modèles de langage). Les auteurs présentent brièvement quelques-unes de ces méthodes puis listent et commentent les méthodes et systèmes existants. Ils présentent également les systèmes à base de règles et d'heuristiques qui sont généralement utilisés pour les divers types d'erreurs abordées dans le chapitre 8.

Le chapitre 9 revient sur la question de l'annotation des erreurs d'apprenants. Les auteurs commencent par rappeler les difficultés inhérentes à l'annotation de corpus d'apprenants (accords interannotateurs bas, multiplicité des corrections possibles, fréquences relativement basses des phénomènes à annoter), puis c'est au tour des schémas d'annotation développés pour l'enseignement et la recherche linguistique qui contrastent avec ceux proposés plus récemment par la communauté TAL pour entraîner et évaluer des approches statistiques. Tandis que les premiers visent l'annotation de toutes les erreurs, les seconds ciblent généralement un type restreint d'erreurs, souvent un seul type comme par exemple les erreurs d'usage dans le choix du déterminant. Les auteurs commentent ensuite deux méthodes permettant de réduire les coûts d'annotation : les méthodes d'échantillonnage et l'utilisation, maintenant assez courante, de *Amazon Mechanical Turk*.

Le chapitre 10 ouvre la discussion sur les perspectives de recherche actuelles pour le domaine de la détection et la correction des erreurs d'apprenants traitant en particulier, de l'utilisation du Web et de la traduction automatique ainsi que de l'importance d'une évaluation « par la tâche » qui permettrait de répondre à la question : « Les systèmes de détection et de correction d'erreurs permettent-ils à l'apprenant de s'améliorer ? ».

L'ouvrage se termine (chapitre 11) par un bref rappel des thèmes traités et une liste de questions ouvertes : développement de « *benchmarks* » communes permettant d'évaluer et de comparer les systèmes développés ; détection d'erreurs pour l'apprentissage de langues autres que l'anglais ; étude d'erreurs autres que celles les plus étudiées (choix des déterminants, des prépositions et des collocations) ; développement de systèmes prenant en compte la langue maternelle de l'apprenant ; collaboration accrue avec le corps enseignant.

« *Automated Grammatical Error Detection for Language learners* » est une bonne introduction à la problématique de la détection d'erreurs dans les textes d'apprenants. La couverture est large, la bibliographie abondante et le niveau de détail approprié pour le public visé (enseignants et chercheurs désireux de s'approprier les grands thèmes du domaine). Une présentation plus conceptuelle

aurait cependant rendu l'ouvrage plus attrayant avec notamment, une présentation détaillée des techniques utilisées et plus d'exemples illustrant les erreurs, les corpus, les méthodes et les systèmes présentés. En particulier, l'intégration, dans le corps des chapitres, de longues listes de systèmes et de méthodes très similaires rend le texte monotone. Ceci aurait facilement pu être évité en utilisant le corps du chapitre pour présenter les méthodes et la dernière section pour lister l'existant. De même le chapitre 4 consiste pour l'essentiel en une longue liste de corpus qui aurait été mieux présentée sous forme d'annexe.

---

**Fiammetta NAMER, Morphologie, lexique et traitement automatique des langues, l'analyseur DériF, *Hermès-Lavoisier*, 2009, 448 pages, ISBN 978-2-7462-2363-9.**

Lu par **Natalia Grabar**

*Affiliation* CNRS UMR 8163 STL, Université Lille 3, 59653 Villeneuve d'Ascq

---

*L'ouvrage de Fiammetta Namer, intitulé* Morphologie, lexique et traitement automatique des langues*, nous conte une heureuse et productive rencontre entre la linguistique et le traitement automatique des langues dans le domaine de la morphologie et de l'acquisition automatique de lexiques sémantiques, tout en proposant une analyse systématique et critique. Le travail présenté montre une grande rigueur dans plusieurs de ses aspects. Tout d'abord, l'auteur ne se limite pas à la description de la morphologie et du lexique du français, mais propose une méthodologie et un outil opérationnel unique dans son genre, l'analyseur morphologique DériF. Lors du développement de DériF, l'auteur s'appuie sur la théorie linguistique issue des travaux de Danielle Corbin. Notamment, grâce à l'exploitation de cette théorie, le développement et l'évolution de l'outil reste cohérent et fluide, pratiquement sans remises en cause du modèle linguistique et de l'approche. De plus, l'outil a pu être adapté à la langue anglaise, ce qui souligne la solidité théorique de sa conception. Finalement, plusieurs tests et analyses sont menés pour vérifier le bon fonctionnement de DériF et de la théorie sous-jacente. L'ouvrage est organisé en trois parties principales, qui peuvent être lues dans l'ordre ou de manière indépendante. Tout au long de l'ouvrage le lecteur est guidé par des introductions et des bilans. Cet ouvrage peut être utilisé comme un support didactique en morphologie computationnelle par les étudiants et les enseignants.*

La première partie est consacrée à la genèse de l'analyseur morphologique DériF. Dans un premier temps, l'auteur décrit de manière très exhaustive des travaux et des théories morphologiques. Le lecteur peut découvrir en particulier une excellente présentation de l'état de l'art relatif aux travaux en morphologie théorique et computationnelle. La richesse de cet état de l'art peut d'ailleurs rendre la lecture difficile à assimiler pour les novices, mais correspond à une étape nécessaire pour découvrir le domaine. Cette présentation peut servir de support didactique aux enseignants et étudiants. Dans un deuxième temps, le passage de la théorie à

l'application est décrit et justifié. Le lecteur peut alors découvrir le fonctionnement de l'analyseur morphologique DériF. Cet analyseur est dédié à l'analyse des lexèmes construits du français : il prend en entrée une liste de lexèmes munis de leurs catégories syntaxiques et, grâce aux règles morphologiques de formation de lexèmes, DériF produit en sortie une analyse morphologique de ces lexèmes, ainsi qu'une glose sémantique associée et calculée automatiquement à partir de l'analyse morphologique effectuée. Le fonctionnement de DériF repose en grande partie sur les travaux de morphologie lexématique de Danielle Corbin en linguistique théorique. Ce travail aborde des notions morphologiques importantes : règles morphologiques, contraintes qui pèsent sur ces règles, opérations sémantiques des affixes, compositionnalité morphologique et calcul de la sémantique. Le travail présenté dans cet ouvrage ne se limite pas à l'implémentation de cette théorie morphologique. L'auteur cherche en effet à étendre l'analyseur DériF à d'autres phénomènes morphologiques, comme la conversion et la composition, peu ou pas décrits dans les travaux théoriques originaux. Les deux parties suivantes décrivent, entre autres, ces études supplémentaires et, par là même, l'extension et la mise en épreuve de la théorie morphologique originale.

