Notes de lecture

Rubrique préparée par Denis Maurel

Université de Tours, LIFAT (Laboratoire d'informatique fondamentale et appliquée)

Ruslan MITKOV, Johanna MONTI, Gloria Corpas PASTOR, Violeta SERETAN. Multiword Units in Machine Translation and Translation Technology. *John Benjamins publishing company*. 2018. 259 pages. ISBN 978-90-272-0060-0.

Lu par **Christian BOITET**

Université de Grenoble-Alpes / GETALP

Multiword Units in Machine Translation and Translation Technology *est un volume de la collection* Current issues in linguistic theory *de John Benjamins. Les références s'arrêtent à 2013, 2014 ou 2015 selon les chapitres du volume. Du coup, il n'y a rien sur la TA neuronale, ce qui est dommage pour un livre sorti en 2018. Il y a d'abord un très long chapitre introductif (trente-sept pages dont douze de bibliographie), dont le but est de présenter le domaine, puis les trois parties, concernant les EML*[1] *en TA, les EML dans des applications multilingues du TAL et l'identification et la traduction des EML. Cette répartition serait à revoir. Par exemple, le dernier article porte sur la traduction des EML dans le système FIPS-2 et serait bien mieux placé dans la partie 1.*

Une remarque préliminaire importante est que, semble-t-il, tous les auteurs, à l'exception notable d'Éric Wehrli et Luka Nerima (dernier article), ignorent les contributions essentielles d'Igor Mel'tchuk à la représentation et au traitement des EML, qui ont commencé… en 1958, avec l'introduction des fonctions lexico-sémantiques (FLS), puis un grand développement de la théorie, l'incorporation dans plusieurs dictionnaires explicatifs et combinatoires (DEC), dans des systèmes de TA, et un très grand nombre de publications, en russe, en français et en anglais (au moins). Dès 1961, le CETA à Grenoble, dirigé par Bernard Vauquois, utilisait certaines FLS (dérivationnelles et productives) dans son système de TA russe français ! De ce point de vue, ce volume est assez décevant. En effet, les FLS sont une vraie clé pour la représentation et le traitement des collocations de type lexico-sémantique. Cependant, on trouve des choses assez intéressantes dans ce volume. Parcourons-le.

1 J'utilise en français le terme EML (expressions multilexicales), de préférence à d'autres termes.

Introduction

La longue introduction du volume, rédigée par les éditeurs, commence, bien sûr, par un petit historique des études sur les EML. On passe directement de Firth (1957) à Sag (2002)… comme si rien ne s'était passé en quarante-cinq ans. Or, il n'y a pas eu que Mel'tchuk à s'y intéresser durant cette période ! On a mentionné le CETA, mais il faut dire que pas mal d'autres systèmes de TA ont abordé le problème et trouvé des solutions pour beaucoup de types d'EML. Peter Toma a introduit dans Systran un « dictionnaire d'expressions » dès 1967, Slocum traitait les mots composés allemands dans METAL dès 1982, etc., et il y a dans cette introduction une section (3.2) qui décrit le traitement des EML dans OpenLogos.

On note aussi des affirmations bien en retard sur l'état de l'art : au 2.2, on nous dit que la recherche à venir dans ce domaine va surmonter les limitations de l'approche dominante « mots-avec-des-blancs ». Mais enfin, la thèse de Violeta Seretan (une des éditrices !) en 2008 a très largement prouvé et surmonté ces limitations, en montrant qu'une bien meilleure approche consiste à travailler sur des arbres linguistiques… C'était donc dix ans avant la sortie de ce volume. On regrette aussi qu'il n'y ait pas d'article sur le projet PARSEME, qui a produit des résultats vraiment très intéressants sur la détection et le traitement des EML en contexte multilingue depuis 2015.

Le côté positif de cette introduction est sa très grande richesse en références bibliographiques et autres pointeurs (projets, systèmes), en ce qui concerne la période 2000-2015 environ. Passons aux différentes parties.

