Received: April 2020 / Accepted: June 2020
Published on line: December 2020

Extraire et encoder l'information lexicale de Wiktionary : quel boulot pour étrangler le goulot !

Franck Sajous
CLLE, CNRS & Université de Toulouse 2
franck.sajous_at_univ-tlse2.fr

Basilio Calderone
CLLE, CNRS & Université de Toulouse 2
basilio.calderone_at_univ-tlse2.fr

Nabil Hathout
CLLE, CNRS & Université de Toulouse 2
nabil.hathout_at_univ-tlse2.fr

Abstract

We present in this article an effort carried out for a decade which consists in using the content of the Wiktionary collaborative dictionary in order to build free lexical resources. Its main result is the design of machine-readable dictionaries and inflectional lexicons for three languages (French, Italian and English). In this paper, we question the usefulness of such lexical resources at a time when mainstream NLP is based on machine learning and readily do without. We compare different methods of producing resources and more specifically of extracting information from Wiktionary. We then discuss the suitability of standard formats for encoding idiosyncratic resources such as Wiktionary and conclude on the need to prioritize, above all, the production and sharing of resources.

Keywords: free lexical resources, machine-readable dictionaries, Wiktionary, information extraction, encoding formats

Résumé

Nous présentons dans cet article une démarche menée depuis une décennie qui consiste à exploiter le contenu du dictionnaire collaboratif Wiktionary afin de construire des ressources lexicales libres. Notre approche a permis de doter trois langues (le français, l'italien et l'anglais) en dictionnaires électroniques et en lexiques flexionnels. Nous questionnons l'utilité des ressources lexicales à un moment où la plupart des systèmes de TAL par apprentissage automatique s'en passent complètement. Nous profitons également de ce retour d'expérience pour comparer différentes méthodes de production de ressources et plus spécifiquement différentes méthodes d'extraction d'informations à partir de Wiktionary. Nous discutons ensuite de la pertinence des formats standards pour l'encodage de ressources idiosyncratiques telles que Wiktionary. Nous concluons sur la nécessité de prioriser, avant tout, la production et le partage de ressources.

Lexique, 27 (2020), 121-144.
ISSN : 0756-7138.

Mots-clefs : ressources lexicales libres, dictionnaires électroniques, Wiktionary, extraction d'informations, formats d'encodage

1. Introduction

« We Desperately Need Linguistic Resources! » s'exclamait Sekine (2010), déplorant le manque de ressources linguistiques indispensables aux tâches de traitement automatique des langues (TAL) dont il dressait la liste. Ce cri du cœur – émanant d'un chercheur qui travaille pourtant sur la langue anglaise, supposée mieux pourvue – était à l'époque de nature à entrer en résonance avec le sentiment de tout TAListe travaillant sur le français. On peut se demander si la situation s'est améliorée en une décennie : pour les langues ne disposant pas, notamment, d'un WordNet de qualité, les réseaux sémantiques construits automatiquement tels que WOLF (Sagot & Fišer, 2008) ou Babelnet (Navigli & Ponzetto, 2010), répondent-ils aux besoins énumérés par Sekine ? Pour De Smedt et al. (2014), les ressources les plus élémentaires, particulièrement celles qui sont libres, sont encore inexistantes pour la plupart des langues, dont l'italien. Selon Sajous, Hathout et Calderone (2014), le français dispose de lexiques flexionnels, mais il n'existe pas de ressource libre pour les autres types d'informations lexicales, comme les transcriptions phonémiques ou les définitions. Pour Pierrel (2013), WOLF et Babelnet privilégient la couverture à la fiabilité des données et ont tendance à être incomplètes, contenir des erreurs et des inexactitudes. Plaidant pour la mutualisation et le partage de ressources, Pierrel souligne à quel point le *Trésor de la Langue Française informatisé* (TLFi) serait une ressource adaptée à des utilisations en TAL. Partageant l'ensemble de ces considérations, nous notons toutefois que le TLFi ne décrit pas le français contemporain et ne peut tenir lieu de réseau sémantique. Surtout, le fait que le TLFi ne soit consultable qu'en ligne par un utilisateur humain rend cette ressource non pertinente pour le TAL. Des conventions passées avec certaines unités de recherche permettent à celles-ci d'obtenir la ressource en intégralité et, sous réserve d'autorisation au cas par cas, d'en distribuer des ressources dérivées. Le cas du *Dictionnaire Électronique des Synonymes* (DES) est similaire : il est accessible en consultation, mais non téléchargeable. Une discussion tenue à propos d'une tâche de substitution lexicale organisée dans le cadre de l'atelier SemDis sur la sémantique distributionnelle (Fabre et al., 2014) interroge sur l'utilité des ressources non libres : certaines équipes participant à la tâche compétitive utilisaient le DES, sans que personne ne sache exactement décrire sa version particulière, ni en préciser le statut juridique, ni même la date et les conditions d'obtention, tandis que d'autres n'y avaient pas accès. Difficile dans ces conditions de comparer les résultats obtenus par des algorithmes intégrant ou non cette source de données exogène, inégalement disponible pour l'ensemble de la communauté, existant dans des versions multiples dont la genèse n'est plus traçable. La réponse à la question que nous posions plus haut est donc négative : 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.

Dans ce contexte, plusieurs méthodes novatrices ont été explorées pour constituer des données lexicales, comme le micro-travail en ligne (*microwork* ou *microtasking*) ou les jeux à objectifs (*games with a purpose*, ou GWAP). Plus ancien, le champ de l'analyse distributionnelle (AD), potentiel pourvoyeur de ressources sémantiques, a été refaçonné par le retour des réseaux de neurones et la conception de nouvelles architectures. Pour notre part, nous avons depuis une décennie adopté une démarche d'exploitation de données issues du « modèle wiki », et plus précisément des différentes éditions de langue de Wiktionary[1], pour produire des dictionnaires électroniques (DE) et des lexiques flexionnels libres. Mais le fameux « goulot d'étranglement » lié à l'acquisition de connaissances lexicales constitue-t-il toujours un frein au traitement de la langue, ou les ressources lexicales (RL) sont-elles rendues caduques par l'équation « TAL = gros corpus + apprentissage automatique » ?