La deuxième partie est consacrée au calcul et à la vérification des contraintes sémantiques des règles morphologiques, qui sont décisives pour qu'une règle puisse être appliquée. Un formalisme pour le calcul automatique des traits sémantiques des lexèmes est mis en œuvre. Il est appliqué aux lexèmes construits et convertis et permet d'enrichir l'annotation de ces lexèmes. Il s'agit de calcul d'une grande variété de traits relatifs par exemple à l'aspect, au statut et à la sémantique des lexèmes de base et des lexèmes construits : verbes transitifs ou intransitifs, accomplis ou non, noms abstraits ou concrets, objet ou instruments ou patients, comptables ou non, animés ou non, d'action ou de résultat, adjectifs de relation ou de propriété, etc. L'auteur s'intéresse ensuite à la formation non affixale de lexèmes, c'est-à-dire à la conversion. Dans la suite d'autres travaux récents, l'accent est mis sur l'orientation de ces procédés. Là aussi, les propriétés et les contraintes sémantiques sont exploitées et permettent de faire des propositions pour les cas qui restent les plus problématiques (Nom/Verbe, Verbe/Nom). Un bilan sur les converts et les traits sémantiques qu'exigent les règles de conversion est effectué. Les expériences alors réalisées et présentées s'appuient sur les travaux de linguistes. Ces derniers sont utilisés comme une base théorique, ou bien sont testés pour confirmer ou infirmer ces propositions théoriques sur les données réelles et de grande taille. De vraies enquêtes et études sont alors menées afin de vérifier des règles et leurs propriétés sémantiques. Finalement, afin de présenter les informations sémantiques calculées automatiquement sous un format standard et exploitable par les outils automatiques, Fiammetta Namer convertit les sorties dans un format inspiré de celui du Lexique Génératif. Ceci permet une utilisation de lexiques sémantiques, de grande taille, extensibles et de format standard au sein de tout traitement sémantique de corpus écrits, quelle que soit la vocation de ce traitement (constitution de ressources de grande taille annotées sémantiquement, analyse de phénomènes

sémantiques à grande échelle, recherche d'information, classification de documents, fouille de textes, Web sémantique, etc.). Le lecteur s'aperçoit ainsi qu'une attention particulière est portée également au passage à l'échelle et à l'utilisation aisée des résultats produits par l'analyseur morphologique DériF.

La troisième partie est consacrée à l'adaptation de DériF au traitement de la composition dite savante ou néoclassique (*otorhinolaryngologiste, buccodentaire),* fréquente dans les domaines de spécialité, comme la médecine, la biologie, ou l'agronomie. Il s'agit d'une extension de la théorie linguistique originale, qui a nécessité également l'ajustement des principes de fonctionnement de DériF. Ainsi, une table spécifique a été créée, qui contient les éléments assimilables à des bases supplétives ou à des radicaux non autonomes de lexèmes. De plus, l'analyse des lexèmes a été ajustée afin de prendre en compte des schémas sémantiques spécifiques aux lexèmes composés. Finalement, le calcul de relations sémantiques propres aux instructions sémantiques des bases et des éléments de composition est ajouté. Dans cette partie, un état de l'art sur la composition morphologique est proposé d'une manière originale et exhaustive. Par exemple, le lecteur est invité à découvrir les liens entre les grammaires du sanskrit (formations morphologiques de classe *dvandva*, *bahuvrihi* ou *tatpurusha*) et les grammaires des langues européennes. Cet exposé peut également servir de support didactique aux étudiants et enseignants. Un autre point fort des travaux présentés dans cette partie correspond à la flexibilité de l'approche. D'un côté, l'adaptation de la théorie originale à la composition est effectuée. Mais comme cette adaptation ne permet pas d'analyser l'ensemble des lexèmes, l'auteur propose une analyse de ces lexèmes, appelés déviants ou récalcitrants, afin d'adapter la théorie morphologique. Cet aspect illustre en particulier l'importance qu'il faut accorder aux données nouvellement collectées pour envisager le prolongement et l'évolution possible des principes d'une théorie. D'un autre côté, l'analyseur est adapté aux données en langue anglaise.

En conclusion, cet ouvrage présente une très grande variété de travaux susceptibles d'intéresser les linguistes, les spécialistes en TAL et les informaticiens. Les spécialistes de différents domaines et travaillant en différentes applications (recherche d'information, fouille de textes, enrichissement et structuration de terminologies...) peuvent exploiter ce travail. L'intérêt que représente cet ouvrage réside autant dans les états de l'art présentés par l'auteur que par les implémentations réalisées et les expériences originales menées. Du point de vue théorique, cet ouvrage montre un bel exemple d'implémentation d'une théorie morphologique, son adaptation et ses extensions aux données réelles et nouvelles. De manière plus fondamentale, cet ouvrage illustre et souligne l'importance et la variété des informations sémantiques que la morphologie est susceptible de produire. Il présente les éléments qui peuvent servir à repositionner la morphologie au sein de la linguistique. Finalement, l'analyseur morphologique DériF est mis à disposition pour les personnes intéressées. Le travail présenté, étant testé sur des

données réelles et de grande taille, montre la robustesse et la richesse d'informations sémantiques que DériF est capable de générer.

---

**Michel ROCHÉ, Gilles BOYÉ, Nabil HATHOUT, Stéphanie LIGNON, Marc PLÉNAT, Des unités morphologiques au lexique, *Hermès-Lavoisier*, 2011, 343 pages, ISBN 978-2-7462-2986-0.**

Lu par **Fiammetta NAMER**

*Université de Nancy2 et UMR 7118 « ATILF » CNRS*

---

*Ce livre aborde différents aspects de la morphologie (flexionnelle et constructionnelle) du français, suivant une approche dont les choix principaux pourraient se résumer ainsi : la forme d'un mot morphologiquement complexe est le fruit de contraintes qui interagissent et dont le poids relatif varie en fonction d'un nombre varié de facteurs, ces contraintes mettent en jeu morphologie, lexique, (morpho)phonologie et leur identification requiert l'étude de données lexicales extensives. Les principes de cette approche originale et novatrice, présentés dans l'introduction de ce livre, sont illustrés dans les cinq chapitres qui constituent l'ouvrage : il y est question, successivement, de flexion verbale, des trois modèles sémantiques de noms en -isme et -iste, de l'effet sur la construction des mots de la contrainte phonologique de dissimilation, de la compétition entre les suffixes -ien et -éen, et d'un modèle théorique de construction des mots illustré par la préfixation en -anti.*

Ce livre est le fruit d'une collaboration entre cinq chercheurs (spécialistes en morphologie, (morpho)phonologie, romanistique, TAL) appartenant ou ayant appartenu à l'axe DUMAL (« Des unités morphologiques au lexique ») de l'UMR CLLE-ERSS à Toulouse. Il présente le fruit des derniers travaux entrepris par cette équipe, témoignant d'une conception paradigmatique de la morphologie, guidée par les données et fondée sur la résolution de compétition de contraintes.

Le premier chapitre sert de guide à la lecture de ce livre, dont il pose le contexte général. Il fait le bilan du cheminement théorique parcouru par la morphologie ces quarante dernières années et présente les choix principaux qui caractérisent l'approche suivie dans tous les chapitres du livre. Celle-ci peut se décliner sous la forme de trois aspects fondamentaux du courant théorique conçu, développé et défendu à Toulouse, et dont se font écho de plus en plus de chercheurs du domaine.

