Notes de lecture

Rubrique préparée par Denis Maurel

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

Philip WILLIAMS, Rico SENNRICH, Matt POST, Philipp KOEHN. Syntax-based Statistical Machine Translation. *Morgan & Claypool publishers*. 2016. 190 pages. ISBN 978-1-62705-900-8.

Lu par **Fabrice LEFÈVRE**

Université d'Avignon / LIA-CERI

Contrat non rempli. Un ouvrage de bonne qualité, rigoureux, mais qui n'offre pas réellement une synthèse du sujet.

Contrat non rempli. Les *Synthesis on Human Language Technologies*» de Morgan & Claypool commencent à former une belle collection (trente-cinq ouvrages publiés jusqu'à présent), très représentative du domaine du traitement de la langue, diverse, mais aussi généralement fidèle à son appellation. Or la *synthesis* est ce qui fait défaut ici à un ouvrage qui a par ailleurs de grandes qualités. Trop ancré dans sa vision algorithmique, trop pointu, le livre y perd en vision globale synthétique sur le sujet.

La structure de l'ouvrage repose sur quatre parties principales. Un premier chapitre introduit l'ouvrage en s'attachant surtout à présenter les modèles utilisés pour la traduction guidée par la syntaxe. Suit un chapitre entier (chap. 2) consacré à la présentation des trois techniques principales d'extraction de grammaire (Hiero, SAMT, GHKM) à partir de corpus parallèles alignés ou pas. Enfin, élément central de l'ouvrage, le décodage est détaillé dans trois chapitres. Le premier permet de poser les bases d'un formalisme (autour de la notion d'hypergraphe) et rend compte des algorithmes associés (chap. 3). Un chapitre entier est ensuite dédié d'abord au décodage d'arbres syntaxiques (chap. 4.), puis au décodage de chaînes (chap. 5.), qui passent en revue les détails des techniques permettant la mise en œuvre de ces approches. Cette partie importante (77 pages) est suivie d'une partie pêle-mêle qui traite de sujets épars (transformation d'arbres, analyse en dépendance, grammaticalité ou encore le problème de l'évaluation) en une vingtaine de pages (chap. 6). Ces quatre parties sont suivies d'une très (trop) brève section de remarques conclusives (4 pages).

On aura noté donc que dans son développement le livre passe en revue de manière très précise et poussée les techniques de décodage pour une traduction aidée par la syntaxe, y consacrant presque trois chapitres entiers. Pour être honnête, à moins d'être en phase de mise au point d'un système, le livre tombe souvent des

mains tant nombre d'éléments relèvent plus des annexes que du corps principal de la présentation. D'ailleurs les auteurs nous mettent en garde dès la préface où ils qualifient leur livre de *solid fundation for beginning experimental work*. On soulignera malgré tout au passage l'intérêt de l'introduction de la notion d'hypergraphes qui permet de présenter ensuite de manière très cohérente et élégante les différentes instances d'algorithmes de décodage.

Au niveau général du livre, il est patent que l'intérêt d'une traduction guidée par la syntaxe n'est jamais vraiment démontré. C'est un postulat. Mais qui aurait pu, et dû, être appuyé par une analyse plus fine des différences linguistiques entre paires de langues. De même les caractéristiques structurelles (VSO *vs* SVO…) ne sont citées qu'une (ou deux) fois (et pas commentées). À cet égard, il est révélateur que la problématique du réordonnancement ne soit même pas présentée et discutée alors qu'elle représente un des problèmes clés de la traduction que la syntaxe est censée aider.

Encore à titre d'illustration, on pourra noter que l'étiquetage syntaxique, pourtant élément pivot de tout l'exposé, n'est jamais clairement introduit ni explicité (en dépit d'un bref et laborieux retour en fin d'ouvrage). Par exemple, les étiquettes elles-mêmes, pourtant utilisées à profusion dans le texte, ne sont pas définies. Bien sûr, ce manque ne posera guère de difficultés à un spécialiste endurci du TALN, mais fera défaut à un novice cherchant à élargir ses compétences.

Pas de discussion non plus vraiment sur la problématique de l'évaluation. Alors même que dès la préface, il est bien admis que l'intérêt de la syntaxe pour la traduction ne se révèle en général pas autant dans les mesures automatiques classiques, à la BLEU, qu'avec des évaluations humaines. Il aurait été souhaitable que quelques conclusions soient tirées de cela. Car sinon comment justifier tant d'efforts dans une approche dont les bénéfices sont condamnés à ne pas (ou difficilement) être démontrables en pratique ?

Dans la même ligne, on regrettera qu'une plus grande part ne soit pas faite à l'implication des nouvelles approches fondées sur les réseaux de neurones et l'impact qu'elles pourront avoir sur la traduction aidée par la syntaxe. Et ce, alors même qu'un certain nombre de travaux sont déjà engagés sur le sujet (par exemple sur la base de Delvin *et al.*, ACL, 2014). Aussi, alors que la notion de grammaire est bien présentée, dans un sens très large, le rôle de la sémantique n'est pas du tout abordé (relégué comme pour les réseaux de neurones à un très court chapitre dans la partie « Et ensuite ? » de la conclusion). Pourtant une quantité non négligeable de travaux, tels ceux menés à Hong Kong par D. Wu, présentent des résultats intéressants et sont très connectés à la traduction guidée par la syntaxe, notamment en partageant un grand nombre de problématiques (étiquetage, représentation structurée, complexité du décodage…). C'était un choix éditorial possible, mais qui concourt à renforcer l'étroitesse de vue du livre.