Partie 1

L'étude du chapitre 1 concerne essentiellement les constructions du type verbe + nom en espagnol. Il y a une petite évaluation de la qualité des résultats de la traduction des quatre-vingt-dix-neuf exemples par le système enrichi avec le traitement des MLE. Dommage qu'il n'y ait aucune précision sur ce traitement (ni de référence au site Apertium, où on devrait pouvoir le trouver).

Le chapitre 2 n'est pas vraiment dans le thème de cette partie et commence mal, avec l'affirmation fausse et non justifiée que la SMT est meilleure que la RBMT pour traduire les EML. Plusieurs articles de ce volume prouvent que c'est le contraire. Les auteurs prétendent que les systèmes SMT (de type Moses ici) peuvent extraire les EML des corpus et donc avoir une meilleure couverture que les systèmes RBMT (qu'on devrait plutôt appeler systèmes fondés sur la connaissance linguistique, ou systèmes « experts », que systèmes à règles, passons sur cette confusion assez généralement répandue). C'est tout à fait faux, car bien des systèmes de TA « experts » incluent des dictionnaires de tournures, idiomes, termes techniques, etc. tout simplement « récupérés » de dictionnaires informatisés et de terminologies en ligne (comme iate.eu). À titre indicatif, ATLAS-II de Fujitsu avait, en 2009 environ, 6,5 M entrées dans ses dictionnaires, Systran entre 600 K et 800 K par langue, etc.). C'est beaucoup plus que ce que l'on trouve dans les tables de

traductions calculées par Moses sur, par exemple, les 20 M de phrases du corpus de l'EU (plus qu'EuroParl).

À part cela, l'étude avec des étudiants post-éditant des résultats de TA est intéressante, mais elle n'apporte rien à l'aspect TAL des EML.

Le chapitre 3 cherche à traiter les EML dans le cadre de la SMT à la Moses, sans utiliser l'approche hiérarchique qui seule permet de travailler sur des arbres (mieux, sur des bi-arbres). Donc les collocations sont toujours traitées dans le cadre d'une fenêtre de quelques mots (mots-formes), ce qui interdit par exemple de traiter convenablement les verbes à particule séparable et séparée de l'allemand et même de l'anglais. Quelques expériences sont présentées, mais la « mesure de qualité » est BLEU ou BLEU 2, et tout le monde devrait savoir, depuis le fameux article de Osborne, Callison-Burch et Koehn, que BLEU ne mesure PAS la qualité (que ce soit la qualité linguistique ou la qualité d'usage), et, pour plusieurs raisons théoriques, ne peut pas la mesurer.

Partie 2

Le chapitre 4 est vraiment très intéressant et à lire en détail. L'auteur a une très longue expérience du TAL pour la TA, depuis ses débuts à Siemens sur le système METAL, puis à Sietech (une filiale de Siemens), puis à GMS, Lernout&Hauspie, Sail Labs (toujours sur METAL), puis sur la version de LMT (IBM, McCord) améliorée par Linguatec (Heidelberg et Munich) en PMT, puis chez GMX… Le contexte est celui de la RI vers l'allemand, à partir de langues européennes, et de langues présentant une grande variabilité au niveau de l'usage des espaces (arabe, persan, pashto, turc). Il s'agit d'extraction d'informations multilingues (des mots-clés et des entités nommées), et pas de TA proprement dite (quoique les réponses doivent être fournies en allemand, mais ce ne sont pas des phrases). Cependant, ce sont bien les techniques efficaces en TA d'énoncés qui sont recherchées et présentées.