(i) Une théorie lexicale doit être guidée par les données. Cela se traduit par l'émergence d'un courant extensif de la morphologie. Grâce aux masses de données lexicales numérisées désormais disponibles (corpus, dictionnaires, mais surtout contenu de la Toile), un grand nombre d'anciennes certitudes sont balayées. Des faits qui passaient inaperçus, ou qui étaient cantonnés parmi les constructions exceptionnelles et « mal formées », s'observent maintenant avec suffisamment de

régularité pour que le morphologue soit obligé d'en tenir compte dans ses raisonnements.

(ii) La construction d'un mot obéit à deux points de vue complémentaires : le point de vue sémasiologique et le point de vue onomasiologique. L'étude des moyens morphologiques ne peut être dissociée de celle des besoins de nomination. Ceux-ci, par ailleurs, ne constituent pas le seul motif de la création lexicale : à la nécessité de donner un nom à un nouveau concept s'ajoutent des besoins d'ordre discursif, énonciatif et diaphasique ainsi que la nécessité d'ancrer une forme nouvelle dans un ensemble lexical bien identifié.

(iii) Le contenu et l'organisation du lexique existant chez un locuteur président à la construction et/ou à la compréhension des nouveaux mots ou des nouvelles formes d'un mot, qu'il énonce ou qu'il entend.

Ces trois principes engendrent un certain nombre de conséquences, qui se manifestent dans les descriptions, raisonnements et analyses proposés dans les cinq chapitres du livre. Les régularités morphologiques, au sein du lexique, se traduisent par des **relations** entre des mots, parfois binaires et/ou orientées, mais pas exclusivement. Contrairement aux règles syntagmatiques, émanant de la grammaire générative qui a longtemps régenté cette discipline, les principes énoncés font appel aux notions de **paradigme**, de compétitions de contraintes, de mécanismes analogiques, qui permettent d'expliquer le régulier en morphologie et les soi-disant exceptions. Un mot, fléchi ou construit, entre à la fois dans une **série** de mots formés au moyen du même procédé, et dans une **famille** dont tous les membres sont morphologiquement apparentés à un même mot. L'appartenance d'un mot à une famille explique que la forme de ce mot **se substitue** à autre forme, pourtant catégoriellement ou sémantiquement plus appropriée, comme base de dérivation. Inversement, l'importance d'une série, inconsciemment identifiée comme telle par un locuteur, conduit celui-ci à **privilégier la sélection** de l'affixe permettant d'intégrer le mot à construire dans cette série. Chacune des cinq monographies que comporte le livre porte en soi une description exhaustive d'un phénomène morphologique, une méthodologie d'analyse, et un modèle explicatif et prédictif rendant compte des données traitées. De très nombreux exemples issus de la Toile étayent, de façon très convaincante, les hypothèses présentées.

Le chapitre 2 présente un modèle de flexion verbale à base de graphes relationnels annotés. Ce modèle met en jeu un schéma cyclique qui relie les différents radicaux d'un verbe, ainsi que des contraintes (bi)orientées qui spécifient chacune des relations. Ce double système rend compte du fonctionnement des verbes réguliers (qui instancient tous les arcs du réseau, au moyen de l'une des contraintes qui y sont définies), comme celui des formes irrégulières, et prédit les hésitations du locuteur face à la conjugaison d'un verbe qu'il ne connaît pas. L'une des caractéristiques innovantes de ce système réside donc dans son aspect purement relationnel, fondé sur des contraintes formelles.

Le chapitre suivant étudie les dérivations en *-isme* et *-iste*. Les noms en *-isme* se partagent en trois modèles, en fonction desquels se décident les conditions d'attestation des noms en *-iste* apparentés. Le premier modèle regroupe les dérivés axiologiques : la forme et le sens ne coïncident pas (*présidentialisme*). Le second réunit les noms d'activités (*parachutisme*), et le dernier, les noms de qualité (*parallélisme*). Ce chapitre est l'occasion de poser la question de l'unicité – et plus généralement, de la nature même – de la règle (ou du modèle) en morphologie.

La contribution du chapitre 4 est l'étude exhaustive, en trois parties, des conséquences qu'entraîne le phénomène de dissimilation sur la forme des mots construits du français : (i) les stratégies d'évitement dictées par la phonologie (dissimilation progressive ou régressive, effacement, haplologie, épenthèse, décalage), (ii) la manifestation morphologique du phénomène (échangisme), qui conduit à l'apparition d'une forme inattendue (thème ou suffixe) dans un mot construit (*endormissement*), (iii) le cas particulier des langages secrets (javanais).

Le chapitre 5 analyse la concurrence entre les suffixes *-ien* et *-éen*, envisagée du point de vue historique (et donc de la pression du lexique existant) et (morpho)phonologique. Les auteurs expliquent dans quelles conditions *-éen* supplante *-ien*, et montrent également la parenté historique de ces deux formes avec *-ain* et *-an*. La distribution de *-ien* et *-éen* est motivée par des facteurs prosodiques, phonologiques, elle est tributaire du type sémantique de la base et des besoins de cohérence lexicale. Cette distribution n'est pas soumise à des contraintes exclusives : c'est ce dont témoigne, dans les exemples proposés, l'hésitation des scripteurs de la Toile (« *chicagoan ? chigagien ? chicagoïen ?*») face à l'éventail des solutions dont il dispose.

Enfin, le chapitre 6 propose un modèle théorique de la morphologie. Cette théorie est fondée sur une organisation du lexique alliant paradigmes morphologiques (séries), paradigmes lexicaux (familles), mécanismes analogiques, et deux groupes de contraintes qui s'affrontent. L'auteur conçoit l'ensemble des mots comme un espace où interviennent relations et distances entre ces mots. Dans ce réseau, chaque mot est caractérisé par un sens, une catégorie et une forme, mais aussi une fréquence, qui en détermine l'accessibilité. La forme et le sens d'un mot morphologiquement construit résultent de faisceaux disjoints de contraintes. Les principes qui gouvernent ce modèle sont illustrés, dans la deuxième partie du chapitre, par une analyse unifiée, originale et solidement argumentée, des adjectifs préfixés par *anti-*. Cette analyse explique à la fois les cas standard (*anti-grippe*) et les différentes manifestations de décalage entre forme et sens (*anti-cancéreux, anti-piratable*).

