# Introduction

Antonio Balvet
Université de Lille – UMR STL 8163
antonio.balvet@univ-lille.fr

## Des connaissances linguistiques pour les algorithmes de TAL

Les systèmes de Traitement Automatique de la Langue (TAL), tels que les analyseurs syntaxiques, les systèmes de reconnaissance vocale, de synthèse vocale à partir du texte, ou encore de traduction automatique, mobilisent des algorithmes, d'une part, et des connaissances linguistiques structurées en grande quantité, pour associer des représentations formalisées, manipulables par une machine, à des unités linguistiques à analyser. Si l'on accepte de considérer que la linguistique aidée de l'ordinateur – qu'on la nomme « linguistique computationnelle », « linguistique informatique » ou « Traitement Automatique des Langues » – s'est constituée dès les années 1960[1], la question peut se poser aujourd'hui de la place des connaissances linguistiques au sein de cette discipline, dans laquelle linguistique et informatique se retrouvent « en contact » (Cori et Marandin, 2001 : 49) depuis plus de soixante ans. Dans leur essai d'histoire rationnelle des liens entre programme générativiste et informatique, Cori et Marandin mettent en lumière comment, au même titre que des langues naturelles, linguistique et informatique ont connu « ignorance réciproque, rivalité plus ou moins belliqueuse, emprunt et collaboration » (p. 49). La rupture, entamée dès la fin des années 1970, semble aujourd'hui définitivement consommée entre grammaire générative et TAL. En effet, celui-ci s'est imprégné tout au long de son histoire des méthodes de l'ingénierie informatique, et encore plus depuis les années 2000 avec le retour en grâce de l'Intelligence Artificielle. Malgré une rupture que d'aucuns nommeraient schisme, TAL et grammaire générative partagent encore largement une conception de la langue, et donc de son traitement (humain comme machine) posant une dichotomie fondamentale entre « grammaire » (règles, algorithmes), d'un côté, et « lexique » (bases de connaissances lexicales) de l'autre. Autrement dit entre un dispositif de traitement (application de règles d'analyse/génération), modularisable en sous-tâches autonomes (découpage en unités lexicales de base, étiquetage en parties du discours, assemblage de constituants) et des ressources dites « lexicales », spécifiant les propriétés formelles (orthographe normalisée et variantes, notation phonétique), flexionnelles (genre, nombre, cas, temps, mode, aspect, etc.), syntaxiques (parties du discours) voire sémantiques (polarité, dénotations, aspect) des unités mobilisées au cours des traitements.

[1] L'Association pour le Traitement Automatique des Langues (ATALA) a été créée dès 1959 en France, l'Association for Computational Linguistique (ACL) dès 1962 aux États-Unis.

*Lexique, 27* (2020), 5-19.
ISSN : 0756-7138.

La vision atomiste et lexicaliste de la langue, défendue par la grammaire générative depuis l'origine, instaure une partition entre propriétés de structure et propriétés des unités atomiques mobilisées par ces structures. Cette vision entre en résonance avec les méthodes éprouvées de l'ingénierie informatique : réduire la complexité d'un problème en le découpant en sous-tâches ou modules (approche « diviser pour régner »), optimiser chaque sous-étape de traitement, identifier les éléments les plus atomiques, factoriser les traitements pour gagner en efficacité. Un langage de programmation n'est pas une langue naturelle, bien évidemment. Mais il s'agit néanmoins d'un langage formel dans lequel la « sémantique » (opérations informatiques, calculs) est dérivée de façon univoque de structures syntaxiques valides (des instructions), correctement analysées par un système informatique. Cette analogie entre l'analyse d'un code informatique en vue de l'exécution des instructions correspondantes et l'analyse des langues naturelles n'est pas fortuite : elle découle des influences mutuelles entre linguistique formelle et informatique, évoquées plus haut[2]. Encore aujourd'hui, malgré les évolutions technologiques récentes laissant augurer de possibles changements de paradigmes à venir, la plupart des systèmes de TAL peuvent être schématisés comme :

- une chaîne de traitements, des plus atomiques aux plus complexes, dans laquelle chaque étape produit de nouvelles informations structurées (représentations formalisées), de façon monotone (chaque traitement ne fait qu'ajouter une nouvelle connaissance), prévisible et dont la cohérence est formellement démontrable. Le système n'a alors typiquement accès qu'au niveau de structure immédiatement précédent ;
- à chaque étape des traitements, des règles (d'analyse ou de génération de structures bien formées) adaptées au niveau de structuration courant font appel à des connaissances structurées.

Cette schématisation, bien que rapide, permet de souligner à quel point la conception de la plupart des systèmes de TAL existants (ex. : étiqueteurs en parties du discours, analyseurs syntaxiques, générateurs de textes, moteurs de Traduction Automatique) suit les mêmes schémas que ceux de l'ingénierie informatique en général : modulariser, spécialiser et ordonner les traitements, dans une logique compositionnelle et itérative. Pour illustrer notre propos, prenons l'exemple d'un analyseur syntaxique, dont l'objectif est d'assigner à une phrase, tirée d'un texte tout venant (ex. : article journalistique, page web, mais également requête utilisateur, message électronique, etc.), une structure syntaxique explicite, formalisée et donc « calculable », pouvant servir de base à d'autres représentations (ex. : identification

[2] Ainsi, tout langage informatique est formellement défini par une grammaire de réécriture, souvent exprimée en « forme normale de Chomsky ». La transformation des instructions d'un programme en actions/opérations est dévolue à un compilateur, qui n'est ni plus ni moins qu'un analyseur syntaxique, chargé de calculer une représentation formelle univoque en langage-machine de bas niveau, à partir des « mots » (mots-clés, instructions de haut niveau) du programme.

des actants et de leur rôle thématique, déclenchement d'une action requise par l'utilisateur). Selon le schéma défini plus haut, un tel dispositif devra, dans l'ordre :