Dans cet article, nous présentons un retour d'expérience sur notre approche, initiée en 2009, qui consiste à extraire et normaliser l'entièreté du contenu de Wiktionary, le rendant ainsi utilisable pour le TAL, la linguistique et la lexicographie. Nous avons en effet développé, pour trois langues, des dictionnaires électroniques et des lexiques flexionnels fondés sur le projet collaboratif :

- GLAWI (Sajous & Hathout, 2015) et GLÀFF (Sajous, Hathout & Calderone, 2013) extraits du Wiktionnaire, l'édition française de Wiktionary ;
- GLAWIT (Calderone et al., 2016) et GLAFF-IT (Calderone, Pascoli, Sajous & Hathout, 2017) extraits du Wikizionario, l'édition italienne de Wiktionary ;
- ENGLAWI et ENGLAFF (Sajous, Calderone & Hathout, 2020), extraits du Wiktionary anglais.

Le présent article n'a pas vocation à présenter ces ressources (pour une description détaillée, nous renvoyons le lecteur aux articles cités ainsi qu'à leurs documentations respectives[2]), mais à exposer notre démarche en insistant sur certains aspects particuliers. Après nous être interrogés sur l'utilité des RL en général et des DE en particulier, nous revenons sur la nature des données issues de Wiktionary, sur l'impact de différentes méthodes d'extraction d'informations, sur la qualité des données produites et sur la pertinence des standards pour l'encodage des ressources lexicales et lexicographiques, souvent atypiques.

2. Les ressources lexicales sont-elles (encore) utiles ?

Dans un MOOC se déroulant en septembre et octobre 2019, Creese et al. (2018) ont exploré « la place des dictionnaires dans le monde actuel » et posé la question : « cela aurait-il une importance s'il n'y

[1] Wiktionary désigne à la fois l'ensemble du projet et l'édition de langue anglaise. Nous parlerons de « Wiktionary » pour le projet et de « Wiktionary anglais » pour son édition de langue anglaise.

[2] http://redac.univ-tlse2.fr/lexiques/

avait plus de dictionnaires ? ». La question vaut pour les utilisateurs humains. On peut l'extrapoler au domaine du TAL et se demander si les DE, et plus généralement les RL, sont (encore) utiles à cette discipline, et dans quelle mesure. Au début des années 1990, Ide et Véronis (1993) revenaient, dans un article intitulé « Extracting Knowledge Bases from Machine-Readable Dictionaries: Have We Wasted Our Time », sur les résultats décevants de travaux visant l'extraction de connaissances à partir de DE. Les auteurs formulaient deux postulats : 1. Les DE contiennent de l'information utile au TAL et 2. Cette information est relativement facile à extraire. Après une revue de la littérature, ils concluaient que « les 15 dernières années de travaux n'ont produit [en 1993] qu'une poignée de taxonomies limitées et imparfaites ». Une explication que les auteurs donnaient alors était que l'information contenue dans les dictionnaires n'est ni totalement fiable, ni parfaitement cohérente. Ils se demandaient également si l'information était aussi facile à extraire que suggéré dans les travaux qu'ils avaient recensés. L'émergence de la linguistique de corpus, le recours au web et la place croissante de l'apprentissage automatique sur des collections de textes toujours plus vastes pourraient reléguer les dictionnaires au rang d'objet du passé. Deux contre-arguments peuvent néanmoins être objectés. D'une part, les travaux recensés par Ide et Véronis utilisaient des DE suffisamment anciens pour être tombés dans le domaine public (en particulier, le *Webster New Collegiate Dictionary*, 7th ed.). Or l'existence d'un dictionnaire sous licence libre, riche, multilingue et mis à jour en continu tel que Wiktionary change la donne. La question de la difficulté d'en extraire de l'information est abordée en Section 4. D'autre part, les RL (notamment extraites des DE) peuvent améliorer les performances des systèmes par apprentissage (Plank & Klerke, 2019).

Deux décennies après l'étude d'Ide et Véronis (1993), Gurevych, Eckle-Kohler et Matuschek (2016) écrivent dans la préface d'un livre sur les bases de connaissances lexicales liées que les connaissances lexico-sémantiques sont vitales pour la plupart des tâches de TAL. Un peu plus loin (p. 65-66), les auteurs écrivent aussi que le bénéfice tiré de l'utilisation des bases de connaissances lexicales n'est souvent pas clairement visible, laissant le lecteur dubitatif. Des travaux montrent pourtant que les RL sont nécessaires à certaines tâches, ou au moins qu'elles améliorent les résultats des systèmes par apprentissage automatique. Pour la conception de phonétiseurs, Rojc et Kačič (2007) montrent que des lexiques phonologiques améliorent les modèles appris automatiquement. Un lexique de variantes diatopiques peut améliorer un système d'attribution d'auteurs (Tanguy et al., 2011). En traduction automatique, des corpus parallèles ou comparables n'étant pas disponibles pour toutes les paires de langues, Klementiev, Irvine, Callison-Burch et Yarowsky (2012) recourent à des corpus monolingues et des dictionnaires bilingues de tailles réduites pour créer des lexiques bilingues. Pour la détection de langues proches, la discrimination passe généralement par le lexique (Tiedemann & Ljubešić, 2012).