Pour le taliste qui s'intéresse au lexique et à la morphologie, la lecture de ce livre offre de précieux atouts. On y trouve, comme dans d'autres ouvrages avant lui, une argumentation solide et étayée en faveur d'une morphologie fondée sur les relations entre mots, plutôt que sur les règles de combinaison entre morphèmes. De plus, s'il

veut appliquer à l'analyse morphologique automatique un modèle empiriquement (et psycholinguistiquement) valide, le spécialiste du TAL peut adapter (voire adopter) les modèles formels présentés dans quelques-uns des chapitres (ou, le cas échéant, exploiter la description minutieuse et exhaustive des données traitées). Ainsi, au chapitre 2, les graphes relationnels annotés de contraintes sont intégralement reproduits et testés pour prédire la forme fléchie des néologismes verbaux. Au chapitre 3, les compétitions de contraintes, conduisant à l'identification du radical approprié de la base des noms en *-isme*, sont proposées dans un format proche de la théorie de l'optimalité. Enfin, le modèle théorique du chapitre 6 a été implémenté, ce qui est une garantie en soi de son opérationnalité. L'important recueil de données récentes servant à illustrer les raisonnements menés aux chapitres 2 à 6 constitue en soi une ressource précieuse pour le TAL, ainsi que la description systématique des phénomènes (morpho-)phonologiques qui se rapporte à ces ressources, dans les chapitres 4 et 5.

Pour conclure, le spécialiste du TAL retirera un grand bénéfice dans la lecture de cet ouvrage consacré à la morphologie du français : pour ses qualités pédagogiques, le caractère novateur de ses analyses, les ressources lexicales qu'il fait découvrir, et les modèles et outils formels qui offrent un traitement unifié de ces données. La symbiose entre TAL et morphologie ne peut que bénéficier aux deux partis : le terrain de la modélisation informatique de la morphologie lexicale, très demandeuse de données et d'outils capables d'explorer le Web, est encore peu exploité, la majeure partie des applications en morphologie relevant de l'apprentissage automatique.

---

**Thierry POIBEAU, Traitement automatique du contenu textuel, *Hermès-Lavoisier*, 2011, 222 pages, ISBN 978-2-7462-3191-7.**

Lu par **Stéphanie Weiser**

*CENTAL, Institut Langage et Communication, Université catholique de Louvain*

---

*L'ouvrage de Thierry Poibeau présente un panorama des questions que pose l'analyse automatique de textes dans le domaine du TAL. Après une introduction très théorique, l'auteur traite de trois applications différentes, portant chacune sur un niveau d'analyse différent. La mise en perspective des questions applicatives avec des considérations théoriques n'est jamais omise.*

Dans l'introduction, l'auteur présente clairement les objectifs de l'ouvrage et situe les problématiques abordées dans le contexte du TAL, aussi bien sur un plan théorique que sur un plan applicatif et historique.

Le premier chapitre de cet ouvrage est le plus théorique. Il présente les théories linguistiques et philosophiques ayant donné naissance au TAL, ou plutôt sur

lesquelles le TAL repose. Plus encore, il permet à l'auteur de placer les fondements sur lesquels les applications présentées dans les chapitres suivants s'appuient. Les travaux de Wittgenstein y tiennent une place importante. L'auteur présente aussi trois courants philosophiques et linguistiques sur lesquels reposent ses travaux actuels. Il s'agit des travaux de Firth sur la collocation, qui rejoignent ceux de Harris avec la notion de sous-langage (ou langage restreint chez Firth) et enfin des travaux du Cambridge Language Research Unit sur la notion de primitive sémantique. Ces travaux permettent d'élargir l'analyse linguistique de phénomènes précis aux questions posées par l'analyse de textes ou de corpus de textes, en se fondant sur l'usage. Ce premier chapitre introduit aussi le débat, incontournable en TAL, du choix entre des méthodes symboliques à base de règles et des méthodes probabilistes.

Les trois chapitres suivants présentent chacun une application avec les questions théoriques qui l'entourent, selon un plan d'analyse différent (micro-, méso- et macrosémantique selon la terminologie de Rastier).

Le deuxième chapitre présente l'annotation sémantique comme une étape nécessaire à de nombreux systèmes de TAL, tout en montrant les questionnements théoriques qu'elle soulève. Ce chapitre présente les principes généraux de l'annotation sémantique avant d'aborder en détail le domaine des entités nommées. Afin d'illustrer les problèmes théoriques posés par la tâche de repérage d'entités nommées, mais surtout afin de montrer son utilité pour des applications réelles, est ensuite présenté TagEN. C'est un système de repérage d'entités nommées fondé sur des règles linguistiques et utilisant des dictionnaires et des grammaires (développées à l'aide d'Unitex). Il intègre des modules multilingues (non évalués) ainsi qu'un moteur de désambiguïsation d'entités. De plus, les problématiques spécifiquement liées aux entités nommées sont présentées : instabilité référentielle, métonymie. Au fil du chapitre, plusieurs évaluations sont présentées, et les résultats comparés à ceux de systèmes similaires.

Le chapitre 3 s'intéresse aux rôles sémantiques et aux relations entre entités. Le contexte est celui de l'extraction d'information et l'application présentée pour illustrer les considérations théoriques est un moteur d'acquisition automatique de schémas de sous-catégorisation. Sur un plan théorique, l'auteur fait un retour sur la notion de prédicat et montre que si la notion de sous-catégorisation a été largement étudiée, elle reste néanmoins floue. Le système ASSCi qui est présenté permet d'acquérir automatiquement des schémas de sous-catégorisation pour l'analyse de verbes français. La phase de prétraitement consiste en une annotation de surface effectuée à l'aide de TreeTagger et Syntex. Les verbes et leurs compléments sont ensuite identifiés par un extracteur de préschémas de sous-catégorisation puis un constructeur de schémas candidats rassemble, pour chaque verbe, les schémas observés en corpus. Enfin, une importante phase de filtrage permet d'éliminer les schémas erronés (méthode statistique). En sortie, le système fournit un lexique de couples verbe-schéma de sous-catégorisation. L'utilité d'ASSCi est illustrée dans la

suite du chapitre par l'expérience consistant à l'utiliser sur un gros corpus afin d'acquérir un lexique de sous-catégorisation, LexSchem. Une place importante est laissée, dans ce chapitre, à l'évaluation des applications présentées. Enfin, dans la dernière partie, ce chapitre explore l'acquisition de familles sémantiques et propose une méthode hybride combinant des méthodes statistiques et une validation humaine.

Les analyses et applications présentées dans le chapitre 4 sont consacrées au texte dans un sens plus global. Deux axes principaux sont explorés. Le premier concerne le traitement de textes procéduraux (guides de bonnes pratiques dans le domaine médical). L'analyse de ce type de textes pose des problèmes car il est nécessaire de dépasser le niveau du syntagme et même de la phrase pour accéder au niveau discursif. Le principal enjeu est de pouvoir déterminer la portée des séquences conditionnelles dans des textes de recommandations *(en cas de X, procéder à Y sinon faites Z).* L'auteur s'appuie, sur un plan théorique, sur la notion de cadre proposée par Charolles et revient sur la notion d'architecture textuelle, ainsi que, de manière plus générale, sur la définition du texte en lui-même. Des questions plus générales sont également soulevées sur la modélisation de textes, sur les genres de textes, et sur l'évaluation de ce type d'application. Le deuxième axe présenté dans ce chapitre explore les problématiques liées au résumé automatique. Après un retour historique sur les méthodes de résumés automatiques, ce chapitre présente le système Cbseas qui peut être utilisé dans différents cadres (analyse de fonds documentaires, résumé de type « mise à jour » et résumé d'opinions) et qui s'appuie sur les classes sémantiques pour sélectionner les phrases les plus centrales dans le texte ou le corpus de textes à analyser. Cbseas a été mis en œuvre à deux reprises dans le cadre de campagnes d'évaluation pour lesquelles des données réelles sont fournies et des comparaisons sont ensuite possibles.