- segmenter le texte en unités linguistiques de base : phrases, mots graphiques (« tokens ») ;
- pour chaque phrase, transformer chaque mot graphique en unité lexématique de base en lui associant un ensemble de connaissances linguistiques explicites, tirées d'une base de connaissances (ex. : un lexique au format électronique). Ainsi, l'analyseur cherchera à déterminer la forme lemmatisée de chaque unité lexématique, ainsi que ses propriétés flexionnelles et sa partie du discours. À cette étape, les unités de base se trouvent donc enrichies d'un ensemble de propriétés linguistiques explicites, et forment un objet informatique de niveau supérieur à l'étape précédente ;
- pour chaque séquence d'unités identifiées à l'étape précédente, identifier les regroupements les plus probables en constituants syntaxiques, en consultant une base de connaissances (ex. : règles/modèle d'analyse) ;
- pour chaque constituant valide, identifier les sous-constituants recteurs et régis (ex. : têtes syntaxiques, noyaux verbaux) ;
- pour chaque constituant régi, typer la relation (fonction syntaxique) qui le lie à son élément recteur.

À chaque étape de traitement, le système prend des décisions en appliquant des règles (algorithme) tenant compte d'informations linguistiques explicites, tirées d'une base de connaissances. Bien que des analyseurs « neuronaux » en dépendances syntaxiques, tels que spaCy[3], semblent adopter des principes de traitement moins linéaires, tous les analyseurs syntaxiques probabilistes considérés comme des références dans la littérature, tels que l'analyseur en dépendances Malt Parser[4], la plate-forme UDPipe[5], ou les analyseurs syntaxiques (en constituants comme en dépendances) produits par le Stanford NLP group[6], adoptent un schéma des traitements similaire à celui présenté plus haut. À titre illustratif, nous rassemblons dans les Figures 1 à 3 ci-dessous illustrent les propositions d'analyses de l'analyseur syntaxique en dépendances UDPipe (Straka, 2018), développé à l'Institute of Formal and Applied Linguistics (Charles University, République Tchèque)[7]. Dans les trois propositions d'analyse présentées, le modèle french-sequoia-ud-2.6-200830, paramétré à partir du corpus Sequoia (Candito & Seddah, 2012), a été activé afin d'associer à chacune des phrases de tests volontairement rédigées dans

[3] https://spacy.io/

[4] http://www.maltparser.org/

[5] http://ufal.mff.cuni.cz/udpipe (Straka, 2018).

[6] https://nlp.stanford.edu/software/lex-parser.html

[7] L'analyseur est disponible sous la forme de programmes exécutables classiques, ainsi que d'un service web (https://lindat.mff.cuni.cz/services/udpipe/) qui permet d'exécuter des analyses automatiques en activant l'un des modèles d'analyse disponibles.

un style « journalistique » une analyse en dépendances universelles[8]. Les visualisations sont assurées par Arborator[9], un outil développé par K. Gerdes (Université Paris-Saclay, Limsi-CNRS).

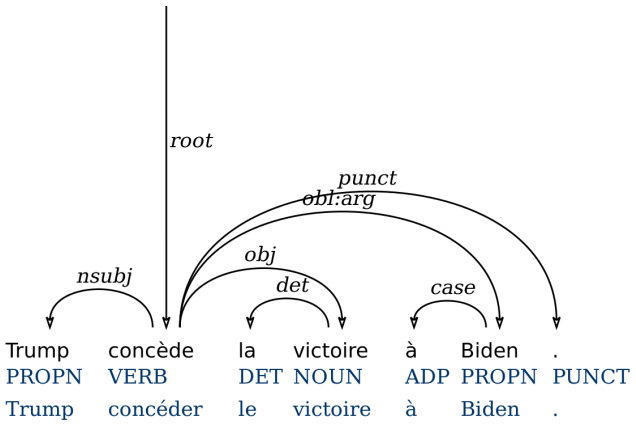


Figure 1. Analyse syntaxique en dépendances d'une construction ditransitive.

Dans les Figures 1 à 3 sont visibles les différents niveaux d'analyse considérés par le programme d'analyse syntaxique, respectivement de haut en bas : le niveau des unités lexicales, puis celui des parties du discours, et enfin les formes lemmatisées. Une fois les unités lexicales de base repérées et associées à leur partie du discours, les paires contrôleur/dépendant sont marquées, selon les fonctions distinguées dans le sous-ensemble des dépendances universelles. Comme on peut le voir, l'analyse pour « Trump accorde la victoire à Biden » est syntaxiquement correcte et conforme aux principes de l'analyse en dépendances universelles : les deux arguments du noyau verbal *accorder* sont correctement identifiés, le premier étant marqué « obj » (complément d'objet), le second « obl:arg » (argument oblique, soit complément d'objet indirect). La préposition « ADP » n'est pas la tête du syntagme prépositionnel, toutefois elle déclenche une marque de cas. Enfin, tant « Trump » que « Biden » sont ici correctement identifiés comme des noms propres « PROPN », alors qu'il est peu probable que les échantillons de paramétrage tirés de Sequoia aient mentionné ces deux personnalités. Ceci démontre la capacité des analyseurs actuels à intégrer, ou à « apprendre », de façon autonome, certaines généralités structurelles[10].

[8] Le projet Universal Dependencies (https://universaldependencies.org/fr/dep/) vise à proposer un cadre descriptif et théorique pour l'analyse en dépendances syntaxiques de nombreuses langues du monde.

[9] Disponible sur https://arborator.ilpga.fr/q.cgi.

[10] Précisons toutefois que la détection et le typage des noms propres constitue une tâche à part entière en TAL, qui repose en grande partie sur des Ressources Lexicales Électroniques (des listes structurées) tout autant que des indices typographiques et structurels.