Ainsi, selon plusieurs auteurs, les RL sont utiles à un certain nombre de tâches. Mais, compte tenu de l'évolution rapide des méthodes par apprentissage, ces travaux peuvent paraître datés, notamment au regard de l'AD fondée sur des modèles neuronaux, sur laquelle se focalise actuellement l'essentiel de l'attention dans le domaine du TAL. Pour Lenci (2018), qui s'interroge sur le pouvoir descriptif et

explicatif de l'AD comme modèle de sens et comme méthode pour l'analyse sémantique, beaucoup de questions restent ouvertes. En effet, si les modèles neuronaux affichent d'excellents résultats sur les jeux de test standards, Lenci rappelle que ces étalons – SimLex-999 excepté – contiennent des jugements sémantiques de proximité (*car/driver*), plutôt que de similarité (*car/van*). Les performances de tous les modèles diminuent considérablement lorsqu'ils sont évalués sur SimLex-999. Une autre lacune relevée par Lenci dans les évaluations habituelles est que celles-ci se focalisent généralement sur les noms. Lorsqu'ils sont évalués sur SimVerb-3500 (qui contient des jugements de similarité entre verbes), les modèles – qu'ils soient prédictifs ou fréquentiels – donnent des résultats très faiblement corrélés avec les jugements humains et obtiennent même des résultats inférieurs à ceux de méthodes non distributionnelles. Malgré quelques tentatives récentes, une autre faiblesse de l'AD est son incapacité à identifier la nature des relations lexicales capturées par les mesures de proximité. Cela montre, selon Lenci, que les modèles distributionnels ne fournissent du sens lexical qu'un modèle de représentation à gros grain, « un proxy superficiel ». Ajoutons enfin que, même si des travaux récents commencent à s'intéresser à la polysémie, ce phénomène reste largement ignoré en AD.

Il ne saurait être question de réfuter les potentiels de l'AD pour la modélisation des connaissances linguistiques sur la base des motifs évoqués ci-dessus, d'autant que les efforts actuellement consentis apporteront probablement de nouvelles avancées. En l'état actuel, l'intégration des RL aux architectures neuronales constitue une piste intéressante : Plank et Klerke (2019) estiment, comme le montrent selon elles de plus en plus de travaux, que, contrairement à l'opinion largement répandue, l'intégration aux modèles neuronaux de connaissances lexicales symboliques améliore les processus d'apprentissage. Les autrices mettent en évidence, pour 21 langues, l'impact positif de la combinaison de RL traditionnelles avec des réseaux neuronaux conçus pour l'étiquetage morphosyntaxique.

3. La piste du crowdsourcing et le « modèle wiki »

3.1. L'indigence, un terrain favorable aux approches innovantes ?

Les conclusions d'Ide et Véronis (1993) sur l'extraction de connaissances à partir de DE coïncidaient à la fois avec le développement par Hearst (1992) d'une méthode d'extraction de relations lexicales par projection de patrons lexico-syntaxiques sur corpus et avec la sortie de la version 1.0 de WordNet (Miller, 1995 ; Fellbaum, 1998). La méthode semi-automatique de Hearst (qui en inspira d'autres) est intéressante dès lors que l'on accepte la nécessité d'une intervention humaine coûteuse (le micro-travail n'était pas encore d'actualité) et que des corpus appropriés sont disponibles. Concernant WordNet, cette ressource a essuyé de nombreuses critiques, mais son utilisation systématique par la communauté TAL en fait un succès dans ce domaine. Les projets de création de ressources similaires dans d'autres langues ont quant à eux connu des succès divers, selon les efforts humains et financiers consentis, ainsi que les ressources et outils préexistants pour la langue visée. Pour le français, la version

d'EuroWordNet (Vossen, 1998) est un échec. Ses nombreuses lacunes sont relevées notamment par Jacquin, Desmontils et Monceaux (2007). Construite par deux acteurs privés et une unique université, elle est d'emblée limitée à un faible nombre de synsets (7500), restreinte aux noms et aux verbes, et reste, depuis la fin du projet, figée, payante et propriétaire (elle figure encore au catalogue d'ELRA). Ajoutons que pour être consultée, EuroWordNet nécessite un outil d'interrogation de bases de données également payant et propriétaire. Ainsi, une institution qui aurait acheté la ressource à l'époque ne serait plus en capacité de l'utiliser, sauf à avoir maintenu une machine et un système d'exploitation « vintage ». Cet échec et l'absence de WordNets libres pour certaines langues ont motivé d'autres initiatives, notamment de construction de réseaux à partir d'agrégation et traduction automatique de ressources libres. C'est le cas de WOLF et BabelNet. Ces approches entièrement automatiques pourraient être intéressantes si elles étaient présentées comme des études de faisabilité et non comme des ressources finalisées utilisables. La qualité – à supposer que l'on sache l'estimer, de près ou de loin – de BabelNet dépend notamment des ressources utilisées. Elle dépend également de la finesse des méthodes d'extraction d'informations appliquées à ces ressources. Or, comme nous le montrons à la Section 4, les initiatives de construction de ressources « massivement multilingues » laissent souvent de côté des ressources pertinentes ou mettent en œuvre des extractions trop frustes. Eckard, Barque, Nasr et Sagot (2012) mettent en évidence des manques et des incohérences dans WOLF. Nous avons pour notre part relevé qu'environ 90% des synsets de WOLF ne contiennent pas de définition en français. Dans une situation où l'inventaire des RL libres disponibles reste indigent, ce type d'initiative retient néanmoins notre attention : couplé à un processus de validation et à la prise en compte d'autres ressources (ou des méthodes d'extraction plus fines), il pourrait donner des résultats intéressants. Bien sûr, formulée de la sorte, l'idée de mener un tel travail à large échelle a tout d'un vœu pieux. Mais le recours au *crowdsourcing* pourrait être une piste.

3.2. Crowdsourcing, jeux à objectifs et micro-travail

Le terme *crowdsourcing* popularisé par Howe (2006) désignait l'externalisation (*outsourcing*) par des entreprises de tâches à (faire) accomplir par « les foules » (*crowds*), *i.e.* des communautés d'internautes. Il désigne aujourd'hui une réalité mouvante et polymorphe que nous nous abstiendrons de définir : nous renvoyons à Estellés-Arolas et González-Ladrón-de Guevara (2012) pour une typologie, à Brabham (2013) pour une formulation des « ingrédients clés » et une description de différents types de mise en œuvre, ainsi qu'à Estellés-Arolas, Navarro-Giner et González-Ladrón-de Guevara (2015) pour un recensement de différentes définitions (souvent divergentes) du concept. Nous commentons ci-dessous deux familles d'approches qui ont été explorées avec des succès divers pour la production de connaissances lexicales.