Le chapitre 5 constitue la conclusion. Les trois chapitres précédents pouvant paraître indépendants les uns des autres, la conclusion permet de les remettre en perspective et montre leur principal point commun : chaque application part d'un système à base de règles et est ensuite étendue par un processus d'adaptation dynamique (à l'aide de méthodes statistiques, de validations manuelles, etc.). Les limites des approches actuelles sont aussi présentées et celles-ci concernent principalement le niveau cognitif : les données à analyser sont complexes et les mécanismes de raisonnements impliqués dépassent le cadre de la logique. Enfin, dans la dernière partie, Thierry Poibeau propose des pistes pour aller plus loin (court terme) et met en avant des besoins de renouveau du domaine du TAL en général (long terme).

Cet ouvrage est intéressant pour la communauté TAL puisqu'il aborde des problématiques générales rencontrées dans la plupart des domaines du TAL. Avec trois domaines et applications différents, les techniques d'analyse présentées sont variées. Il se lit facilement, dans le sens où il très bien structuré, et contient un index général et un index des auteurs, ainsi qu'un glossaire. De plus, le contexte théorique

est toujours très bien posé. Néanmoins, il est destiné à des personnes ayant de bonnes bases en linguistique et TAL. Davantage d'illustrations à l'aide d'exemples linguistiques précis auraient pu aider la lecture. La bibliographie très riche illustre la place importante laissée à l'état de l'art, que ce soit par des retours historiques ou pour un positionnement vis-à-vis des recherches actuelles. De plus, si cela n'apparaît pas clairement dans cette note de lecture, il faut néanmoins relever que l'auteur présente de nombreux travaux collaboratifs ayant pris place dans différents projets de recherche.

---

**Marine CAMPEDEL, Pierre HOOGSTOËL, Sémantique et multimodalité en analyse de l'information, *Hermès-Lavoisier*, 2011, 423 pages, ISBN 978-2-7462-3139-9.**

Lu par **Patrick SAINT-DIZIER**

*IRIT – CNRS, Toulouse*

---

*Il est indispensable de créer des outils capables de gérer les différents médias pour appréhender des données de plus en plus nombreuses et de plus en plus riches sémantiquement. Ces dispositifs doivent permettre d'annoter automatiquement les données de manière pertinente afin que les recherches et les analyses puissent se faire avec un niveau sémantique élevé.*

*Cet ouvrage décrit les solutions de la littérature, proposant ainsi un état de l'art de chacun des cinq domaines multimodaux : texte, image, audio, parole et vidéo. Il analyse également les difficiles problèmes de la multimodalité vraie et les concepts ontologiques associés, puis présente de nombreuses applications concrètes mises en œuvre pour différentes problématiques (chaînage, fusion, analyse du risque, crises, indexation, etc.), notamment dans le cadre du projet Infom@gic (Thalès).*

*S'adressant aux étudiants, aux ingénieurs et aux chercheurs,* Sémantique et multimodalité en analyse de l'information *apporte des réponses aux défis que le multimodal et le multimédia posent aux différents acteurs de l'économie du numérique.*

Cet ouvrage, de plus de quatre cents pages, est composé de onze chapitres, de dimensions à peu près équivalentes, qui explorent les différentes facettes de la sémantique dédiée au traitement de la multimodalité. Il est rédigé par cinquante-cinq auteurs, chaque chapitre étant rédigé par un nombre important d'auteurs, de six à dix. Certains auteurs se retrouvent donc dans plusieurs chapitres, ce qui permet de garantir une meilleure cohésion à l'ensemble du texte.

L'ouvrage suit un cheminement assez classique : la modalité texte étant le point de départ, car c'est probablement ce champ qui a fait l'objet du plus grand nombre d'études et qui est le plus avancé. Suivent les modalités image et vidéo, audio et

parole. Suivent enfin des analyses à visée davantage méthodologique ou applicative : multimodalités et ontologies, fouille de données conversationnelles, analyse sémantique dans les moteurs de recherche multimédia, indexation sémantique de vidéos, extraction automatique d'événements, fusion en recherche d'informations visuelles, indexation audiovisuelle, et apprentissage pour l'annotation d'images. Chaque chapitre se termine par une très abondante bibliographie qui couvre les vingt dernières années environ. Ces onze chapitres couvrent à peu près l'ensemble de la discipline, on aurait pu, toutefois, y trouver aussi la musique, associée à l'audio, qui devient une problématique de recherche importante par exemple en indexation.

Tout chercheur un tant soit peu impliqué en sémantique sait à quel point ce champ disciplinaire est complexe et encore peu avancé du point de vue des applications. Par ailleurs, les liens avec la représentation des connaissances et certaines formes de raisonnement en complexifient encore l'approche. Les travaux les plus avancés en sémantique sont probablement ceux liés au texte, encore faut-il s'entendre sur une définition de ce terme : notons la distance qui sépare une sémantique formelle de type montagovien (véri conditionnel) d'une sémantique conceptuelle qui tente de représenter l'essentiel par le biais de primitives. Notons aussi que la sémantique peut difficilement se passer de la syntaxe. La notion de sémantique reste stable sur les autres médias, mais celle-ci devient plus complexe à caractériser en multimodalité du fait de la complexité des dispositifs d'analyse à mettre en œuvre et de leurs imperfections inhérentes.

Cet ouvrage émane assez directement du vaste projet Infom@gic, soutenu par le pôle de compétitivité Cap Digital. Il en reflète donc les problématiques et les orientations, même si les chapitres développent des considérations générales qui vont bien au-delà de ce qui a été effectivement étudié et réalisé. Examinons à présent quelques éléments de cet ouvrage. Cette analyse étant réalisée pour une revue de TAL, nous nous focaliserons bien entendu d'abord sur cette composante, en en reliant les autres.

L'introduction est quelque peu décevante. On y attendait une présentation rapide des principaux objectifs, des problématiques et des orientations scientifiques et techniques du projet, elle se contente de naviguer entre quelques problèmes, certes conséquents comme le transcodage, la fusion des données, etc.