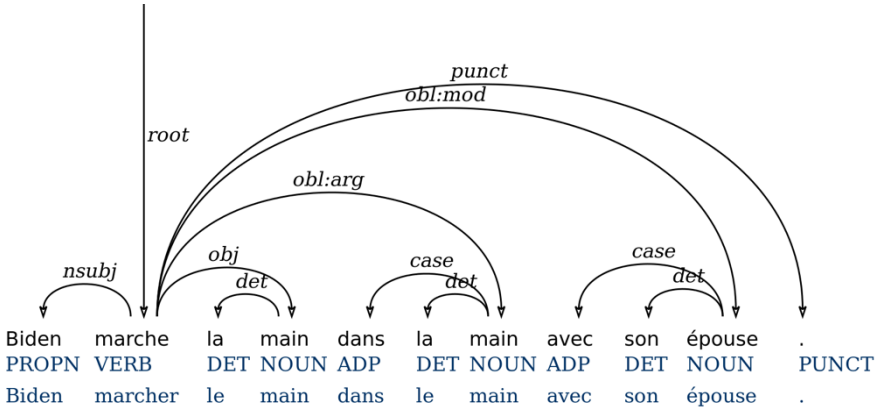


Figure 2. Analyse syntaxique en dépendances d'une unité polylexicale couplée à un verbe peu fréquent

Autant la première phrase pouvait être considérée comme correctement analysée[11], autant la deuxième phrase pose plus de difficultés à l'analyseur, comme le montre la Figure 2. Le modifieur adverbial *la main dans la main* est ainsi incorrectement segmenté en deux sous-séquences, dont la première moitié est considérée comme un argument objet de *marcher*, alors que *dans la main* est annotée exactement de la même façon que *à Biden* dans la phrase précédente : un argument oblique, dont la marque de cas est déclenchée par la préposition. En d'autres termes, l'analyseur considère ici *marcher* comme un noyau verbal à trois arguments : un sujet, un objet direct et un objet indirect. L'erreur d'analyse est due non seulement à la présence d'une unité polylexicale, mais également à la présence d'un noyau verbal relativement moins fréquent, dans les corpus de paramétrage utilisés, que *concéder*. En effet, n'importe quelle phrase construite avec *marcher* suivi de modifieurs prépositionnels (*dans la rue*, etc.) produit la même analyse : des arguments obliques, alors que le noyau verbal *manifester*, par exemple, produit les analyses attendues (Figure 3)[12].

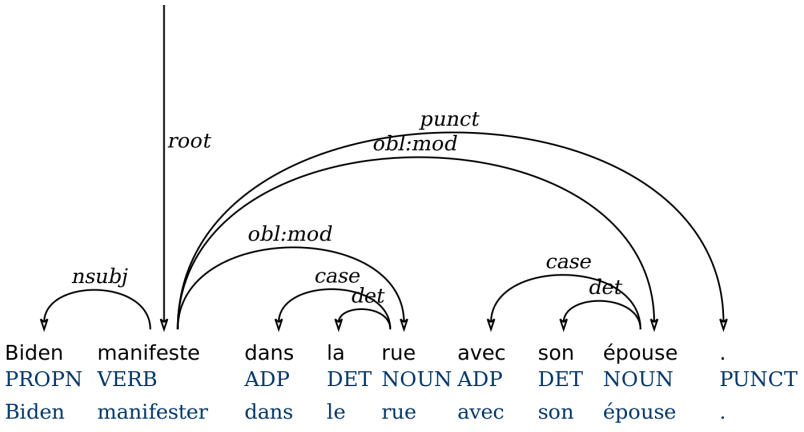


Figure 3 : Effet du choix du noyau verbal sur l'analyse d'une phrase

[11] Conformément aux principes de l'analyse en dépendances, aucune structure syntaxique hiérarchisée n'est proposée (arbre), tous les arguments du noyau verbal sont représentés au même niveau.

[12] Pour un traitement plus classique de la valence verbale, voir la présentation de la grammaire FrenchTAG dans la soumission de A. Savary et collègues (ce numéro).

Ces exemples d'analyse automatique illustrent le potentiel des outils logiciels développés au cours des deux dernières décennies, qui intègrent des algorithmes d'apprentissage automatiques sophistiqués, afin de généraliser les règles d'annotation à partir d'exemples d'analyses de référence. Toutefois, ils illustrent également le fait que des propriétés déterminantes, telles que la structure argumentale, ou à défaut la valence d'un noyau verbal, sont potentiellement inaccessibles à ces systèmes[13].