Dans un autre registre, il faut noter que la qualité littéraire de l'ouvrage est indéniable. Le texte est agréable à lire dans un anglais parfois très insulaire, mais qui nous sort agréablement de nos habituelles formules technico-scientifiques si formatées, sans perdre de sa précision ni entamer notre compréhension.

En résumé, il s'agit d'un ouvrage scientifiquement très solide, bien étayé par une bibliographie également solide, mais qui semble rater la cible de la série dans laquelle il s'inscrit. Un ouvrage à réserver donc aux lecteurs désireux d'acquérir rapidement les techniques liées à la mise en œuvre d'un système de traduction guidé par la syntaxe sans trop s'inquiéter des raisons qui les y ont conduits et justifient cette option. Les autres devront se tourner vers des présentations plus anciennes, comme la trentaine de pages qu'y consacre le livre de référence d'un des coauteurs, P. Koehn, « *Statistical Machine Translation* ».

Céline POUDAT, Frédéric LANDRAGIN. Explorer un corpus textuel. Méthodes – pratiques – outils. *De Boeck*. 2017. 240 pages. ISBN 978-2-80730-563-2.

Lu par **Chantal ENGUEHARD**

Université de Nantes – LS2N

Cet ouvrage synthétise l'expérience d'un groupe de chercheurs en linguistique en ce qui concerne la méthodologie d'exploitation d'un corpus. Il présente l'originalité d'avoir été construit à partir des pratiques et de répondre à un objectif pédagogique bien identifié : il s'agit de connaître un ensemble d'outils afin d'effectuer un choix fondé sur leurs fonctionnalités, sur les méthodes mises en œuvre, ainsi que les interprétations qui en découleront.

Les auteurs distinguent deux méthodologies principales d'exploration de corpus. L'analyse peut être fondée sur le corpus. Il s'agit alors d'une démarche déductive dans laquelle les données constituent un appui à une théorie linguistique. Ou bien l'analyse peut être guidée par le corpus dans une démarche inductive, sans hypothèse préalable. D'autres ambivalences sont également expliquées : l'analyse peut être dirigée par l'outil ou dirigée par l'utilisateur ; l'utilisateur peut préférer une analyse de linguistique qualitative ou une analyse de linguistique quantitative.

Le premier chapitre présente des définitions du domaine de l'exploration textuelle et introduit quelques notions telles la recherche de mots dans un texte (simple ou à l'aide d'expressions rationnelles), l'annotation de corpus (permettant l'élaboration de filtres), ou encore la recherche dans une base de données à l'aide de requêtes. Il s'agit de visualiser le corpus et d'être en mesure de l'interroger ou d'élaborer des statistiques.

Le chapitre 2 est consacré aux annotations. Les auteurs en donnent une définition, expliquent que leur pose peut être manuelle ou automatique. Les difficultés, notamment dues aux différents formats de fichiers ainsi que le respect des recommandations et standards sont évoqués. Des cas d'annotations d'une grande diversité sont présentés, telles la correction d'erreurs (langage SMS, textes médiévaux), la transcription de l'oral, l'anonymisation, l'annotation de la structure du texte ou encore la pose d'annotations enrichissant le texte (syntaxe, entités nommées, références). Différents modèles d'annotations sont détaillés. La nécessité

de phases d'expérimentation lors de la phase d'annotation est argumentée. Le chapitre aborde ensuite l'exploitation des annotations à l'aide de statistiques textuelles, des segments répétés et des cooccurrences.

Le chapitre 3 intitulé « exploration de la structure d'un corpus » présente différentes approches pour explorer le contenu d'un corpus (brut ou annoté) en faisant ressortir les attractions et oppositions qu'il recèle *via* une analyse statistique (comme l'analyse factorielle des correspondances ou l'analyse en composantes principales), et la construction d'un tableau de données. Trois niveaux de données sont distingués : les mots-formes, les lemmes et les parties du discours. Les limites des approches statistiques sont expliquées (absence de traitement des hapax, par exemple) et des notions plus élaborées sont abordées comme le test du *bootstrap* ou l'effet Guttman.

Le chapitre 4 traite de l'exploration d'une hypothèse élaborée par le chercheur en se fondant sur un corpus. Deux grandes démarches sont distinguées. La première est fondée sur l'examen d'une hypothèse au regard d'une structuration du corpus ; c'est l'occasion de rappeler quelques statistiques élaborées à partir d'un tableau de contingence : test de l'écart réduit, du Khi 2, etc. La seconde est focalisée sur l'examen d'une unité linguistique spécifique en regard de la structuration du corpus (concordances, segments répétés, cooccurrences). Les limites des mesures sont expliquées.