La première, celle des jeux à objectifs, consiste à faire accomplir des tâches à des internautes, motivés par l'aspect ludique du système proposé. Une des difficultés consiste pour une équipe de

recherche à trouver un bon compromis entre un coût de développement limité et une attractivité susceptible d'attirer des foules de joueurs. En la matière, *JeuxDeMots* (Lafourcade, 2007) est un exemple de jeu à la conception relativement simple, qui a attiré un nombre conséquent de participants. Diverses récompenses sous forme de bons d'achat ont en revanche dû être ajoutées à la version initiale de *Phrase Detective* (Chamberlain, Kruschwitz & Poesio, 2009), un jeu conçu dans le but d'annoter des anaphores (le jeu devenant ainsi un système hybride à mi-chemin entre *GWAP* et micro-travail). L'idée de s'amuser en résolvant des anaphores n'était effectivement peut-être pas immédiate pour la plupart des internautes. Selon Jurgens et Navigli (2014), la plupart des *GWAP* consistent en une interface textuelle qui s'apparente trop à une tâche d'annotation traditionnelle. Les auteurs proposent le développement de jeux vidéos avec un graphisme proche de ce que connaissent les joueurs en ligne. Avec *Puzzle Racer*, ils obtiennent de meilleurs résultats qu'avec une plateforme de micro-travail, et à moindre coût (75% plus bas). Mais, d'une part, la participation est assurée par le recrutement d'étudiants rétribués par des bons d'achats. Difficile donc de juger réellement de l'attractivité du jeu. D'autre part, le coût – financier – de développement du jeu est nul : ce sont également des étudiants qui l'ont développé dans le cadre d'un cours de Java… La question éthique du travail gratuit accompli par ces étudiants n'est pas abordée dans l'article. À titre de comparaison, Poesio, Chamberlain, Kruschwitz, Robaldo et Luca (2015) révèlent que le développement de *Phrase Detective* a coûté £60 000 de salaires (en plus des £18 000 de bons donnés aux participants).

La seconde famille d'approches est celle du micro-travail, système distribué visant à faire accomplir des micro-tâches (*microtasks*) à des travailleurs contre de faibles rémunérations. Elle a été une piste largement exploitée en TAL et plus récemment en lexicographie. Par exemple, Snow, O'Connor, Jurafsky et Ng (2008) ont obtenu de meilleurs résultats dans plusieurs tâches de nature sémantique avec un système entraîné sur les annotations de plusieurs annotateurs naïfs qu'avec un système entraîné sur celles d'un expert unique. De nombreux projets d'annotation ou de création de jeux de données standard ont été développés depuis pour le TAL. Le micro-travail pose principalement deux problèmes : l'éthique et l'évaluation de la fiabilité des données. L'aspect éthique amène à s'interroger sur l'exploitation potentielle des travailleurs en ligne. Rumshisky, Botchan, Kushkuley et Pustejovsky (2012), par exemple, constatent froidement que restreindre la participation aux États-Unis (*i.e.* bannir les contributeurs indiens) permet d'augmenter la qualité des annotations, mais nécessite d'augmenter le montant des rétributions, faute de quoi les internautes boudent les micro-tâches proposées. Bederson et Quinn (2011) s'interrogent sur les conditions à partir desquelles le micro-travail serait éthiquement acceptable et proposent un ensemble de bonnes pratiques visant à satisfaire à la fois les travailleurs et leurs commanditaires. La question de l'évaluation de la qualité des données reste problématique : la métrique employée pour évaluer les données produites (également utilisée pour éliminer les travailleurs non fiables sans les rétribuer) reste l'accord inter-juges. Or cet accord est corrélé, selon Murray et Green (2004), à un niveau de compétence homogène parmi les annotateurs, et non à un niveau de compétence élevé. Ajouter les annotations produites par un expert (supposées de bonne qualité) à celles

produites par un groupe de naïfs fait chuter l'accord (bien qu'augmentant *a priori* la qualité globale des annotations). Les questions que pose l'utilisation du micro-travail en TAL restent donc ouvertes. Cette approche constitue néanmoins une piste intéressante, notamment en lexicographie, pour des tâches difficiles à automatiser telle que l'alignement des occurrences en corpus d'unités lexicales avec un inventaire de sens existant (Rumshisky, 2011) ou l'induction de sens à partir de corpus (Rumshisky et al., 2012). Čibej, Fišer et Kosem (2015) envisagent la possibilité d'intégrer, de manière modulaire, le micro-travail dans l'ensemble des tâches nécessaires à la conception d'un dictionnaire. Selon Rundell (2017), il serait insensé d'ignorer le potentiel du micro-travail pour la lexicographie étant donné les expériences positives menées dans ce domaine. Il reste à trouver un modèle de processus qui optimise à la fois la partition du travail entre systèmes informatiques, naïfs et lexicographes, et la satisfaction des travailleurs, des maisons d'édition et des utilisateurs de dictionnaires.