L'article est à la fois très érudit et très technique, avec des références précises aux niveaux X0, X1, X2 de la théorie X-barre. Les conclusions sont très intéressantes, et effectivement démontrées. (1) Il est important (et possible) de traiter de façon uniforme les unités lexicales, qu'elles soient simples ou complexes. (2) Le défi du traitement des EML se situe essentiellement en analyse et est monolingue, pas bilingue, le transfert lexical devant être le même, qu'il s'agisse d'unités simples ou complexes. (3) Dans les dictionnaires monolingues, les EML doivent avoir tous les attributs (traits) des unités lexicales (UL) simples, plus la spécification de la tête de l'EML et la liste des UL composantes, avec leur lemme et leur partie du discours (POS). (4) Il ne faut pas traiter les EML en prétraitement (approche « mots-avec-des-blancs »), ni après l'analyse (trop tard !), mais comme une partie intégrante de l'analyse. Du point de vue linguistique, l'approche consiste à ajouter un niveau au schéma X-barre, niveau où on appliquera les règles relatives aux EML. (5) Enfin, la décision du choix d'utiliser un sens non compositionnel (EML) ou un

sens compositionnel (synthétique) dépend du contexte et peut en pratique être fondée sur des scores (probabilistes, possibilistes ou préférentiels).

Le chapitre 5 pose le problème de la terminologie : sous prétexte que les mots composés allemands n'ont pas de séparateurs internes, on ne les considère pas comme de type EML. C'est d'autant plus injustifié qu'en allemand la même unité, un verbe à particule séparable, peut se présenter en deux mots (typographiques) séparés (« *er kam am Morgen zurück* ») ou collés (« *er ist am Morgen zurückgekommen* »).

Cela dit, le contenu de ce chapitre est intéressant. On utilise une convention pour noter la segmentation d'un mot composé en ses parties lexicales et dérivationnelles : « Brauch~er#schutz#dienst#list~ung » (service de protection du consommateur). Dommage que l'on n'aille pas jusqu'à une description hiérarchique (nécessaire pour la synthèse vocale d'ailleurs), comme : « [[Haupt#[bahn#hof]]#[[ge~päck]#[auf~be[#wahr~ung]]]] » (consigne à bagages de la gare principale).

Ces alignements sont intéressants car ils montrent que, comme dans les énoncés complets, ils ne sont pas un à un. Par exemple, « Korruption~s#be~kämpf~ung#ein~heit » est en français « unité anti-corruption » (plutôt que « unité de lutte contre la corruption », mais « *órgano contra la corrupción* » (« *contra* » vient de ε et « *bekämpfung* » va à ε). Au total, c'est une étude intéressante, mais pas directement reliée au thème de cette partie, ni d'ailleurs à celui de la TA.

Le but du chapitre 6 est ici d'obtenir de meilleurs alignements que quand on aligne des mots formes (mots typographiques). Le processus de segmentation est proche de celui de l'allemand décrit dans le chapitre précédent. Une amélioration est obtenue quand on aligne les « sous-mots » néerlandais avec les mots (typographiques) de l'autre langue (ici, anglais pour un corpus médical, français pour un corpus automobile). Le point intéressant est qu'une amélioration plus importante est obtenue en appliquant aux deux alignements obtenus une heuristique d'intersection, puis en fusionnant tous les points d'alignement, et enfin en ajoutant des points d'alignement venant du résultat du module grow-diag-final dans le cas du modèle d'alignement entraîné sur les alignements des sous-mots. Les F-mesures relatives à l'extraction de terminologies bilingues restent cependant assez faibles (les meilleures à 56 % ou 54 %). Cela nous renforce dans l'idée que, pour obtenir de bons résultats, il faut passer des alignements entre chaînes à des alignements entre arbres.

Partie 3

Le chapitre 7 est assez court, et n'apporte pas grand-chose de nouveau. Surtout, on voit bien que les règles de correspondance font intervenir un modèle d'alignement de chaînes, et pas d'arbres, alors que la thèse de V. Seretan datait de 2008, six ans avant la rédaction de cet article. Il y a bien un article de Seretan de 2011 cité dans le texte, mais pas pour le fond, seulement pour étayer le fait que

les collocations sont la majorité des ELM. Dans le même ordre d'idée, le chapitre devrait tout de même discuter les résultats présentés par Estelle Delpech, en 2013, dans sa thèse thèse (prix de l'ATALA). Elle y montre, en résumé, que l'extraction de terminologie technique bilingue (de qualité) à partir de corpus comparables ne fonctionne pas et ne peut pas fonctionner…