Les auteurs ont pris soin de définir et d'expliquer les termes techniques du domaine (repérés en gras), ce qui rend cet ouvrage très utile aux personnes qui aborderaient l'exploration de corpus. De plus, l'équivalent anglais de chaque terme est signalé afin de faciliter la compréhension de l'abondante bibliographie anglophone du domaine. Le texte est bien écrit, parsemé d'exemples, certains étant développés dans des encadrés, ce qui le rend très abordable. L'ouvrage aborde également des aspects très spécialisés de l'exploration de corpus (comme les « structures de traits récursives »). Il est donc à la fois utile aux novices du domaine et aux praticiens confirmés.

Une liste d'outils d'exploration de corpus mentionnés dans l'ouvrage figure en annexe, chaque outil y est sommairement décrit. Toutefois, cette liste est incomplète, ainsi R , SATO et Alceste, bien que présents dans le cours des pages, n'y figurent pas. De nombreuses fonctionnalités, présentées dans l'ouvrage, n'apparaissent pourtant pas dans l'index ou le répertoire des outils. On pourrait regretter également l'absence d'une grille d'analyse croisant les outils et les fonctionnalités afin d'en avoir une vision synthétique. Cette absence est probablement due au foisonnement d'outils. D'ailleurs, les auteurs signalent les difficultés qu'entraînent la diversité des formats manipulés par les outils et le besoin d'une interopérabilité ou d'un outil polyvalent. Ce souhait pourra inspirer la communauté TAL œuvrant dans le domaine des outils d'exploration de corpus.

Horacio SAGGION. Automatic Text Simplification. *Morgan & Claypool publishers*. 2017. 121 pages. ISBN 9-781-62705-968-1.

Lu par **Yannis HARALAMBOUS**

IMT – UMR CNRS 6285 Lab-STICC

Cet ouvrage montre de manière très convaincante que la simplification, en tant que transformation d'un texte en un autre, est une opération très difficile. Diverses approches sont présentées, et même celles qui ne prétendent fournir que des résultats très partiels sont loin d'être efficaces. On peut citer deux principales difficultés : (a) dans la mesure du possible, le sens du texte doit rester inchangé, (b) on doit être capable de mesurer la difficulté du vocabulaire et de la syntaxe utilisés et de les comparer à ceux du texte original. Le point (b) est d'autant plus difficile que l'on a du mal à définir ce qui peut être la difficulté d'un mot ou d'une structure syntaxique.

L'auteur, Horacio Saggion, professeur à l'université Pompeu Fabra de Barcelone, a, depuis sa thèse à l'Université de Montréal en 2000, travaillé sur le résumé automatique. Il s'est ensuite spécialisé dans la simplification de textes espagnols, en particulier à destination de personnes souffrant de dyslexie, dans le cadre de projets de recherche favorisant l'intégration de personnes avec des déficiences intellectuelles.

L'ouvrage est divisé en huit chapitres. Après une introduction générale à la problématique de la simplification de texte (chapitre 1) et un historique des travaux sur la lisibilité de textes (chapitre 2) arrivent les deux chapitres les plus importants, à savoir la simplification lexicale (chapitre 3) et la simplification syntaxique (chapitre 4). Le cinquième chapitre concerne la possibilité de la simplification en tant qu'opération d'apprentissage artificiel. Enfin, les trois derniers chapitres énumèrent des systèmes de simplification, des applications de la simplification, des ressources existantes, ainsi que des méthodes d'évaluation. Le lecteur animé par la curiosité scientifique trouvera son bonheur dans les chapitres 3, 4 et 5. Celui cherchant des solutions concrètes et implémentables (voire même implémentées) pourra consulter avec profit les chapitres 6 à 8.

Chapitres 1 et 2 : introduction et la question de la lisibilité

La simplification peut cibler trois catégories principales de lecteurs : (a) les personnes souffrant de dyslexie ou de déficiences intellectuelles, (b) les personnes avec un quotient intellectuel plutôt bas, (c) les apprenants d'une langue donnée. Il y a eu des travaux dans ce sens en anglais, portugais du Brésil, japonais, français, italien, basque et espagnol. D'autre part, il existe en ligne des journaux, ou magazines, simplifiés en suédois, norvégien, français de Belgique[1], flamand, danois, italien, finnois et espagnol. Enfin, il existe une version de Wikipédia en anglais

1 Le journal *L'Essentiel*, dont la version papier a été fondée à Charleroi en 1990 par la FUNOC, un centre de formation de jour pour adultes en difficulté d'insertion socioprofessionnelle.

simplifié, qui a – comme on le verra – beaucoup servi en tant que ressource linguistique.

La *lisibilité*, introduite ici puisqu'il s'agit d'évaluer la difficulté du texte produit, est le serpent de mer de la linguistique. En effet, elle a intéressé les pédagogues et les linguistes depuis le début du XX^{e} siècle, sans jamais donner de résultats satisfaisants. Déjà dans les années 80, il y a eu des centaines de formules de lisibilité dont le but est de caractériser le niveau éducatif requis pour lire un texte donné. L'application première de ces formules était l'affectation de textes dans les manuels scolaires de différents niveaux, ainsi que l'évaluation de la difficulté des textes donnés en examen. L'auteur donne une description très sommaire du domaine (le chapitre n'occupe en tout et pour tout que treize pages), mais avec une bibliographie bien fournie et des conseils de lecture pertinents.