Ce qui vaut pour l'analyse syntaxique vaut également pour d'autres niveaux de traitement : que ce soit pour la segmentation en unités linguistiques de base, l'étiquetage en parties du discours, l'identification des actants et de leur rôle thématique pour l'extraction d'informations structurées à partir de texte, ou l'identification de la polarité (positive, neutre ou négative) d'un constituant, tout système de traitement de la langue doit faire appel à des connaissances linguistiques structurées, autrement dit des « Ressources Lexicales Électroniques » (désormais RLE). Comme dans tout domaine de technicité élevée, dans lequel un rendement important est attendu (en termes de volume et de qualité des traitements), le TAL est le lieu d'une spécialisation des tâches : certains développeurs concentrent leurs efforts sur les algorithmes principaux (le « moteur »), alors que d'autres se consacrent à la constitution, la structuration, l'annotation et l'évaluation de ressources lexicales. C'est ainsi que la totalité des analyseurs syntaxiques automatiques de référence intègrent désormais des algorithmes d'apprentissage automatiques indépendants des langues particulières, alors que les systèmes dits « symboliques » développés jusqu'à la fin des années 1990, à base de règles écrites par des linguistes informaticiens pouvaient être spécifiques à chaque langue. Les analyseurs génériques doivent toutefois être paramétrés, pour les langues à traiter, sur des corpus de référence annotés manuellement, tels que le Penn Treebank (Taylor et al., 2003) pour l'anglais américain, le French Treebank (Abeillé et al., 2003) ou le corpus Sequoia (Candito & Seddah, 2012) pour le français. Autrement dit, loin de s'affranchir de connaissances linguistiques, les systèmes contemporains de traitement de la langue ont, plus que jamais, besoin d'exemples d'annotations, ainsi que de RLE à large couverture, tout en restant précises, standardisées et inter-opérables. Le défi à relever pour le TAL semble désormais moins de développer de nouveaux algorithmes dédiés au traitement de la langue, que de proposer des méthodes efficaces, robustes et indépendantes des langues pour constituer, vérifier, et étendre des RLE de bonne qualité. L'autre défi, qui découle de l'inter-opérabilité et de l'inter-connexion nécessaire aux infrastructures techniques qui nous offrent des services de traitement de la langue activables à l'envi (smartphones, assistants intelligents), est la maintenance, dans la durée, de vastes ensembles de données linguistiques, complexes et fortement structurées. Faire exister dans la durée des données structurées, des corpus de référence n'est pas chose aisée. En effet, combien de lexiques syntaxiques ou sémantiques, de bases de données morphologiques, de réseaux lexicaux sont-ils aujourd'hui tombés dans l'oubli, soit parce qu'ils étaient la propriété d'une entreprise privée, soit parce que ces données, bien que développées par des organismes de recherche publics, étaient tributaires d'un format

[13] En effet, il est impossible de déterminer si ces propriétés pourront être généralisées un jour, moyennant un nombre suffisant d'exemples d'annotation.

propriétaire ou d'un logiciel fermé (ex. : bases de données) ? Le paysage du TAL se trouve aujourd'hui bouleversé par l'adoption, y compris par de grands groupes privés, de normes et standards internationaux, tant la valeur ne se crée plus sur la propriété des données mais bien sur l'adoption, par le grand public, d'écosystèmes technologiques dans lesquels les services linguistiques constituent un argument de vente parmi d'autres.

Comme évoqué plus haut, le TAL a, dès l'origine, intégré des méthodes et approches issues de l'ingénierie informatique, auxquelles s'ajoutent désormais des méthodes de la science des données : algorithmes d'apprentissage automatique sophistiqués, méthodes d'optimisation poussées, de manière à réduire les quantités de données de référence nécessaires, infrastructures matérielles avancées (serveurs de calcul en réseau, processeurs parallèles) et méthodes complexes de gestion des flux de données. On assiste donc à une convergence technologique qui suit, de fait, le mouvement général de l'ingénierie informatique. Cette diffusion sans précédent des méthodes d'ingénierie informatique, rendue possible par un mouvement généralisé de « libération » du code informatique, y compris chez des acteurs privés (Microsoft, Google, Facebook et Amazon en tête), s'est accompagnée d'une évolution des modes de traitement, et donc de la nature et de la logique de structuration des ressources lexicales nécessaires. Aux traditionnels « lexiques électroniques », c'est-à-dire des listes structurées d'informations linguistiques associées à une entrée, s'ajoutent désormais :

- des réseaux lexicaux inter-connectés (*Linked Lexical Resources*), conçus pour le web sémantique[14]. Exprimées dans un formalisme logique normalisé et standardisé[15], les connaissances encyclopédiques et linguistiques sont destinées à former des ressources dématérialisées, exploitables par tout service du web sémantique ;
- des modèles statistiques standard en apprentissage automatique : notamment Machines à Vecteur Support (*Support Vector Machines*), offrant des performances supérieures aux classiques chaînes de Markov (*Hidden Markov Models*) des années 1990, pour des tâches de catégorisation (ex. : classification d'images, reconnaissance de formes). Les SVM ont ainsi beaucoup été exploitées en TAL, pour des tâches de classification de mots ou de messages selon leur polarité, ou encore pour l'étiquetage en parties du discours ;
- des modèles dits « distribués », ou *word embeddings* (*vecteurs de mots*, *plongements lexicaux* en français) : essentiellement, des représentations optimisées des profils distributionnels des unités lexicales, extraites de corpus de très grande taille[16]. Ces

[14] Une présentation des enjeux du web sémantique pour le TAL serait hors de propos ici. Indiquons simplement que la vision de T. Berners-Lee, l'un des concepteurs de l'architecture du web, consiste en un monde où les machines seraient à même d'analyser tout le contenu (dont les textes) disponible sur les réseaux (Berners-Lee & Fischetti, 2001 ; Hendler & Berners-Lee, 2010).

[15] Par exemple : Web Ontology Language (OWL), Resource Description Framework (RDF).

[16] Plusieurs milliards de mots, collectés sur le web. Les plongements lexicaux s'appuient sur une représentation vectorielle des mots dans un espace à *n* dimensions (300 ou plus). Cette représentation vectorielle permet

ressources permettent, entre autres, d'identifier des relations lexicales entre lexèmes, sur la base de leurs contextes d'occurrence : termes sémantiquement proches, hyperonymes, hyponymes, termes morphologiquement apparentés ;

- des modèles neuronaux dits « profonds » : *à convolution*, *récurrents*, ou *récursifs*, selon l'architecture des différentes couches de neurones. Dans ces approches, c'est la capacité à généraliser des représentations complexes à partir de données particulières, incomplètes, voire bruitées, qui est principalement recherchée. Après avoir fait leurs preuves pour des tâches de reconnaissance de formes, notamment l'identification de visages sur des photos ou la reconnaissance des panneaux de signalisation pour les véhicules autonomes, ces modèles neuronaux sont de plus en plus exploités pour des tâches réputées complexes en TAL, comme la Traduction Automatique[17].