Le chapitre 1 est celui qui nous est le plus proche. L'introduction y est sommaire : on aurait voulu connaître les problématiques et les composants du TAL qui sont les plus cruciaux dans le cadre de la sémantique et de la multimodalité. On est poussé assez rapidement vers le traitement des entités nommées et le problème de leur extraction. Suit une présentation, déconnectée de la précédente et un peu abstraite, des modèles statistiques pour le traitement du texte sans que l'on sache véritablement ce qui peut être traité avec telle ou telle méthode ni comment cela se gère en multimodalité. Ce chapitre est très décevant à plusieurs titres. Tout d'abord,

la seconde partie est un exposé abstrait, sans illustration, de modèles statistiques pour le TAL. Ensuite, et surtout, il est très peu question de sémantique, et on ne voit pas comment les éléments fournis sont utiles en traitement de la multimodalité. Certes, il est important de pouvoir caractériser les entités nommées reconnues dans des images ou des vidéos, mais l'information sémantique est plus large, on peut citer l'expression de l'espace et du mouvement, l'expression d'actions élémentaires telles que l'on puisse les segmenter sur des vidéos, etc.

Le chapitre 2, image et vidéo, est plus convaincant pour un lecteur peu expérimenté, il résume, de façon toujours un peu abstraite un certain nombre de techniques simples. On voit bien qu'image et vidéo sont des réalités assez différentes. Ce qui est difficile à percevoir, c'est comment des signatures mathématiques peuvent constituer de la 'sémantique' et comment elles peuvent être exploitées dans des situations d'indexation et de recherche d'information, voire de résumé de contenu. Le chapitre 3 ouvre la problématique des modalités audio et parole, en fait essentiellement parole. À ce stade de la lecture, on découvre à nouveau des exposés sur les méthodes statistiques idoines des traitements de la parole : on se dit alors qu'il aurait été peut-être préférable d'avoir un chapitre entier dédié aux méthodes statistiques (apprentissages, etc.) et qu'ensuite les chapitres dédiés au texte, la vidéo et la parole auraient pu être plus orientés vers le 'métier'. Cependant, ce chapitre est assez convaincant et utile au lecteur ayant un bagage modeste dans ces domaines.

Le chapitre 4 est dédié à l'intégration d'ontologies dans le multimodal, l'objectif étant l'extraction d'informations sur ces médias. Le chapitre commence assez logiquement par développer le problème de la fusion d'informations très hétérogènes en contenu et forme. L'approche est essentiellement fondée sur des métriques pour obtenir des classifications qui soient consensuelles. Interviennent alors les ontologies, utilisées pour des tâches de normalisation. Le terme ontologie couvre un vaste panel d'approches et de systèmes de représentations. Celles évoquées ici sont relativement simples. De fait, ce chapitre, en contraste avec les précédents demeure introductif et très peu formel.

Les chapitres qui suivent sont davantage orientés vers des applications. Notons les développements en sémantique pour les moteurs de recherche multimédia, domaine très actuel mais encore peu exploré. Des fonctionnalités simples et utiles y sont présentées. Le chapitre 7 développe des éléments centraux en sémantique à partir d'indexations vidéo : les résultats, simples, peuvent constituer un riche ensemble de données pour la sémantique de l'action et du mouvement, en particulier.

Cet ouvrage est à la fois intéressant, utile, et pourtant un peu décevant. Les aspects formels et sémantiques du TAL, de la vidéo et de la parole sont quelque peu éludés au profit d'exposés très académiques peu mis en perspective. Le lien n'est pas immédiat avec les chapitres orientés applications, qui, eux, renferment de

bonnes idées, intéressantes, bien illustrées et susceptibles de faire évoluer ce champ disciplinaire particulièrement complexe.

---

**Delphine BATTISTELLI, Linguistique et recherche d'information : la problématique du temps, *Hermès-Lavoisier*, 2011, 244 pages, ISBN 978-2-7462-2582-4.**

Lu par **Marie-Hélène LAY**

*Université de Poitiers, Laboratoire ForeLL*

---

*L'ouvrage situe son objet à la croisée des deux champs disciplinaires que sont la recherche documentaire et le traitement automatique des langues. La question abordée est celle de la structuration des informations repérées dans les textes et de leur représentation : la composante temporelle est envisagée comme une dimension structurante permettant la circulation dans une organisation de l'information par ancrage calendaire. Les contextes applicatifs évoqués sont ceux des systèmes de questions/réponses et des frises chronologiques.*

De nombreuses conférences et campagnes d'évaluation soulignent l'intérêt du traitement automatique de l'information temporelle pour toute une série d'applications dans le domaine de la recherche d'information. Détecter l'organisation temporelle des informations impose de prendre en compte un certain nombre de paramètres complexes : non-coïncidence de la temporalité événementielle et de la temporalité textuelle, modalisation (présentant un événement comme « réalisé-certain-pris en charge par l'énonciateur » ou « hypothétique douteux »), etc. Tout ceci sur des unités textuelles dépassant le cadre de la phrase. Partant de cette posture, l'ouvrage s'organise en quatre chapitres.

Le premier chapitre permet de faire le point sur les campagnes d'évaluation menées dans le champ de la RI, distinguant deux paradigmes tous deux ancrés dans un principe de référence calendaire : (1) la représentation d'événements sur des lignes temporelles, sous-jacente à toute une famille de logiciels de visualisation comme Time Wall ou SIMILE's TimeLine ; (2) l'analyse et la modélisation des mécanismes inférentiels ainsi que le degré de « certitude » concernant la réalisation effective des événements. Identifier un événement, lui attribuer un degré de « certitude » (de « factualité »), le dater, l'intégrer dans une série chronologique, tous ces aspects sont à prendre en compte et mobilisent à divers degrés les catégories linguistiques de temps, d'aspect, de modalité et de modalisation de l'énoncé comme de l'énonciation.

Deux chapitres sont alors consacrés au traitement de la temporalité en linguistique. Le deuxième chapitre aborde les choses d'un point de vue théorique allant du niveau morphosyntaxique à celui des opérations énonciatives. Au niveau morphosyntaxique, les valeurs aspecto-temporelles comme modales ne sauraient se

limiter aux unités qui les véhiculent. Par définition, elles portent à tout le moins sur la proposition ou sur des unités textuelles plus vastes, et elles peuvent être intriquées, contribuant à « brouiller » l'ordonnancement des séquences. Les textes sont constitués de séquences d'objets composites articulés de façon cohérente, fondés sur une logique temporelle. Mais comment identifier les expressions temporelles, comment en déterminer la portée : que retenir de *deux jours après moi* ? Comment traiter l'enchâssement des propositions ? S'ensuit une discussion autour de l'ordonnancement temporel des procès, allant de nouveau du niveau du « mot » à celui du texte. L'auteur s'emploie alors à la présentation des adverbiaux temporels de datation et se focalise sur le traitement privilégié à réserver aux expressions calendaires. L'objectif est de décrire des opérations de référenciation à l'univers calendaire propre au texte, et de proposer un schéma d'annotation dédié, autorisant la circulation entre différents sous-systèmes calendaires. La dynamique interprétative du texte pourra se faire sous forme de graphe. Les nœuds du graphe correspondent aux référentiels identifiés dans le texte, les arcs correspondent aux transitions entre les référentiels. L'ensemble des référentiels locaux s'articule pour constituer le référentiel global du texte. Les unités textuelles dégagées sont donc caractérisées par des marqueurs de rupture ou de continuité temporelle et/ou discursive (identification de citations, par exemple). Toutes ces unités sont en relation les unes avec les autres dans un espace référentiel propre au texte et au positionnement énonciatif qu'il traduit (discours asserté, rapporté, assumé ou pas, etc.). La modélisation retenue s'inscrit dans la tradition énonciative investie explicitement en termes d'opérations, les quatre opérations retenues étant la visée aspectuelle, le positionnement temporel, la catégorisation modale et le cadrage (positionnement énonciatif, temporel, spatial, thématique), cette dernière opération permettant de créer des « sous-cadres discursifs ». La position ici retenue est qu'il faut distinguer le niveau prédicatif du niveau énonciatif : seul le niveau énonciatif procède d'une opération de repérage dans le temps par rapport à un quelconque repère et le niveau prédicatif se calcule par rapport à ce « moment » de l'énonciation.