Chapitre 3 : simplification lexicale

On entre ici dans le vif du sujet. La méthode classique consisterait à (1) détecter les mots difficiles, (2) les désambiguïser et en trouver des synonymes, (3) choisir le synonyme le plus simple. Or, le sens et donc aussi la difficulté dépendent du contexte. Pour cette raison, la plupart des travaux se sont plutôt appuyés sur l'*alignement* de corpus, et en particulier sur l'alignement entre la Wikipédia anglais standard et celui en anglais simplifié. Ainsi, Yatskar utilise l'historique de modification de la Wikipedia simplifié pour former une base de synonymes simplificateurs (selon l'hypothèse tout à fait plausible qu'un contributeur à la Wikipedia simplifié remplacera toujours un mot par un synonyme *plus simple*). Biran utilise plutôt des vecteurs de contexte pour détecter les (quasi-)synonymes et définit la complexité d'un mot comme étant le ratio de sa fréquence dans la Wikipedia standard divisé par sa fréquence dans la Wikipedia simplifié. D'autres, enfin, utilisent des méthodes d'apprentissage profond en ce qui concerne la synonymie, et des corpus spécifiques pour l'évaluation de la difficulté, comme le corpus LexMTurk qui contient cinq cents phrases anglaises tirées de la Wikipedia contenant chacune un mot marqué pour lequel cinquante utilisateurs d'*Amazon Mechanical TUrk* ont proposé, indépendamment les uns des autres, des versions simplifiées. Toutes ces méthodes ont donné des résultats plutôt moyens, il y a juste un domaine où la simplification semble bien se passer, c'est le domaine des expressions numériques : on réussit assez bien à remplacer des expressions comme « un quart » ou « 57 % » par « ¼ » respectivement « plus de la moitié ». Mais il reste toujours des pièges lorsque, par exemple, il faut tenir compte d'un seuil de précision donné pour comparer des valeurs. L'auteur donne l'exemple de la phrase « *l'inflation au Royaume-Uni est passée de 1,2 % en septembre à 1,3 % en octobre* », simplifiée par « *l'inflation au Royaume-Uni est passée d'environ 1 % en septembre à environ 1 % en octobre* » qui pose un léger problème d'ordre pragmatique...

Chapitre 4 : simplification syntaxique

L'approche la plus classique consiste à découper les phrases longues, avec propositions subordonnées, en phrases courtes, si possible du type sujet verbe complément. Cela peut être effectué en appliquant des règles de transformation aux

arbres syntaxiques. Or, il peut y avoir des problèmes de cohésion référentielle. Voici un exemple caractéristique, tiré de l'ouvrage : *Mr. Anthony, who runs an employment agency, decries program trading, but he isn't sure it should be strictly regulated.* On y trouve donc une subordonnée relative, suivie de la proposition principale, suivie d'une deuxième proposition reliée par une conjonction de coordination. Les règles vont tout naturellement découper ces trois propositions en trois phrases, dans l'ordre (1) proposition principale, (2) subordonnée relative, (3) deuxième proposition : *Mr. Anthony decries program trading. Mr Anthony runs an employment agency. But he isn't sure it should be strictly regulated*. La permutation fait que le référent canonique du *it* de la troisième phrase n'est plus *program trading*, mais *employment agency*. Il faut donc passer par une analyse du discours, ne serait-ce que pour obtenir l'ordre des phrases découpées.

Une autre approche, qui rejoint le résumé automatique, consiste à extraire du texte les événements clés en évitant les informations de moindre importance. La technique est encore une fois fondée sur des règles de transformation.

Chapitre 5 : la simplification en tant qu'apprentissage artificiel

Ce chapitre traite de la possibilité d'appliquer des méthodes de *machine learning* sur les corpus, entre-temps assez volumineux, du Wikipedia anglais standard (plus de 5,5 millions d'articles) et du Wikipedia anglais simplifié (environ 131 milliers d'articles). Il s'agit d'appliquer des méthodes de traduction automatique, en considérant la langue simplifiée comme une langue différente de la langue de départ. Au niveau des arbres syntaxiques par constituants, les principales opérations envisagées sont la scission, la suppression, le réordonnancement, et la substitution. On calcule un modèle probabiliste fondé sur ces opérations, et on apprend à les appliquer à de nouveaux textes. Dans des travaux plus récents, on trouve également une composante sémantique, et l'utilisation de la théorie de représentation du discours de Kamp qui fournit un nouveau graphe permettant à son tour des calculs de probabilités.

Chapitres 6 à 8 : les outils et les ressources

La lecture de ces chapitres est passionnante puisque l'on y découvre un fourmillement d'idées, mais aussi une montagne de difficultés et de problèmes à résoudre. La dernière partie du chapitre 8 est consacrée à l'évaluation des systèmes de simplification. Bien évidemment, l'approche classique est de passer par des juges humains, mais certains systèmes utilisent aussi des mesures d'évaluation empruntées à la traduction automatique comme BLEU, TERp et ROUGE.