3.3. Le modèle wiki en lexicographie

Le « modèle wiki », sur lequel reposent Wikipédia et Wiktionary, fait partie de ce que certains auteurs appellent les « commons-based peer production ». Brabham (2013), comme d'autres, considère que les wikis, en l'absence d'une instance hiérarchisée de management descendant, ne constituent pas une sous-catégorie du *crowdsourcing*. Concernant Wikipédia et Wiktionary, ce dernier point est complexe et discutable – Lew (2014) évoque au contraire une organisation à structure pyramidale – mais la question qui nous occupe en premier lieu est de savoir si une approche particulière de production de données, parmi celles faisant intervenir les foules, est pertinente pour la conception de RL. De l'article « Wisdom of Crowds versus Wisdom of Linguists » de Zesch et Gurevych (2010), on ne cite généralement que la conclusion : les résultats obtenus avec des ressources fondées sur la « sagesse des foules » ne sont pas supérieures, mais comparables à ceux obtenus avec celles fondées sur la « sagesse des linguistes », les premières complétant les secondes et les surpassant en termes de couverture. On oublie souvent la deuxième partie du titre « Measuring the Semantic Relatedness of Words » : les comparaisons sont établies à travers le prisme de leur pertinence pour le calcul de proximité sémantique. Mais quelle est la qualité intrinsèque de ces ressources envisagées comme source primaire de connaissance ? Quand on se sert de Wikipédia pour développer des algorithmes de calcul de proximité sémantique, l'exactitude encyclopédique des articles importe peu. Si en revanche on utilise Wiktionnaire comme source de description du lexique, pour en extraire des transcriptions phonémiques, des relations lexicales, des définitions, un inventaire de sens, des vocabulaires de mots marqués (mots techniques, datés, emprunts, etc.), la question se pose en d'autres termes. Que vaut cette ressource en tant que dictionnaire, pour un utilisateur humain d'une part, et comme DE potentiel d'autre part ? Dans sa thèse, Meyer (2013) donne une description des éditions anglaise et allemande de Wiktionary en adoptant « une perspective à la fois TAL et métalexicographique », mais son approche est entièrement quantitative. Par ailleurs, Meyer compare Wiktionary à des ressources disponibles au format électronique telles que, par exemple, WordNet, mais pas à de « vrais

dictionnaires », qu'ils soient au format papier ou électronique. Les descriptions et évaluations qualitatives de Wiktionary ne s'appuient souvent que sur une quantité infime d'exemples, rendant toute généralisation hasardeuse. Lew (2014), par exemple, fonde sa comparaison entre Wiktionary et le *Longman Dictionary of Contemporary English* sur l'étude d'une seule entrée prise au hasard, le verbe *handle*. L'analyse de lexicographes chevronnés, comme celle de Rundell (2017), est souvent moins critique que celle de contempteurs moins experts. Rundell évoque par exemple un niveau de contrôle éditorial tel qu'une contribution non pertinente sera rapidement supprimée. Il constate en revanche une manière obsolète de décrire le sens des mots (par ex. par des définitions morphologiques), ou l'absence de consultation de corpus par les internautes, « comme si les 30 dernières années de développement de la lexicographie n'avaient pas existé », ou « comme si Wiktionary perpétuait ou faisait revivre d'anciennes pratiques lexicographiques ». Ces études sont menées sur l'édition anglaise de Wiktionary. Généraliser les analyses au français serait méconnaître les différences qui existent entre les différentes versions de langue du projet collaboratif d'une part, et (surtout) entre la lexicographie française et anglo-saxonne (qu'il s'agisse du processus éditorial ou du produit fini qu'est le dictionnaire). La première a connu des évolutions majeures, auxquelles Rundell fait référence, alors que la seconde peine à innover et semble « en panne » (Corbin, 1998), à tel point qu'on peut, comme Corbin (2008), s'interroger sur son avenir. Cette différence est également perceptible lorsqu'on compare l'offre dictionnairique, notamment les dictionnaires contemporains disponibles gratuitement en ligne (Corbin & Gasiglia, 2020). Une critique du Wiktionnaire, positive ou négative, devrait toujours se faire avec une bonne connaissance de ce que sont réellement les dictionnaires du français[3], en oubliant un temps le prestige inconditionnel que l'imaginaire collectif leur prête souvent (imaginaire bien sûr non amendé par les maisons d'édition et entretenu par une presse peu exigeante). Des études qui comparent le Wiktionnaire à des dictionnaires professionnels, adoptant l'angle de la néologie (Sajous, Josselin-Leray et Hathout, 2018), du vocabulaire (plus ou moins) spécialisé (Sajous, Josselin-Leray & Hathout, 2020), ou de la neutralité de point de vue (Sajous, Hathout & Josselin-Leray, 2019), montrent d'une part que le Wiktionnaire est un objet complexe, qui hérite notamment d'anciennes pratiques lexicographiques dues à des imports automatiques (faisant écho à l'observation de Rundell), et d'autre part que des dictionnaires issus du secteur privé, tel que le *Petit Robert*, présentent de nombreuses lacunes et incohérences. Concernant le Wiktionnaire, les différentes études s'accordent en général sur sa nomenclature hors norme et sa propension à décrire les néologismes et les termes spécialisés. Si l'on ajoute que, pour le français hexagonal contemporain[4], c'est la seule ressource sous licence libre à

[3] Par exemple, les définitions morphologiques qui, pour Rundell, appartiennent au passé, sont encore largement répandues dans les dictionnaires français.

[4] Le dictionnaire en ligne *Usito*, qui décrit le lexique québécois, est gratuitement accessible depuis octobre 2019, mais n'est pas téléchargeable ni interrogeable automatiquement.

fournir des définitions, on comprendra qu'on l'envisage comme une source de données potentielle pour la conception de RL.

Avant de présenter nos méthodes d'extraction et nos choix d'encodage, nous présentons dans la section qui suit différents travaux qui ont exploité le contenu de Wiktionary.