Le troisième chapitre fait le point sur les méthodologies mises en œuvre au sein du TAL et en examine les aspects « appliqués » à la recherche d'information, essentiellement dans les domaines de la défense, du médical et du juridique. La référence calendaire associée à l'identification des phénomènes citationnels et modaux semble aujourd'hui indispensable à la qualification de l'information recherchée. Cette approche est présentée comme caractéristique des années 2000 et située par une mise en perspective historique remontant aux années 70. La principale difficulté rencontrée aujourd'hui tient à l'absence de consensus sur la standardisation des étiquettes et des procédés d'annotation. Dans ce contexte, l'annotation est guidée par des finalités applicatives : plutôt que de fournir une analyse temporelle complexe des textes, il s'agit de détecter l'information qu'il est utile d'annoter... Tout en essayant de finaliser des schémas d'annotation qui ne soient pas strictement dépendants d'une application visée : d'où l'ambiguïté des

métalangages ayant une visée de standardisation, comme TimeML. Enfin, pour que la recherche d'information soit efficace, il faut aussi caractériser la validité de l'information. À nouveau, les travaux sur les marqueurs de temporalité vont de paire avec ceux sur les marqueurs de modalisation et rejoignent les travaux sur l'annotation des opinions.

Le chapitre 4, enfin, problématise la description linguistique du temps en s'appuyant sur des applications concrètes, l'une relevant de l'analyse d'opinion, l'autre de la navigation dans les textes. Il s'agit donc de mettre en regard les acquis du TAL et les besoins spécifiques de communautés d'utilisateurs. L'adaptation à une communauté d'utilisateurs pourrait, par exemple, intégrer les habitudes rédactionnelles propres à telle sphère d'activité pour en inférer des stratégies de recherche d'information fondées linguistiquement.

Valider l'information, évaluer le degré de la réalité d'un fait évoqué, la crédibilité qu'on lui accorde, voire la fiabilité de la source, conditionnent la performance d'un système de recherche d'information. Le processus de validation repose essentiellement sur trois mécanismes : la réputation de la source, le contexte du document et l'analyse de son contenu, dernier point pour lequel le TAL est mobilisable. L'exemple détaillé est celui du peuplement d'une ontologie en biologie, domaine où les modèles évoluent vite et où de nombreuses publications permettent d'accéder à l'information nécessaire à la mise à jour des ontologies. L'objectif peut être aussi de simplement assister l'exploration d'une littérature scientifique très abondante. Dans ce cas, il faut rapprocher les segments comparables concernant un fait biologique et y associer des informations pour déterminer si ce fait est présenté comme certain *vs* incertain. À l'ontologie biologique s'associe donc une ontologie de la modalité.

Pour ce qui est de la navigation dans les textes sur base calendaire, l'étude se situe dans le cadre de l'ANR *conique* et porte sur l'extraction de connaissances par inférence. Le système cherche à évaluer la relation entre les références temporelles de la question posée et celles de la réponse proposée, ce qui permet de donner des indices de pertinence : à la question « Quel est le nom de la célèbre comète apparue *pour la dernière fois* près de la terre *entre 1985 et 1986* ? », les informations temporelles associées permettent d'éliminer un certain nombre de comètes. En l'état le système ne traite que les expressions calendaires datatives (*en 1985*). Un système d'information dédié aux informations touristiques d'accessibilité à des lieux a, par ailleurs, été développé, pour répondre à des recherches comme « *musées ouverts à Venise le week-end du 1$^{er}$ mai* »

En conclusion, l'ouvrage rappelle que les travaux menés se situent au point de rencontre entre le *temps en tant que catégorie sémantique* et le *temps en tant qu'espace d'interrogation et de représentation de l'information*. Un projet en cours, ChronoLines, permettra d'intégrer les différentes briques présentées jusque-là : il s'agit de visualiser les informations reçues par l'AFP. Conciliant les approches de la

linguistique énonciative et de la linguistique textuelle, une définition opérationnelle de la temporalité linguistique est proposée : elle *donne à voir* des situations selon certaines caractéristiques aspectuelles, temporelles et modales ; de plus elle les relie à l'intérieur de cadres référentiels ou discursifs, les *donnant alors à voir* comme des segments dénotationnels homogènes, au sein desquels on peut circuler. On peut envisager différents types de navigations et rompre ainsi avec la linéarité des textes. Ce mode d'accès à de l'information ne se substitue pas à la lecture linéaire, mais met l'accent sur la gestion et la personnalisation des savoirs. Il s'agit de tirer parti du TAL et des acquis méthodologiques de la RI pour proposer le développement de parcours interactifs de navigation textuelle.

La lecture de l'ouvrage est parfois malaisée. Pour autant, il propose une posture intellectuelle intéressante. Outre l'aspect « État de l'art », les implémentations proposées au lecteur, les familles d'applications présentées, donnent à voir les différentes facettes de cette question : peut-on utiliser les informations temporelles linguistiques pour identifier des briques informationnelles « événements » et leur associer toutes les informations de « validité » nécessaires à leur exploitation ? Ce qui me semble intéressant, c'est que la dimension applicative est vue comme porteuse des bonnes questions théoriques, celles qui permettent d'interroger les modèles existants, de les éprouver, et, à terme, peut-être d'en proposer d'autres.

---

**Étienne BRUNET, Ce qui compte. Écrits choisis, tome II. Méthodes statistiques, *Champion*, 2011, 373 pages, ISBN 9782745322258.**

Lu par **Gérald PURNELLE**

*CIPL, Université de Liège, Belgique*

---

*Choix de dix-sept articles d'Étienne Brunet, de 1970 à 2009, qui abordent les questions théoriques et méthodologiques dans le domaine de la statistique lexicale et de la lexicométrie. À la fois histoire de la discipline et guide méthodologique, le volume constitue l'ouvrage de réflexion d'un des meilleurs spécialistes de la discipline.*

Dans la suite directe de *Comptes d'auteurs, Études statistiques, de Rabelais à Gracq*, volume dans lequel Damon Mayaffre avait tiré de la bibliographie d'Étienne Brunet un choix d'études appliquées et de monographies consacrées à divers auteurs, ce deuxième tome de ses *Écrits choisis* rassemble dix-sept contributions de portée théorique ou méthodologique – un choix judicieux, opéré « avec l'approbation du maître », et qui se signale par sa richesse et sa diversité.