Conclusion

Cet ouvrage est une mine d'informations. Il est agréable à lire et donne un très grand nombre de pointeurs vers des publications et des ressources. C'est une synthèse claire, complète et bien argumentée du domaine par un de ses spécialistes notoires.

Le seul point négatif est d'ordre psychologique : toutes les voies empruntées dans ce livre tombent tôt ou tard sur des obstacles, ce qui engendre une certaine

frustration. On a l'impression que toutes les difficultés du TAL se sont réunies dans ce domaine et que le chercheur courageux qui veut y travailler aura fatalement à se battre contre vents et marées. Comme si cela ne suffisait pas, dans la dernière page de l'ouvrage, l'auteur souligne le fait qu'il n'y a pas une *seule* simplification, et que selon les besoins de la population ciblée, *la simplification de l'un peut aggraver la situation pour l'autre* : entre dyslexiques, apprenants d'une langue, et déficients intellectuels, il y a, du moins partiellement, incompatibilité des besoins et donc aussi des solutions.

Nous espérons que cette note de lecture attisera la curiosité du lecteur intéressé par le domaine des transformations textuelles (résumé, traduction, paraphrase, simplification) et l'incitera à la lecture de cet ouvrage, qu'il ne regrettera pas !

Émilie NÉE. Méthodes et outils informatiques pour l'analyse des discours. *Presses universitaires de Rennes*. 2017. 248 pages. ISBN 978-2-75355-499-3.

Lu par **Nadia MAKOUAR**

INALCO

Le présent ouvrage, dédié à l'analyse des données textuelles, se propose d'aborder des méthodes pratiques sous l'angle de l'analyse du discours. Les auteurs de l'ouvrage, coordonné par Émilie Née, sont tous spécialistes de l'analyse outillée du discours. Ce manuel comporte six chapitres, ponctués par des encadrés détaillant au lecteur des informations de type documentaire, bibliographique et notionnel, ainsi que des résumés de recherches menées sur des données textuelles. Un index en fin d'ouvrage reprend les notions et les termes techniques.

L'ouvrage dédié à l'analyse des données textuelles ancre d'emblée son positionnement théorique du point de vue de l'analyse du discours et développe les choix méthodologiques et pratiques qui en découlent.

Dans l'introduction les auteurs présentent de façon concise et claire les prémices de l'analyse des données textuelles. Ceci est l'occasion pour le lecteur de découvrir ou de se rappeler l'évolution des pratiques de l'analyse outillée en statistique linguistique et lexicale. Les auteurs expliquent comment la lexicométrie a progressivement laissé place à la textométrie. La mise en commun des connaissances qualitatives et quantitatives a donc permis d'étendre l'observable de l'unité lexicale au texte. Ces précisions permettent de mettre en évidence les points de divergences entre l'analyse du discours et la sémantique interprétative, mais aussi de montrer que certains travaux se trouvent à l'intersection de ces deux théories. Pour dépasser ces clivages théoriques, la communauté regroupe ces disciplines et ces courants sous l'appellation « Analyse des données textuelles » (ADT) ; l'ADT se distingue de la linguistique de corpus et du traitement automatique des langues.

L'originalité de l'ouvrage est qu'il met en regard les outils et leurs méthodes en lien avec des problématiques concrètes en analyse du discours. Tout le long du manuel, les auteurs insistent sur la nécessaire réflexion à porter sur les données

avant, pendant et après leur regroupement en corpus. Il s'agit d'une démarche itérative où les hypothèses sont constamment mises à l'épreuve.

Le chapitre 1 introduit la question du décompte des unités textuelles et les méthodes implémentées dans chacun des logiciels étudiés. Par exemple, pour un même texte, les auteurs montrent que Word, Notepad et la commande CAT sous Unix intègrent des paramètres de comptage et donc de découpage et d'identification des formes graphiques différentes. Ce qui a une incidence sur l'analyse des données du chercheur. Ils proposent également l'exemple des nuages de mots et la nécessité de prendre en compte l'ensemble des éléments du texte (notamment les mots-outils) pour mieux saisir et interpréter les données soumises au logiciel. Dans cette même perspective de sensibilisation aux données étudiées, les auteurs ont recours à Ngram Viewer (conçu par Google) qui permet d'observer l'évolution chronologique d'un mot ou d'une séquence de mots sur la base de livres numérisés par Google Livres en français. Malgré l'obtention d'un graphique, il est difficile d'en faire une quelconque interprétation en raison de l'inaccessibilité des ressources analysées. Cet exemple illustre parfaitement la problématique de la connaissance du corpus, nécessaire à une analyse objective.