Le chapitre 8, une étude, assez spéculative sur « la possibilité d'utiliser des faisceaux lexicaux bilingues pour améliorer le degré de naturel et de fidélité (*textual fit*, littéralité ?) des textes traduits [automatiquement, par des systèmes Moses] ». Les faisceaux lexicaux sont des séquences de trois à sept mots ayant des fonctions discursives similaires, apparaissant dans un corpus comparable anglais-polonais de tracts d'information à destination de patients. À cause des différences translinguistiques, on applique de plus un certain nombre de critères formels pour filtrer les faisceaux dans chaque sous-corpus. Les résultats montrent que les faisceaux lexicaux bilingues extraits de corpus comparables ont un potentiel inexploré pour la TA, les aides à la traduction, et la lexicographie bilingue. Au total, on voit mal comment on pourrait, comme l'affirme l'auteur, intégrer les résultats de cette étude dans des systèmes de TA et de THAM (aide à la traduction).

Le chapitre 9 est une description très précise, intéressante et illustrée de la représentation et du traitement des tournures en croate à l'aide de NooJ. La représentation correspond à la partie « haute » des transducteurs, et les actions (balisage, traduction…) à la partie basse des transitions. Comme NooJ permet de traiter non seulement des transducteurs classiques (réguliers), mais aussi des réseaux de transition récursifs (équivalents aux grammaires hors contexte, mais plus puissants quand on utilise des contextes) permettant de travailler, en fait, sur des arbres, on a la possibilité de traiter des ELM non connexes.

Du point de vue de la traduction, ou d'autres applications multilingues, rien n'est dit. La bibliographie est assez courte, et on s'étonne de ne pas y trouver en bonne place au moins une référence à l'article de Šandor Dembitz *et al.* qui décrit une énorme collection allant jusqu'à des octogrammes, qui doivent contenir beaucoup de tournures.

Le chapitre 10 est une étude très détaillée, certes, mais on ne voit pas ce que cela apporte à la TA ou à la THAM. En fait, la méthode utilisée montre en un sens ce qu'il ne faut pas faire ! Ici, par exemple, on prend le verbe « faire » et on cherche à détailler tous ses sens possibles. Mais (et c'est dit dans le texte), son sens dépend le plus souvent du lexème en collocation à sa droite. Cela veut simplement dire que, le plus souvent, ce verbe est un verbe support.

L'étude linguistique présentée ici gagnerait donc à être structurée autour du ou des « sens propres» (comme créer, agir…), des FLS dont doći peut être une valeur, et de son ou ses rôles et sens possibles dans des idiomes (comme « faire un tabac », « faire le beau », « faire fort », etc.).

Le chapitre 11 est de loin le chapitre le plus intéressant et le plus avancé du volume. L'identification des collocations, la résolution des anaphores et la

traduction sont connues comme des problèmes parmi les plus difficiles du TAL. Ici, ils sont traités ensemble, et avec succès.

Comme dit plus haut, la meilleure méthode d'identification des collocations dans des corpus est due à É. Wehrli et son équipe. Ici, il s'agit de les identifier quand des éléments en sont élidés ou pronominalisés. Exemple : « *Paul broke the word record. He broke it by a large margin* ». Même si un système reconnaît les deux EML (*break* est verbe support de record, et large ∈ Magn(*margin*)), il produit actuellement au mieux : « Paul a battu le record du monde. Il l'a brisé avec un grand écart ». La solution présentée, implémentée dans l'analyseur Fips-5, consiste à faire évoluer une mémoire des centres de reprises anaphoriques possibles, et à la transmettre d'un énoncé au suivant au cours de l'analyse. On arrive alors à améliorer la traduction, et à produire : « Il l'a battu avec une grande marge ».

Ne serait-ce que pour ce dernier article, il est donc tout à fait intéressant pour un taliste de lire ce volume, tout du moins les chapitres signalés ici comme les plus intéressants.

Pierre M. NUGUES. Language Processing with Perl and Prolog. Theories, Implementation, and Application. *Springer*. 2014. 662 pages. ISBN 978-3-642-41463-3.