3.4. Exploitation de Wiktionary en TAL : travaux antérieurs

Depuis la fin des années 2000, Wiktionary est exploité dans des buts variés. Avant d'être considéré comme un matériau pour la construction de RL, il a servi à la mise au point d'algorithmes de calcul de proximité sémantique (pouvant le cas échéant être utilisés pour la conception de RL) : Zesch, Müller et Gurevych (2008b) l'ont utilisé pour évaluer des mesures de proximité sémantique ; Navarro et al. (2009) ont mis à profit sa structure de petit monde hiérarchique pour enrichir son réseau de synonymie par marches aléatoires ; Weale, Brew et Fosler-Lussier (2009), se fixant le même objectif, ont exploité la structure hypertextuelle de ses articles. Wiktionary a ensuite été utilisé pour extraire et enrichir des données lexicales destinées la plupart du temps à une tâche particulière. Par exemple, Schlippe, Ochs et Schult (2010) ont créé pour plusieurs langues des dictionnaires de prononciation à destination d'un système de reconnaissance et de synthèse de la parole et validé la bonne couverture et la bonne qualité des transcriptions extraites. De Smedt, Marfia, Matteucci et Daelemans (2014) ont exploité la nomenclature et les parties du discours de Wiktionary pour développer un étiqueteur morphosyntaxique faiblement supervisé, rapide et libre, qu'ils estiment être d'une qualité raisonnable. À partir d'une approche semi-supervisée exploitant les patrons de flexion du wikicode, Metheniti et Neumann (2018) ont produit des paradigmes flexionnels pour 150 langues. Segonne, Candito et Crabbé (2019), dont le but est de développer une méthode de désambiguïsation en contexte pour des langues ne bénéficiant pas de corpus annotés au niveau sémantique (c'est-à-dire, écrivent les auteurs, les langues autres que l'anglais), se tournent également vers Wiktionary. Se focalisant plus particulièrement sur la désambiguïsation des verbes français, ils exploitent l'inventaire des sens du Wiktionnaire ainsi que les exemples correspondants comme corpus d'entraînement. Leur démarche est encourageante, les auteurs concluant notamment à une granularité de l'inventaire de sens du Wiktionnaire appropriée à une annotation sémantique de bonne qualité. Si l'on admet que les RL sont encore utiles (cf. Section 2), il semble que Wiktionary puisse servir à la conception de beaucoup d'entre elles. Encore faut-il être en capacité d'extraire correctement les données qu'il contient. C'est la question abordée en Section 4.

4. Extraction d'informations

Dans leur article évoqué en Section 2, Ide et Véronis (1993) reviennent sur le postulat qu'ils ont eux-mêmes formulé : « on peut extraire facilement des connaissances lexicales à partir des DE ». Pour des raisons différentes que nous exposons plus bas, l'extraction d'informations à partir de Wiktionary est loin d'être aussi simple qu'il est souvent implicitement suggéré dans la littérature. La syntaxe du

wikicode (le format d'encodage des wikis) n'est pas définie formellement, évolue dans le temps et utilise abondamment des *templates* (« patrons » ou « modèles » paramétrables) qu'il faut « émuler », *i.e.* analyser et réimplémenter, afin d'extraire de manière fiable et exhaustive les informations ainsi encodées (Liebeck & Conrad, 2015). Hormis certains éléments comme les hyperliens, l'italique et la graisse, la syntaxe du wikicode varie fortement d'une édition de langue à l'autre. De nombreux auteurs qui recourent à Wiktionary n'en extraient qu'un type d'information spécifique. Parce que des données partielles et bruitées sont suffisantes pour les tâches qu'ils envisagent, ils ne font pas état de la nature problématique du wikicode[5]. La plupart d'entre eux utilisent le *dump* de l'édition anglaise pour extraire des données dans une langue cible en ignorant purement l'édition correspondant à cette langue. D'autres conçoivent des extracteurs à très gros grains pour les appliquer aux *dumps* de différentes éditions de langue. Notre approche consiste au contraire à développer des analyseurs finement ajustés à chaque langue traitée afin d'extraire aussi fidèlement que possible l'intégralité des informations contenues dans chaque édition, sans *a priori* sur l'utilisation ultérieure qui sera faite des données ainsi produites. Nous analysons ci-après les avantages et les limites des différentes méthodes d'extraction d'informations à partir de Wiktionary.

Zesch, Müller et Gurevych (2008a) ont développé et mis à disposition JWKTL, une API permettant d'extraire de l'information à partir des *dumps* des éditions allemande et anglaise de Wiktionary. Navarro et al. (2009) ont mis à disposition WiktionaryX, contenant les versions française et anglaise de Wiktionary sous forme de DE. Un avantage de WiktionaryX est d'être prêt à l'emploi et de ne nécessiter que peu de compétences en programmation (il suffit de savoir manipuler un document XML volumineux). Un inconvénient est son obsolescence : son contenu, une fois extrait, est une image passée, désynchronisée, de Wiktionary. À l'inverse, JWKTL permet de traiter des *dumps* récemment téléchargés. L'inconvénient majeur est que cette API ne traite que le balisage wiki le plus simple, comme celui encodant les hyperliens et le style typographique, mais ne prend pas en charge la plupart des *templates* utilisés par le moteur MediaWiki. Par exemple, les marques lexicographiques des définitions sont ignorées[6] et les *templates* imbriqués ne sont pas supportés[7]. À titre d'exemple, le wikicode d'un des sens de l'adjectif *sweet* est :

```
{{lb|en|informal|followed by {{m|en|on}}}} [[romantic|Romantically]]
[[fixate|fixated]], [[enamor|enamoured with]], [[fond|fond of]]
```

JWKTL produit le résultat suivant :

[5] Le fait que l'extraction d'informations à partir du wikicode ne soit ni une activité intellectuelle épanouissante, ni « bankable » en termes de h-index, y est peut-être également pour quelque chose.

[6] La méthode `getMarker()` de la dernière version (1.1.0) de JWKTL appliquée au *dump* anglais du 1/11/2019 retourne toujours une chaîne vide.

[7] Le texte correspondant est supprimé, remplacé par des accolades fermantes excédentaires.