C'est dans cette perspective que les auteurs proposent dans le chapitre 2 d'éclairer le lecteur sur les principes de base liés à la constitution et à la structuration du corpus. Cette partie insiste sur la réflexion à porter sur le corpus à constituer ou en voie de constitution. Ainsi, les données se construisent à partir d'hypothèses de travail. À travers ces différents principes de constitution et de structuration, les auteurs fournissent des méthodes clés que le chercheur pourrait adapter à ses propres données. Les méthodes et les stratégies d'analyse y sont explicitées. Les notions de « hors corpus » ou de « sous-corpus à géométrie variable » éclairent sur l'éventail des méthodes applicables. Les auteurs illustrent également avec quelques recherches que la diversité des approches du corpus est possible à condition de bien définir leur « statut » durant la phase d'analyse. Un rappel nécessaire est fait sur le va-et-vient permanent entre l'aspect qualitatif et le retour au texte. Les auteurs proposent d'illustrer la démarche d'analyse en fonction de plusieurs types de structuration de corpus : par genre, par sources énonciatives, par sphères d'activités et par moments discursifs. Même si un seul critère peut présider à la constitution du corpus, les spécialistes rappellent néanmoins le nécessaire croisement des critères. La question des données issues du Web y est aussi abordée ainsi que la dimension « techno-langagière » qui interroge sur le rôle joué par le support (Twitter par exemple) dans l'interprétation des données.

Le chapitre 3 propose une mise en pratique des principes explicités dans le chapitre qui le précède. Ce troisième volet de l'ouvrage s'apparente à un tutoriel où le lecteur est guidé étape par étape dans la démarche d'analyse qui commence dès l'élaboration des premières hypothèses. Ces étapes articulent les données à étudier et l'outillage convoqué pour répondre aux questionnements sur l'objet d'étude. Trois scénarios de constitution de corpus sont proposés afin d'illustrer la démarche : à partir 1) d'un corpus médiatique construit autour d'une forme langagière, 2) d'un autre corpus sociopolitique construit autour d'un thème et enfin, 3) d'un corpus politique construit autour d'un genre. Cet éventail permet alors au lecteur de prendre

connaissance des principes méthodologiques et pratiques et des pièges qui lui faut éviter afin de ne pas fausser son analyse. Le chapitre explique minutieusement le parcours de la constitution et de l'analyse en termes de délimitation, de codage, de formatage, de structuration et de balisage, et ce, en fonction du logiciel utilisé. C'est pourquoi les auteurs insistent sur la connaissance du logiciel que le lecteur souhaitera utiliser avant d'y intégrer les données. Cette partie nous éclaire aussi sur l'évolution de l'ADT et ce qu'elle a permis en termes de disponibilité et d'accessibilité des corpus « réservoirs » (comme ESLO) ou « partagés » (comme Textopol), utiles pour tout chercheur.

Le chapitre 4 qui précise la problématique des formes graphiques évoquée dans le premier chapitre, nous éclaire précisément sur le comptage des unités. Il existe en effet plusieurs façons d'aborder les données textuelles : à partir d'une forme graphique, d'un lemme, ou d'une catégorie morphosyntaxique, notamment. Plusieurs illustrations, avec des copies d'écran, montrent comment certains logiciels sont capables de traiter ces types d'unités. Les principes et les usages des segments répétés et des cooccurrences en analyse des données textuelles sont également abordés et approfondis dans ce chapitre.

Le chapitre 5 offre un panorama sur la typologie des logiciels disponibles pour l'ADT. Il introduit tout d'abord quelques repères historiques et épistémologiques sur l'apparition et le développement de différents outils d'analyse, et ce, en fonction des préoccupations et des observations des chercheurs qui les ont conçus. Les auteurs donnent l'exemple de Max Reinert qui, en cherchant à étudier des données en psychanalyse a développé le logiciel Alceste lui permettant de faire émerger les thématiques dominantes dans un texte. À partir de corpus accessibles et disponibles en ligne, les auteurs proposent d'illustrer les différentes fonctionnalités offertes principalement par les outils proposant une approche structurante, d'une part (Alceste, IRaMuTeQ), et ceux contrastifs et longitudinaux, d'autre part (Lexico, TXM, Le Trameur, etc.). Les nombreuses illustrations détaillées et expliquées précisent de façon complète les différentes approches possibles des données *via* l'utilisation de ces logiciels, toujours en fonction des hypothèses formulées et mises à l'épreuve.

Le sixième et dernier chapitre éclaire et précise au lecteur l'articulation entre les questionnements méthodologiques et le traitement des données avec les outils d'analyse. Les auteurs mettent l'accent sur les problématiques herméneutiques. Ils y abordent la question de l'approche thématique d'un corpus et proposent des points d'entrées et plusieurs méthodes quantitatives. En maintenant l'attention du lecteur sur la question de l'interprétation des données, les auteurs montrent, par exemple, les implications de l'utilisation d'une approche inductive et déductive dans le cadre d'une recherche thématique. Les recherches qui illustrent ces démarches éclairent le lecteur sur les difficultés qu'il peut lui-même rencontrer et donc, contourner. Ces analyses montrent aussi quel type de cheminement interprétatif pourrait compléter la recherche de l'analyste et ainsi enrichir ses questionnements.

En fin d'ouvrage, les auteurs consacrent quelques pages sous forme de « fiches pratiques » essentielles à la manipulation et au traitement automatique des données.