## Des RLE pour le traitement du français

Bien que l'appel à soumissions du présent numéro ne ciblait pas spécifiquement le français, les soumissions ici rassemblées présentent des ressources, méthodes et formalismes pensés principalement pour l'outillage de la langue française. Ce numéro de *Lexique* permettra donc, nous l'espérons, de mieux faire connaître les ressources et méthodes qui y sont présentées en priorité à un public de chercheurs travaillant sur la langue française, sans avoir jusqu'à présent intégré des RLE dans leur boîte à outils, en raison de leur complexité technique et du volume de données disponibles. Nous pensons, bien entendu, aux collègues linguistes désireux d'exploiter des corpus au format électronique et des services de Traitement Automatique des Langues, mais également aux traducteurs cherchant des alternatives aux classiques thesaurus (au format papier ou électronique) pour alimenter leurs mémoires de traduction, aux psycholinguistes et psychologues à la recherche de base de données lexicales de référence pour leurs expérimentations, aux sociologues cherchant à exploiter des questionnaires d'enquêtes, et au-delà, à l'ensemble des chercheurs en Humanités Numériques, travaillant sur le matériau linguistique textuel. Nous n'oublions pas les collègues TAListes, à la recherche de ressources de qualité pour le traitement du français, qui pourraient être rebutés de prime abord par des ressources dans lesquelles une part de travail manuel reste présente.

Nous avons choisi de présenter ces soumissions en regroupant, en premier lieu, deux projets de réseaux lexicaux : la version mise à jour du *Dictionnaire Électronique des Synonymes* du CRISCO, et le réseau lexical JeuxDeMots. En effet, bien que construits selon deux approches complémentaires[18],

d'identifier les « voisins » d'un mot donné (mots sémantiquement liés), sur la base de leurs profils distributionnels.

[17] Dans ces modèles, la frontière entre représentations structurelles et lexicales tend à s'estomper.

[18] Compilation de dictionnaires existants puis traitements algorithmiques pour l'un, intégration des réponses à des « jeux à objectifs » puis extension et vérification automatiques pour l'autre.

ces deux projets constituent des ressources dont le potentiel nous paraît insuffisamment exploité, tant en TAL que dans les domaines connexes de la linguistique outillée et de la linguiste de corpus. Les réseaux lexicaux électroniques, dont le parangon est le Princeton WordNet (Fellbaum, 1998), constituent, avec les lexiques électroniques morphosyntaxiques, tels que les dictionnaires électroniques du français élaborés au Laboratoire d'Automatique Documentaire et Linguistique (Courtois & Silberztein, 1990) l'exemple prototypique de ce que nous nommons « Ressources Lexicales Électroniques ». En effet, ces bases de connaissances encodent, au niveau lexical, soit des relations sémantiquement typées vers d'autres lexèmes, soit des propriétés linguistiques renseignées de façon systématisée et structurée (ex. : forme lemmatisée, partie du discours, genre, nombre, etc.). En tant que connaissances linguistiques exprimées au niveau lexical, elles sont typiquement amenées à être consultées par différents algorithmes au cours de la chaîne des traitements schématisée plus haut.

Dans « Les vedettes du *Dictionnaire Électronique des Synonymes* et les relations d'adjacence entre leurs synonymes », L. Chardon et J. François présentent les dernières évolutions du réseau lexical constitué au laboratoire CRISCO (Université de Caen) à la fin des années 1990 par la compilation de sept dictionnaires sources. Cette base a, par la suite, été progressivement complétée par des propositions venant d'internautes, contrôlées et vérifiées avant d'être intégrées à la ressource. En ce sens, le DES constitue un réseau sémantique en constante évolution, qui intègre une dimension participative, sans toutefois aller jusqu'à proposer des « jeux à objectifs » comme dans la ressource JeuxDeMots, ni rémunérer des contributeurs. Le fait que la ressource ait été construite à partir d'une base de référence, à savoir des dictionnaires de synonymes du commerce, offre une garantie quant à la qualité et à la couverture des relations lexicales initiales qui s'y trouvent. Toutefois, comme le reconnaissant les auteurs : « le DES a été conçu jusqu'à présent comme le réceptacle d'une multitude d'intuitions sur la proximité entre des dizaines de milliers de paires de mots en termes de fusion et non d'intersection. Le résultat est un noyau de jugements de proximité largement partagés et une périphérie de jugements isolés. » Il s'avère donc nécessaire, après plus de vingt ans d'existence, de proposer des méthodes objectivables et reproductibles pour piloter la gestion de la ressource qui, bien que relativement moins dense et volumineuse que d'autres réseaux lexicaux (JeuxDeMots, mais également WOLF, BabelNet ou DBPedia) n'en reste pas moins suffisamment complexe pour exclure un travail d'édition uniquement manuel des liens sémantiques.

La contribution de L. Chardon et J. François présente plus particulièrement une approche tirant parti de la topologie du réseau sémantique ainsi construit, afin d'améliorer la qualité de la ressource, en identifiant par exemple les nœuds isolés. La topologie des sous-graphes, c'est-à-dire des sous-ensembles de nœuds lexicaux plus ou moins fortement connectés, était déjà exploitée dès l'origine du projet : les nœuds lexicaux se trouvant dans une configuration de « clique »[19] permettaient de proposer

[19] C'est-à-dire un sous-graphe dans lequel tous les points sont connectés, soit directement, soit par transitivité de la relation de synonymie.

une hiérarchie de « meilleurs » synonymes. Les configurations correspondant à des « composantes connexes », c'est-à-dire des sous-graphes les plus fortement connectés à l'intérieur d'un nuage de mots, ont également été exploitées, en ce qu'elles mettent en évidence des facettes de sens (ou sous-sens) des lexèmes de départ. Dans l'article de L. Chardon et J. François, des opérations de transformation des sous-graphes du réseau lexical permettent, par ex., de révéler de façon automatique deux familles de sens pour *campagne* : une composante « stratégique » (*guerre*, *expédition*, *offensive*, etc.) et une composante « topographique » (*nature*, *champ*, *cambrousse*, etc.), qui émergent après élimination des « synonymes précaires », c'est-à-dire des nœuds lexicaux faiblement connectés à la vedette. D'autres manipulations permettent, plutôt que d'éliminer des synonymes dits « précaires », de proposer de nouveaux liens : des synonymes « probables », par l'examen des connexions primaires et secondaires entre nœuds du réseau. En d'autres termes, la topologie du graphe sémantique est exploitée dans le but de faire évoluer la ressource, en garantissant la cohérence des nouveaux liens, tout en motivant l'élagage du graphe de mots.