}} Romantically fixated, enamoured with, fond of

au lieu de :

(informal, followed by on) Romantically fixated, enamoured with, fond of

Pour construire le réseau multilingue DBnary, Sérasset (2012) se concentre sur l'extraction des entrées « facilement extractibles » de Wiktionary. Le graphe résultant pour le français, comptant 260 467 nœuds, est loin d'être exhaustif. Sérasset écrit d'ailleurs ultérieurement que le but de sa démarche n'est pas de produire un résultat qui reflète le contenu de Wiktionary de manière extensive (Sérasset, 2015). L'inconvénient principal des extracteurs qui ne traitent pas le wikicode dans toute sa complexité ne réside cependant pas dans la quantité moindre de données extraites mais plutôt dans le fait qu'ils génèrent des données erronées. Les définitions (avec les étymologies) sont les éléments de Wiktionary faisant intervenir la plus grande diversité de *templates*. Ceux-ci encodent des marques lexicographiques, des styles typographiques, mais également de nombreux éléments micro-structurels des gloses et des exemples. Une extraction correcte de l'information requiert donc un traitement approprié de l'ensemble de ceux-ci. Dans DBnary, 9% des définitions (68 524 sur 760 184) sont vides (elles contiennent seulement une marque lexicographique, un point ou une accolade) et, surtout, d'autres diffèrent de celles de Wiktionary. Par exemple, la définition de *children* consiste en un seul point dans DBnary alors qu'une gestion correcte des templates permet à ENGLAWI d'avoir une définition conforme à celle de Wiktionary : *plural of child*[8]. Certains sens des définitions présentes dans DBnary sont manquants, comme le cinquième sens du nom anglais *pseudo*, du fait de la non prise en compte par l'extracteur utilisé du *template* {{clipping of|en|pseudoephedrine}}, qui produit dans Wiktionary *clipping of pseudoephedrine*[9]. Outre la non prise en compte de nombreux *templates*, DBnary ne traite pas non plus les définitions imbriquées, ce qui entraîne la perte de plusieurs sens pour certaines entrées. Par exemple, DBnary ne donne « que » 18 sens (sur 43) du verbe *strike* : le sens de premier niveau *to have a sharp or severe effect* est bien présent, mais le sens *to steal money* est absent de la ressource (les 43 sens sont correctement extraits dans ENGLAWI). Notons pour finir que, si dans les fichiers morphologiques de DBnary mis à disposition au format *turtle*, les paradigmes distinguent bien lemmes et flexions, ce n'est pas le cas dans les fichiers « core data » : *cats* et *cat*, ou *children* et *child*, y apparaissent comme des entrées canoniques, non reliées entre elles.

BabelNet (Navigli & Ponzetto, 2010), constitué par agrégation et traduction automatique de différentes ressources, mélange connaissances encyclopédiques et lexicographiques et donne, pour les mots anglais, des définitions extraites de Wikipédia, Wiktionary et WordNet. Pour le français, les

[8] La version XML de la définition de ENGLAWI est : <inflectionOf><inflectionType>plural</inflectionType><lemma>child</lemma></inflectionOf>.

[9] La définition textuelle de ENGLAWI est conforme à celle de Wiktionary et sa version XML est : <formOf type="clipping">pseudoephedrine</formOf>.

définitions sont extraites la plupart du temps de Wikipédia et jamais du Wiktionnaire. Le mot *alphabète* n'est défini que par un équivalent (*lettré*), alors qu'il possède une définition dans le Wiktionnaire. Le québécisme *divulgâcher*, absent de Wikipédia, est également absent de BabelNet, bien que l'entrée existe et soit définie dans le Wiktionnaire. Un autre effet du choix de ne pas utiliser le Wiktionnaire est la sous-représentation dans BabelNet des mots français autres que les noms (les articles de Wikipédia correspondant majoritairement à cette catégorie). Par ailleurs, les mots autres qu'anglais sont souvent définis par une glose en anglais provenant de la définition, extraite de WordNet, d'un équivalent traductionnel. Par exemple, *consensuel* est défini dans BabelNet par *existing by consent* (définition de *consensual* dans WordNet) alors qu'il pourrait l'être par la définition tirée du Wiktionnaire *issu d'un consensus*. L'adjectif italien *consensuale* a la même définition que le nom *consenso* dans BabelNet, alors qu'une définition existe dans le Wiktionary italien : *che si fa col consenso della o delle altre parti*.

Les approches les plus proches de la nôtre en termes d'extraction d'informations sont celles utilisées pour produire knoWitiary (Nastase & Strapparava, 2015) et IWNLP (Liebeck & Conrad, 2015). Nastase et Strapparava se donnent pour objectif d'obtenir une ressource lexicale cohérente et fidèle à Wiktionary, contenant autant d'information que possible sur les mots et les relations qu'ils entretiennent. Liebeck et Conrad (2015) développent un lemmatiseur pour l'allemand et se concentrent sur l'extraction des informations flexionnelles pour cette langue. Ils insistent sur l'importance, pour cette tâche, d'une réimplémentation correcte des *templates*. Les deux articles se comparent aux résultats obtenus par JWKTL et concluent à des manques de ce dernier. Contrairement à IWNLP, dont le code source et les données sont mis à disposition, knoWitiary n'est pas accessible. Une autre approche intéressante est celle de Kirov, Sylak-Glassman, Que et Yarowsky (2016), qui exploitent la structure des pages HTML de différentes éditions de langue de Wiktionary pour en extraire les paradigmes flexionnels. Cette méthode a pour avantage de nécessiter peu d'adaptation de leurs programmes d'une langue à l'autre. Selon les auteurs, les résultats sont comparables en quantité et en qualité, pour trois langues testées, à ceux obtenus avec des extracteurs existants, spécifiques et finement ajustés à une langue donnée. Les auteurs concluent que la ressource développée, UniMorph, couvrant 350 langues, est une ressource morphologique dont la taille est sans précédent. ENGLAWI n'existait pas à cette époque. Des ressources existaient cependant pour le français, notamment GLÀFF et GLAWI, qui affichent une meilleure couverture, mais n'ont pas été prises en compte dans cette comparaison.

Ainsi, une extraction d'informations finement ajustée a un impact positif, au plan à la fois qualitatif et quantitatif. La question se pose ensuite du choix du format pour encoder les données extraites.