De nombreuses copies d'écran permettent au lecteur de suivre pas à pas les commandes proposées pour le traitement des données. Une dernière fiche illustrée par des schémas et des copies d'écran propose d'approfondir la notion d'analyse factorielle des correspondances ; fonctionnalité que l'on retrouve dans plusieurs logiciels d'analyse textuelle. Ces éléments sont complétés par une bibliographie et une sitographie dédiée aux notions détaillées dans ces fiches.

Cet ouvrage didactique et accessible constitue un apport scientifique et méthodologique précieux pour quiconque souhaiterait mener une recherche en analyse des textes. Il représente une source considérable d'informations théoriques et pratiques sur les différents types d'approches en ADT.

Jean-Marc Leblanc. Analyses lexicométriques des vœux présidentiels. *Wiley-Iste*. 2017. 386 pages. ISBN 978-1-78405-210-2.

Lu par **Daniel YAO**

Université Jean L. Guédé de Daloa (Côte d'Ivoire)

L'ouvrage de Jean-Marc Leblanc traite des avantages liés à l'utilisation des outils de traitement de données textuelles selon une approche lexicométrique. Il présente au niveau formel, une subdivision en six chapitres encadrés en amont, par une partie introductive et en aval par une conclusion. L'ouvrage ambitionne de présenter un traitement transversal des textes en mobilisant des expérimentations plurielles grâce à divers logiciels de traitement textuel.

L'introduction pose les grands principes qui orientent le traitement du matériau analysé en l'occurrence, les vœux présidentiels des présidents de la République française depuis Charles de Gaulle jusqu'à François Hollande. Elle insiste aussi sur les balises tant théoriques que méthodologiques à observer afin de conduire une analyse lexicométrique exploitant les multiples fonctionnalités offertes par la pluralité des logiciels de traitement textuel. Cette étape liminaire précise enfin, la nécessité de revenir toujours au texte, à la suite des manipulations statistiques, car pour reprendre l'auteur, la lexicométrie ne saurait être un raccourci méthodologique pour des analyses clé en main.

Le chapitre 1 explicite la nature du corpus à analyser (l'ensemble des discours liés aux vœux présidentiels), une brève approche historique de la lexicométrie en tant que discipline et les travaux princeps y afférents. Les premiers travaux fondateurs avaient ainsi pour ambition d'examiner la ventilation du vocabulaire, la fréquence des occurrences, d'opérer des retours réguliers au texte et de mesurer son homogénéité stylistique. Cette partie décrit et compare également les différents logiciels, selon qu'ils sont contrastifs ou longitudinaux (Lexico 3, Hyperbase, WebLex, etc.), structurants (Alceste, IRaMuTeQ, etc.) ou catégorisateurs et évaluateurs sémantiques (Tropes, Cordial, etc.), même si la tendance générale actuelle des concepteurs évolue vers leur interchangeabilité.

Le chapitre 2 opère, de manière factuelle, le traitement lexicométrique des quarante-trois messages des septennats relatifs aux vœux présidentiels. Il procède aux premières analyses des vœux présidentiels des locuteurs selon une approche diachronique et comparative. J.-M. Leblanc mobilise à ce niveau, diverses techniques telles que les AFC et des mesures de proximité sur le corpus comme la distance sur V et sur N. Ces éléments soulignent une originalité gaullienne au sein de l'homogénéité des locuteurs, et spécifient la rupture et la continuité dans lesquelles s'inscrit François Mitterrand par rapport à Charles de Gaulle. Le logiciel Lexico 3 convoqué, révèle un contraste entre les discours des septennats et ceux des quinquennats avec Jacques Chirac qui affiche une indistinction, car il se démarque peu de ses prédécesseurs. Valéry Giscard d'Estaing est proche de Georges Pompidou, avec qui il partage la même connexion du vocabulaire, tandis que François Hollande se rapproche, sur la base de son premier discours, de Nicolas Sarkozy. Ce dernier partage des liens de proximité avec Chirac sur la distance lexicale au regard des expérimentations issues des logiciels TextObserver et Hyperbase.

Dans le chapitre 3, l'auteur procède à l'étude des catégories grammaticales, des modes et des temps lexicaux liés au corpus. Il examine aussi, par le biais des logiciels Lexico, Cordial et Hyperbase, les aspects morphosyntaxiques, et les emplois des marques personnelles. Les stratégies discursives indiquent, par exemple, que Chirac privilégie les constructions du type *il faut + infinitif*, les emplois déontiques, volitifs et le présent de l'indicatif. Chez de Gaulle, le temps présent est sous-employé, les volitifs sont également présents avec une absence significative des pronoms *je, nous* et *vous* . Giscard d'Estaing, quant à lui, privilégie le *je*, le futur de l'indicatif, les emplois explicatifs dans une approche didactique et répétitive comme *cela veut dire*. Chez Pompidou, les modes subjonctif et impératif sont valorisés en des accents gaulliens sur un ton argumentatif et un équilibre entre le *je* et le *vous*. Mitterrand, tout comme Chirac, favorise la simplicité et la connivence, la personnalisation du discours où saillit l'incarnation de la fonction. Les thématiques liées à l'Europe, à la sécurité et au dialogue » sont de même prégnants. Au final, cette partie 3 expose des stratégies discursives de légitimation, de justification, d'appel à l'unité nationale avec un vocabulaire stable ou conjoncturel au sein de ruptures syntaxiques, ainsi que des profils lexico-énonciatifs contrastés chez les auteurs.