De son côté, « JeuxDeMots : un réseau lexico-sémantique pour le français, issu de jeux et d'inférences » de M. Lafourcade et N. Le Brun présente le réseau lexical ouvert et gratuit le plus étendu, et le plus dense en termes de relations lexicales, pour la langue française[20]. M. Lafourcade et N. Le Brun distinguent réseau sémantique et réseau lexico-sémantique : « [u]n réseau lexico-sémantique est un réseau qui permet de relier des connaissances du monde à des informations lexicales sur le vocabulaire qui les véhicule ». Autrement dit, le réseau JeuxDeMots est conçu comme un objet informatique représentant à la fois des connaissances encyclopédiques et des relations sémantiques, voire morphologiques, entre unités lexicales. Le réseau, construit sur une fondation constituée de réseaux lexicaux tels que le *Dictionnaire Électronique des Synonymes*, ou encore Verbaction (Hathout et al., 2002), (Tanguy & Hathout, 2002), a, depuis son lancement en 2007, été étendu et complété grâce à de nombreux dispositifs inspirés des « jeux à objectifs » (*Games With A Purpose*, ou GWAP). En d'autres termes, JeuxDeMots fait appel à « l'intelligence des foules » ou « peuplonomie » (*crowdsourcing*), sans toutefois recourir au principe des micro-tâches de plateformes telles que Amazon Mechanical Turk. En ce sens, JeuxDeMots constitue, avant tout, une expérimentation réussie dans le domaine de la constitution de ressources collaboratives, élaborée par collecte indirecte. En effet, les concepteurs de JeuxDeMots, qui est à la fois une ressource et une plate-forme web de jeux à objectifs à finalité lexicographique électronique, prennent soin de présenter les relations lexicales autour desquelles les jeux sont construits en évitant de présenter aux joueurs un vocabulaire métalinguistique (ex. : « hyponyme », « méronyme », etc.). Une fois les propositions de joueurs

[20] Nous considérons à part des ressources telles que le WOLF (Sagot & Fišer, 2008), ou BabelNet (Navigli & Ponzetto, 2012), essentiellement constituées par traduction automatique d'autres ressources et réseaux lexicaux, dont le Princeton WordNet (Fellbaum, 1998) et Wikipedia. Quant au Réseau Lexical du Français (Polguère, 2013), bien que les données soient disponibles (https://www.ortolang.fr/market/lexicons/lexical-system-fr), il est annoncé comme « un projet en cours » destiné à être révisé et augmenté régulièrement.

collectées, des processus de vérification de la cohérence des relations lexicales proposées, de correction, voire de propagation d'informations lexicales en tirant parti de la structure du réseau sont déclenchées.

Outre la présentation de la ressource elle-même, la soumission de M. Lafourcade et N. Le Brun expose les approches « ludo-contributives » appliquées au problème de l'acquisition et de la maintenance d'une base de connaissances linguistiques dense et étendue. Le lecteur intéressé pourra donc puiser dans cette contribution des éléments méthodologiques quant au recours aux jeux à objectifs pour la recherche, et au contrôle de la qualité des propositions des joueurs.

Les contributions précédentes mettaient l'accent sur l'enrichissement contrôlé de liens (sémantiques au sens large, mais également morphologiques) entre les nœuds d'un réseau sémantique. Toutefois, bien que les expressions polylexicales (*multiword expressions*) soient présentes aussi bien dans le *Dictionnaire Électronique des Synonymes* du CRISCO que dans JeuxDeMots, leur caractérisation, ainsi que leur représentation formalisée ne font pas l'objet d'une attention spécifique. Pour cette raison, nous avons choisi de présenter, à la suite des deux soumissions présentant des réseaux sémantiques, les travaux de A. Savary et collègues, qui abordent spécifiquement la problématique de la représentation formalisée des expressions polylexicales, habituellement délaissées au profit des unités simples ou, dans le meilleur des cas, des mots complexes[21].

La contribution de A. Savary, S. Petitjean, T. Lichte, L. Kallmeyer et J. Waszczuk, « Object-oriented lexical encoding of multiword expressions: Short and sweet » met l'accent sur ces expressions polylexicales, qui posent des difficultés à la fois de représentation et de traitement aux approches lexicalistes en TAL. A. Savary et ses collègues proposent une approche adaptée aussi bien aux expressions polylexicales continues (mots complexes) que discontinues (constructions à verbe support, unités phraséologiques, expressions idiomatiques figées), en recourant à une grammaire formelle LTAG pour représenter ces unités à la frontière entre syntaxe régulière et figement, tout en réglant les inévitables contraintes d'accord, de conjugaison et d'ordre apparent des constituants de ces unités complexes.