Certes, Étienne Brunet n'appartient pas tout à fait à la génération des tout premiers pionniers de la statistique linguistique, mais il a rapidement rejoint ceux qu'il considère comme ses maîtres, à commencer par le père fondateur de la discipline en France, Charles Muller. Littéraire venu à l'informatique puis à la

statistique à la fin des années 60, Étienne Brunet fait depuis longtemps figure de référence dans le domaine, respecté et écouté, auquel les deux volumes forment à la fois « hommage et témoignage » (Céline Poudat dans l'avant-propos).

On trouve en tête d'ouvrage deux véritables documents : le tout premier article d'Étienne Brunet, intitulé « Programme » et daté de 1970, où était décrit un programme d'ordinateur calculant la fréquence théorique et l'écart réduit pour chaque mot d'un corpus constitué de seize œuvres de Giraudoux ; et l'introduction de la thèse qu'Étienne Brunet a soutenue en 1976 et qui était consacrée au même auteur. Par-delà le côté anecdotique et historique de ces deux textes, on y observe d'emblée la convergence immédiate et définitive de trois compétences en un seul homme : le littéraire, l'informaticien et le statisticien, que l'on sent en plus d'un endroit passer avec un plaisir intellectuel intact de la formule statistique au clavier et du code au graphique. À deux reprises, l'auteur justifie le choix de conjoindre les trois compétences en un seul homme, plutôt que de confier la programmation à un informaticien. (Et lui-même continue à illustrer ce point de vue, à travers le développement permanent de son logiciel d'exploration et de statistique textuelle Hyperbase, dont une version figure sur un DVD joint à l'ouvrage.)

Mais ces premiers textes montrent en outre les grandes qualités pédagogiques dont Étienne Brunet fera toujours preuve et qui se retrouve dans chaque chapitre. Dans chaque article il explique, détaille, illustre, se met à la portée du lecteur, quel qu'il soit. C'est donc aussi à un grand pédagogue que cet ouvrage rend hommage.

Classés dans l'ordre chronologique, ces dix-sept textes esquissent par l'exemple une histoire de la discipline, de l'informatisation des textes littéraires (constitution des grandes banques données, le TLF, Frantext) à leur analyse statistique, vite baptisée lexicométrie.

C'est aussi une histoire, partielle mais instructive, des débats, voire des polémiques, théoriques et méthodologiques, qui ont agité théoriciens et praticiens de l'analyse statistique textuelle. Chacune de ses interventions montre combien Étienne Brunet fut et reste animé par une réflexion permanente sur la discipline, une vision globale de son évolution et de son avenir. Constamment il revient sur les méthodes de ses collègues pour les discuter, les comparer, les amender parfois, les utiliser et les implémenter dans son propre logiciel. Ses prises de position sont toujours courtoises et ouvertes, mais fermes.

Citons à cet égard l'article très technique (sur le plan mathématique) où il prône l'utilisation de la loi normale plutôt que la loi hypergéométrique dans le traitement des grands corpus (chapitre 4), ou celui où le « schéma d'urne » est discuté, pour avoir été la source d'un débat qui remettait en cause le rôle (statistique) du hasard dans la constitution du vocabulaire d'un texte.

La dimension à la fois historique et réflexive de l'ouvrage en fait presque l'équivalent d'un manuel ou d'une introduction théorique ou méthodologique, même

si le degré de difficulté varie d'un texte à l'autre. Au point qu'il aurait peut-être été intéressant de classer ces textes dans un ordre « pédagogique » plutôt que chronologique, en allant des questions générales et des textes introduisant la discipline jusqu'aux plus pointus. Mais l'ordre chronologique réserve un joli effet de bouclage, l'avant-dernier chapitre, « Plaidoyer pour la statistique linguistique », renvoyant comme une mise à jour au deuxième chapitre, qui « dessine les contours d'une statistique linguistique » (Céline Poudat).

Quant aux méthodes et objets auxquels Étienne Brunet s'est intéressé tout au long de sa carrière, on sait qu'il est l'homme des amples corpus et de la statistique lexicale. Celle-ci mobilise comme objets et concepts le mot, la fréquence, le vocabulaire, la spécificité lexicale et la distance intertextuelle (chapitres 12 et 17).

Les articles rassemblés illustrent combien la banque de données Frantext constitue le corpus de référence qu'inlassablement il explore, découpe, et prend pour référence depuis de nombreuses années. Soit il étudie toute la littérature à partir de Frantext, notamment dans son découpage chronologique (chapitre 3), soit il se concentre sur des corpus exhaustifs d'auteurs, en comparant les sous-corpus dont ils sont constitués ou en confrontant l'auteur au corpus de référence. À tous égards, les dimensions qui sont observées sont régulièrement le genre, la diachronie, la forme (vers ou prose).

Mais Brunet ne s'en est pas tenu au vocabulaire du texte (plus précisément, l'ensemble des « graphies » qui le constituent, chacune étant considérée indépendamment des autres, dans de simples dénombrements) comme substance de ses études et méthodes. Signalons sa contribution à la question de la lemmatisation : faut-il lemmatiser les textes (ce qui est coûteux ou source d'erreur) ? Le gain de précision statistique est-il assuré ? Le chapitre 9 (« Qui lemmatise dilemme attise ») apporte un éclairage surprenant : le volume des corpus mobilisés a pour conséquence que les tests statistiques sont peu sensibles à l'opposition lemmatisation/non-lemmatisation. Mais en d'autres cas le lemme est un réel apport, auquel l'auteur ne renonce pas.

Le lemme n'est pas la seule information linguistique attachée à la forme du mot : les critères morphosyntaxiques, que les méthodes modernes d'étiquetage peuvent produire, apportent un enrichissement à la statistique textuelle.

Une troisième voie, particulièrement riche, par laquelle Étienne Brunet dépasse le mot, tend à appréhender la linéarité du texte en passant du mot comme simple occurrence dans le texte (que l'on peut dénombrer, pondérer, etc.) à la séquentialité de celui-ci : cooccurrences, rafales, séquences et enrichissement du vocabulaire (chapitres 6, 13, 15)

Enfin, on relèvera combien Étienne Brunet apprécie l'exercice complexe de la comparaison des méthodes (aux chapitres 11, 12 et 14, ou dans le chapitre 17, qui est un hommage à Charles Muller).

Terminons par le meilleur : le style clair et naturel d'Étienne Brunet s'enrichit d'un humour subtil et sans complexe : clins d'yeux et jeux de mots émaillent plus d'un texte. À cet égard, laissons la parole à l'informaticien qui double le littéraire : « La machine doit s'irriter très fort qu'on la compare une fois de plus à une personne et qu'on lui prête des sentiments ».