L'auteur étudie de manière spécifique, dans le chapitre 4, la notion d'*ethos* présidentiel qui caractérise l'implication personnelle du locuteur dans le discours. À travers divers outils (Lexico 3, Hyperbase, WebLex), l'approche poly-cooccurrentielle indique que le « je » présidentiel est fortement ancré dans le rituel et le métadiscursif. Le « nous » des messages est de type argumentatif accompagné des verbes d'action. Les *ethos* varient selon les présidents de la République avec une forte tendance à l'empathie chez Chirac et Pompidou tandis qu'elle est absente chez de Gaulle. Le lexicogramme récursif confirme le caractère argumentatif chez Mitterrand qui lui permet de renforcer sa légitimité. Au final, la mobilisation du logiciel Alceste, *via* les lemmatisations et les tris croisés, donne des distributions

statistiques et linguistiques faisant de Giscard, le plus grand contributeur des énoncés du « je » avec une forte représentation du rituel.

Dans le chapitre 5, en mobilisant l'outil Alceste, J.-M. Leblanc examine les « mondes lexicaux » pour obtenir des lexèmes et extraire *in fine*, les thématiques dominantes. L'auteur procède ensuite à une analyse récursive pour approfondir la compréhension des composantes du rituel *via* la structure des phrases. Il opère enfin des analyses comparatives sur la base des partitions locuteurs. Les différentes expérimentations indiquent, de prime abord, avec le logiciel Tropes que, pour de Gaulle et Mitterrand, le thème le plus représenté est « nation » avec une place importante accordée à « l'Europe » chez le second. À l'inverse, Giscard et Pompidou n'évoquent nullement ces items, leur préférant les énoncés « Français et France ». L'outil Alceste, dont l'algorithme, fondé sur la classification descendante le distingue des logiciels classiques, fournit les cinq classes sémantico-thématiques du corpus : la classe 5, la plus massive dominée par Giscard, est relative au « rituel » ; la classe 1 est liée à « la politique internationale » où de Gaulle et Mitterrand sont les plus expressifs ; la classe 4 est inhérente aux « valeurs républicaines, démocratiques, un vocabulaire incitatif et volontaire » où de Gaulle est le plus contributif ; la classe 3 relative à la « politique intérieure, économique et sociale » est le lieu commun de tous les locuteurs ; la classe 2 établit les « énoncés constatifs et bilans » mobilisés surtout par les présidents de la République du septennat et Chirac. J.-M. Leblanc effectue grâce à l'indice du Khi 2, des retours réguliers au texte (concordance), après les résultats issus de Lexico 3. Des analyses récursives sont conduites sur les sous-corpus de chaque locuteur afin d'affiner les formes significatives qui leur sont associées.

Le chapitre 6 analyse et compare les deux derniers présidents : *Sarkozy, Hollande, questions de style ?* Dans cette ultime section, l'auteur se propose de dégager les principaux profils langagiers des deux locuteurs identifiés par le biais des instruments tels que l'AFC, les classes sémantiques, les histogrammes et des recours itératifs aux textes (concordances). Il faut noter ainsi que les messages de Sarkozy deviennent chronologiquement centraux sur l'axe factoriel. Les spécificités lexicales insistent sur la permanence du rituel, indépendamment de la transformation des styles des locuteurs : de Gaulle (Algérie, coopération) ; Pompidou (nier, situation, Français) ; Giscard (vœux, bonheur, Français) ; Mitterrand (droit, Europe, désarmement) ; Chirac (avenir, solidarité, emploi) ; Sarkozy (crise, urgence, avenir) ; Hollande (vœux, compétitivité, décision). Les classes thématico-sémantiques sous Alceste, expriment avec l'analyse récursive et l'indice Khi 2, l'individuation des messages des présidents de la République. Leur vocabulaire et leurs *ethos* respectifs impriment une image spécifique à chaque discours, quoique l'influence de l'événementiel sur le rituel ne soit pas insignifiante. Le positionnement énonciatif « je, nous, vous » semble enfin déterminant dans les rapprochements perceptibles sur l'analyse factorielle entre les locuteurs. Il souligne une prégnance de l'empathie chez Sarkozy dans le rituel, caractéristique qui n'est pas étrangère chez Hollande non plus.

En conclusion, l'ouvrage de J.-M. Leblanc s'inscrit dans une double perspective : à la fois didactique, car il nous initie aux bases fondamentales de la

lexicométrie, et médium de comparaison des outils de traitement textuel. Il a recours à des cas pratiques tant dans la manipulation que dans l'interprétation des résultats tout en renvoyant le lecteur à des extensions numériques pour approfondissement sur un site Internet. Le style de rédaction de J.-M. Leblanc est accessible, nimbé d'humour et les références bibliographiques sont actuelles, riches et variées. Il s'agit donc d'une contribution pertinente dans le champ du TAL.