La formalisation proposée par A. Savary et collègues constitue un type de base de connaissances linguistiques formalisées distinct des réseaux lexicaux présentés plus haut, en ce qu'elle cherche à représenter, de façon non redondante, des unités dans lesquelles « le principe de l'idiome » (Sinclair, 2001) amène à devoir dépasser la dichotomie stricte grammaire / lexique. Les unités traitées représentent un défi en TAL en ce que certains sous-constituants sont fixes (et définitoires de l'expression polylexicale), alors que d'autres présentent une variabilité dans leur forme : soit que ces

[21] Par leur degré de figement (séquences de mots contiguës) et la présence d'indices typographiques explicites (tiret, en particulier), les mots complexes sont relativement plus « faciles » à identifier dans un texte, et à représenter du point de vue de la lexicographie électronique, que les expressions polylexicales non contiguës, ou qui sont le lieu de phénomènes d'accord ou de flexion.

sous-constituants soient sujets à des phénomènes d'accord ou de conjugaison, ou que leur ordre apparent puisse être modifié. À ces contraintes internes de forme s'ajoutent des phénomènes d'ambiguïté entre expressions polylexicales et constituants syntaxiques relevant de « principe du choix ouvert », dans certains contextes. La réponse apportée par A. Savary et collègues est une preuve de concept de la faisabilité de l'approche décrite. Elle fait appel au formalisme des grammaires d'arbres adjoints lexicalisés (LTAG, *Lexicalized Tree-Adjoining Grammars*) dérivé du formalisme TAG (*Tree Adjoining Grammar*) de (Joshi & Schabes, 1997). Les auteurs s'appuient sur eXtensible MetaGrammar (XMG) présenté dans (Crabbé et al., 2013 ; Petitjean et al., 2016), une méta-grammaire permettant de compiler des analyseurs syntaxiques spécifiques à partir de métarègles abstraites. Dans l'approche de Savary et collègues, les formes simples et complexes des entrées lexicales sont dérivées par le biais de mécanismes de fusion/extension de sous-arbres lexicalisés, afin de produire une ressource lexicale conforme à l'orientation-objet (héritage, surcharge), visant la précision sans toutefois perdre en concision. Bien que ces travaux fassent appel à des outils formels relevant du courant « symbolique » en TAL[22], soulignons ici que les expressions polylexicales ont pu être considérées comme « a pain in the neck for NLP » par Sag et al. (2002). Elles semblent « [s]till a pain in the neck for NLP », d'après Shwartz & Dagan (2019). A. Savary co-anime, par ailleurs, l'action PARSEME dédiée aux expressions polylexicales (Savary et al., 2017), l'approche proposée est évaluée sur un corpus de référence d'expressions polylexicales (Savary et al., 2018), ce qui donne à cette contribution une hauteur de vue appréciable. Là encore, nous espérons contribuer à mieux faire connaître à la fois les problèmes descriptifs et théoriques posés par ces expressions polylexicales, dont le degré de figement vient bouleverser les conceptions habituelles du « mot », et les solutions apportées par Savary et collègues qui, bien que formulée dans le cadre d'une grammaire LTAG, est pensée pour être adaptable dans n'importe quel cadre formel.

Nous clôturons ce numéro par la contribution de F. Sajous et collègues, qui constitue à la fois une présentation des ressources développées à CLLE-ERSS depuis plus de dix ans, et une prise de position sur les stratégies de constitution de RLE à partir de sources de connaissances lexicographique (semi) structurées, telles que le Wiktionnaire.

F. Sajous, B. Calderone et N. Hathout présentent, dans « Extraire et encoder l'information lexicale de Wiktionary : quel boulot pour étrangler le goulot ! », une approche visant à acquérir des ressources lexicales libres en s'affranchissant en grande partie du travail manuel, sans pour autant tomber dans les écueils des approches misant sur la quantité plutôt que sur la qualité des connaissances linguistiques. En effet, les auteurs font le constat que « la situation en matière de ressources disponibles pour le TAL ne s'est guère améliorée en dix ans. Ces dernières sont toujours soit inexistantes, soit inadaptées ou invisibilisées par leur statut légal. » Ils proposent donc d'exploiter le contenu semi-structuré du dictionnaire collaboratif Wiktionary, en allant à rebours des tendances actuelles, qui ont

[22] Par opposition au courant « apprentissage automatique ».

plutôt recours à des giga-corpus, faiblement structurés, et toujours plus grands, ou encore à des plateformes de micro-tâches et d'annotation par les foules (*crowdsourcing*). Ils proposent ainsi une approche autant qu'une stratégie, exploitable pour toute langue disposant d'un Wiktionnaire, afin d'automatiser l'assignation de propriétés de forme (notation phonétique, orthographe), d'informations morphologiques flexionnelles, et de propriétés syntaxiques aux unités lexématiques. Leur approche permet également d'exploiter les définitions présentes dans les articles du Wiktionnaire, et donc de tirer parti des distinctions de sens, marques d'usage et autres informations lexicographiques détaillées, absentes de la plupart des ressources induites à partir de grands volumes de textes[23]. L'approche proposée permet, en outre, de dépasser les lacunes des ressources constituées à partir d'une autre ressource collaborative, l'encyclopédie Wikipédia[24].

Au-delà des ressources[25] ainsi créées pour le français, l'italien et l'anglais, les auteurs font également un bilan de plusieurs décennies de travaux visant à dépasser les limites de la saisie manuelle d'informations structurées, et font le constat de l'indigence généralisée de ressources lexicales, y compris pour des langues bien pourvues comme l'anglais : « les ressources les plus élémentaires, particulièrement celles qui sont libres, sont encore inexistantes pour la plupart des langues ». Cette contribution est donc autant la présentation d'une méthode applicable à toute langue, pour peu qu'on dispose d'un fonds d'articles du Wiktionnaire, qu'un engagement concret en faveur de la mise à disposition de ressources lexicales de qualité, ouvertes et réexploitables, y compris au-delà des frontières du TAL[26]. Pour toutes ces raisons, nous avons choisi de clôturer ce numéro de la revue *Lexique* par cette contribution, qui rejoint par bien des aspects notre propre constat. Nous espérons, par ce numéro, contribuer à combler le fossé entre producteurs de ressources lexicales et utilisateurs finaux, tout en donnant aux ressources existantes, mais parfois méconnues, la visibilité qu'elle méritent.