Lu par **Caroline BARRIERE**

Université d'Ottawa, École de science informatique et génie électrique

---

De prime abord, le livre de Nugues est sans contredit « intimidant ». C'est d'abord son titre qui m'a intimidée. Pourquoi Perl ou encore Prolog dans une ère où le Java ou encore plus le Python prévalent? C'est ensuite sa taille qui m'a intimidée, me retrouvant devant un volume de 650 pages sans savoir trop par où débuter.

Alors, j'ai débuté, par le début…. Le premier chapitre « *An Overview of Language Processing* » est vraiment un tour de force en tant que résumé de divers aspects du traitement automatique des langues (TAL). J'ai beaucoup aimé. De façon concise et claire, Nugues nous parle autant de syntaxe que de sémantique et de dialogues. Je recommande la lecture de ce premier chapitre à toute personne désirant avoir un aperçu du TAL.

Poursuivant ma lecture, j'ai constaté que, de façon soutenue, l'auteur écrit bien, de façon claire et précise, et qu'il structure ses chapitres de façon uniforme, présentant d'abord les concepts théoriques, suivis d'exemples pratiques, parfois d'implémentation en Prolog, terminant par une section de « *further reading* »et quelques exercices. J'apprécie cette cohérence.

Quant à l'ordre de lecture, j'ai rapidement abandonné un ordre linéaire. Même si l'auteur nous dit qu'il s'agit d'un livre de cours, j'ai plutôt trouvé que la structure du livre ne suivait pas tant un ordre naturel pour l'enseignement du TAL, mais couvrait

plutôt beaucoup de sujets différents, tels qu'on le verrait dans une encyclopédie. Un avantage de cette organisation encyclopédique est que nous pouvons lire les chapitres qui nous intéressent dans l'ordre que nous désirons.

Ma déception a été le manque de références vers des articles récents et des applications récentes. Il est dommage que cette deuxième édition de *Language Processing with Perl and Prolog*, datant de 2014 (la première édition datant de 2006), pointe encore vers plusieurs applications des années 90, voire même des années 80. Lors de la lecture des sections « *Further Reading* », ou même la lecture des sections applicatives, le manque de références récentes laisse un arrière-goût de désuétude. C'est d'autant plus dommage que l'auteur sait bien présenter et expliquer les concepts théoriques de tout temps, parlant autant d'automates à états finis, que de réseaux de neurones, que de modèles markoviens, ou que de grammaire de Chomsky. Non seulement la couverture théorique est large en termes de sujets et d'époques (des années 1970 à 2010), mais en plus, pour chaque sujet, Nugues sait donner juste assez d'informations pour la compréhension des principes sans se perdre dans les détails.

Comme tout livre en TAL, le champ est tellement vaste que l'auteur y injecte ses préférences… Le livre de Nugues a un biais vers la syntaxe. Sur l'ensemble des dix-sept chapitres du livre, on retrouve sept chapitres touchant l'analyse syntaxique Alors, les amoureux de syntaxe, c'est un livre pour vous!

Quant aux langages de programmation suggérés dans le titre du livre, Perl et Prolog, j'ai rapidement constaté que Perl était relégué à l'arrière-plan, n'étant présenté que dans une section discutant des expressions régulières, mais qu'en contrepartie, Prolog était omniprésent. En effet, presque chaque chapitre contient une section sur l'implémentation en Prolog des concepts présentés. De plus, une annexe « *An Introduction to Prolog* », de cinquante pages, pourrait faire en soit le contenu d'un excellent tutoriel sur l'utilisation de Prolog pour le TAL.

En conclusion, je pense que malgré mon intimidation initiale, j'ai su apprécier la clarté et la concision dont fait preuve Nugues sur les divers sujets. Je suggérerais aux lecteurs hésitants de ne pas se laisser dérouter par le titre mentionnant le Perl et le Prolog, et de profiter du contenu de ce livre plutôt comme un manuel de référence sur les concepts en TAL, que ce soit pour découvrir la sémantique, le dialogue, l'analyse de corpus, ou encore, plus certainement, l'analyse syntaxique.