## Bibliographie

Abeillé, A., Clément, L., & Toussenel, F. (2003). Building a Treebank for French. In A. Abeillé (Ed.) *Treebanks* (pp.165-187). Dordrecht, Springer.

Berners-Lee, T., & Fischetti, M. (2001). *Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor*, DIANE Publishing Company.

[23] Qui constituent de ce fait plus des lexiques structurés que des dictionnaires électroniques.

[24] Wikipédia étant, par nature, une ressource encyclopédique, les informations qui s'y trouvent ont essentiellement trait aux propriétés des objets du monde, et non à leurs propriétés linguistiques.

[25] Et des outils de traitement des différentes versions du Wiktionnaire.

[26] Les auteurs ont également développé le PyschoGLÀFF (http://redac.univ-tlse2.fr/lexiques/psychoglaff.html), un « lexique à tout faire » pour les psychologues.

Candito, M., & Seddah, D. (2012). Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. *TALN 2012 – 19e conférence sur le Traitement Automatique des Langues Naturelles*, Juin 2012, Grenoble, France. http://talnarchives.atala.org/TALN/TALN-2012/index.html

Cori, M., & Marandin, J. M. (2001). La linguistique au contact de l'informatique : de la construction des grammaires aux grammaires de construction. *Histoire Épistémologie Langage*, *23*, 49-79.

Crabbé, B., Duchier, D., Gardent, C., Le Roux, J., & Parmentier, Y. (2013). XMG: extensible metagrammar. *Computational Linguistics*, *39*(3), 591–629. https://doi.org/10.1162/COLI_a_00144.

Fellbaum, F. (Ed.) (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Goldberg, A. E. (1995). *A Construction Grammar Approach to Argument Structure*. Chicago Press.

Hathout, N., Namer, F., & Dal, G. (2002). An Experimental Constructional Database: The MorTAL Project. In P. Boucher (dir.), *Many Morphologies* (pp. 178-209). Cascadilla, Somerville, Mass.

Hendler, J., & Berners-Lee, T. (2010). From the Semantic Web to social machines: A research challenge for AI on the World Wide Web. *Artificial intelligence*, *174*(2), 156-161.

Navigli, R., & Ponzetto, S. (2012). BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*, *193*, 217-250.

Joshi, A. K., & Schabes, Y. (1997). Tree-adjoining grammars. In G. Rozenberg & A. Salomaa (Eds), *Handbook of Formal Languages,* Vol. 3: *Beyond Words* (pp. 69–123). Berlin: Springer.

Petitjean, S., Duchier, D., & Parmentier, Y. (2016). XMG 2: Describing description languages. In M. Amblard, P. de Groote, S. Pogodalla & C. Retoré (Eds.), *Logical Aspects of Computational Linguistics. Celebrating 20 Years of LACL (1996-2016) – 9th International Conference, LACL* (pp. 255–272). Nancy, France / Berlin: Springer. https://doi.org/10.1007/978-3-662-53826-5_16

Polguère, A. (2013). Tissage du Réseau Lexical du Français (RLF) : buts et méthodes. In É. Buchi, J.-P. Chauveau & J.-M. Pierrel (dir.), *27ème Congrès International de Linguistique et de Philologie Romanes (CILPR 2013)*, Nancy, France. http://www.atilf.fr/cilpr2013/

Pollard, C., & Sag, I. A. (1994). *Head-Driven Phrase Structure Grammar*, CSLI, Stanford. Chicago, Chicago University Press.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In A. Gelbukh (Ed.) *Computational Linguistics and Intelligent Text Processing. Proceedings of the Third International Conference CICLing 2002* (pp. 1-15). https://doi.org/10.1007/3-540-45715-1_1

Sagot B., & Fišer D. (2008). Building a free French wordnet from multilingual resources. In *Proceedings of Ontolex 2008* (pp. 14–19). Marrakech, Maroc https://www.researchgate.net/publication/242710164_Proceedings_of_OntoLex_2008

Savary, A., Ramisch, C., Cordeiro, S. R., Sangati, F., Vincze, V., Qasemi Zadeh, B., Candito, M., Cap, F., Giouli, V., Stoyanova, I., & Doucet, A. (2017). The PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expression – EACL 2017* (pp. 31-47). https://www.aclweb.org/anthology/W17-1704.pdf

Savary, A., Candito, M., Barbu Mititelu, V., Bejček, E., Cap, F., Čéplö, S., Cordeiro, S. R., Eryiğit, G., Giouli, V., van Gompel, M., HaCohen-Kerner, Y., Kovalevskaitė, J., Krek, S., Liebeskind, C., Monti, J., Parra Escartín, C., van der Plas, L., QasemiZadeh, B., Ramisch, C., Sangati, F., Stoyanova, I., & Vincze, V. (2018). PARSEME multilingual corpus of verbal multiword expressions. In S. Markantonatou, C. Ramisch, A. Savary & V. Vincze, (Eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop* (pp. 87–147). Berlin: Language Science Press.

Shwartz, V., & Dagan, I. (2019). Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, *7*, 403-419.

Sinclair, John (1991). *Corpus, Concordance and Collocation*. Oxford: Oxford University Press.

Straka, M. (2018). UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In *Proceedings of CoNLL 2018: The SIGNLL Conference on Computational Natural Language Learning* (pp. 197-207), Association for Computational Linguistics, Stroudsburg, PA, USA, https://doi.org/10.18653/v1/K18-2020

Tanguy, L., & Hathout, N. (2002). Webaffix : un outil d'acquisition morphologique dérivationnelle à partir du Web. In *Actes de la 9e Conférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN'2002)*, Nancy. ATALA. http://talnarchives.atala.org/TALN/TALN-2002/index.html

Taylor, A., Marcus, M., & Santorini, B. (2003). The Penn Treebank: an overview. In A. Abeillé (Ed.), *Treebanks* (pp. 5-22). Dordrecht, Springer.