5. Normes et formats : encoder Wiktionary, tout Wiktionary, rien que Wiktionary

Il existe différentes familles de normes et de formats permettant d'encoder de façon plus ou moins standard les RL. Depuis 2009, nous mettons à disposition des ressources (lexiques et DE) libres. Enjoints à plusieurs reprises de justifier notre renoncement à encoder ces ressources dans tel ou tel format standard, il nous est apparu durant cette période que la moindre réserve émise sur la capacité d'une norme à encoder un type particulier d'information est irrecevable alors que le choix, même motivé, d'un format *ad hoc* pour la mise à disposition d'une ressource place d'emblée celle-ci sous les feux de la critique. Nous tentons ici une mise au point.

5.1. Normes et orthodoxie

Les *guidelines* de la *Text Encoding Initiative* (TEI), dans leur version P5, ont paru en 2007. La TEI permet d'encoder divers éléments des corpus textuels tels qu'on les conçoit habituellement, mais également une variété d'objets complexes : dictionnaires, fac-similés, poésies et pièces de théâtre, oral retranscrit, etc. En 2008, le nouveau *Lexical Markup Framework* (LMF), conçu pour encoder différents types de lexiques, devient un standard ISO. Une décennie plus tard, des travaux sont menés pour raffiner ou simplifier ces standards. Bański, Bowers et Erjavec (2017) reviennent sur la TEI, dont les *guidelines* qui visaient à être en mesure de représenter toute ressource existante, fournissant pour cela de multiples solutions d'encodage, ont été critiquées pour leur trop grande complexité. Pour « garantir l'interopérabilité », une stratégie est de fournir, dans le cadre de la TEI-Lex0, un format « qui ne pourra peut-être pas gérer toutes les variations potentielles, mais qui traitera à la place 90% des phénomènes, 90% du temps ». En bref : pour garantir l'interopérabilité, réduisons la quantité de données à rendre interopérables. En parallèle de la TEI-Lex0, Romary et al. (2019)[10] ont initié une « révision en profondeur » de LMF, la « norme *de jure* qui constitue un cadre pour la modélisation et le codage des informations lexicales ». Selon les auteurs, l'objectif de cette version actualisée, qu'ils baptisent « LMF reloaded » dans leur article, est de proposer une suite plus modulaire, flexible et durable de la norme LMF originale, jugée trop riche et trop complexe. Or les principaux problèmes de LMF ne sont pas, selon nous, sa richesse ni sa complexité, mais le discours prosélyte qui l'accompagne et le statut de ses spécifications. D'une part, LMF est un méta-modèle dont chacun peut donner une instanciation spécifique. Même « reloaded », il ne peut donc *garantir* l'interopérabilité. Celle-ci est rendue possible seulement si une même instance de modèle est largement adoptée par une communauté pour l'encodage d'un certain nombre de ressources. D'autre part, contrairement à la TEI dont les *guidelines* détaillées sont librement accessibles dans un document d'environ 2000 pages, on ne peut se

[10] Article cosigné par deux des trois auteurs de celui de Bański et al. (2017).

procurer les spécifications ISO de LMF qu'en déboursant 178 CHF[11]. Il faudra débourser 88 CHF supplémentaires pour obtenir les spécifications sur le registre de catégories de données (*data category registry*) utilisé pour générer les couples attribut-valeur de LMF[12]. On pourra enfin regretter que l'ouvrage édité par Francopoulo (2013), qui présente LMF et en donne des cas d'utilisation, ne soit pas diffusé librement.

En plus des conventions d'encodage, les réflexions menées autour de la TEI, encadrées par le TEI Council, ont permis la mise en commun d'outils et de pratiques ainsi que la structuration d'une communauté à la fois TAL et linguistique. Les travaux actuels de refaçonnage de la TEI et de LMF sont justifiés par leurs auteurs par la nécessaire interopérabilité des ressources, ou, plus récemment, par le cadre des « données liées ». Pour les langues qui manquent de ressources (par ex. de DE), comme le français et l'italien, l'interopérabilité n'est cependant pas un problème : rappelons le truisme que la question de l'interopérabilité ne se pose que lorsque plusieurs ressources existent.

5.2. Des formats, pour quoi faire ?

La question nous a été plusieurs fois posée de savoir pourquoi les ressources que nous développons ne sont pas disponibles dans tel ou tel format RDF. Bien que les formats suggérés soient une option possible (on peut toujours trouver une bijection entre structure de graphe, base de données relationnelles, format tabulé, etc.), une réponse tentante est « pourquoi le seraient-elles ? ». En effet, si l'on se demande parfois quel format est approprié pour encoder quel type de données, on se demande rarement « pour faire quoi ? ». La TEI et LMF ont été conçus pour représenter des données linguistiques et sont adaptés à des ressources construites dans cette visée. RDF est en revanche adapté aux ontologies dans le cadre du web sémantique (ou web des données, ou *linked data*), mais toutes les données, selon leur utilisation, n'ont pas vocation à être représentées sous forme de graphes représentant un ensemble de triplets sujet-prédicat-objet. Si nous avons diffusé, par exemple, une version de GLAFF-IT en texte tabulé et une autre dans une implémentation de LMF, c'est d'une part parce que les deux formats se prêtent à l'encodage d'un lexique flexionnel, mais également parce que linguistes-informaticiens (ou l'inverse) et TAListes sont habitués à manipuler ces types de représentation. LMF aura la préférence d'un utilisateur à l'aise avec l'exploitation de documents XML (par programme ou via un document XSL), le texte tabulé celle des adeptes d'outils Unix en ligne de commande. Néanmoins, le format de nos ressources rendant l'extraction d'informations très simple, tout utilisateur intéressé à produire des données dans un format RDF donné (ou tout autre format), pourra aisément écrire un convertisseur à cet effet.

[11] https://www.iso.org/standard/37327.html

[12] https://www.iso.org/standard/69550.html

Concernant les DE construits à partir de Wiktionary, nous les avons encodés dans un format spécifique, pour les raisons que nous développons ci-après.

5.3. Encodage de ressources idiosyncratiques

Ide et Véronis (1995), travaillant au sein du TEI Dictionary Working Group, ont noté que les dictionnaires étaient parmi les types de texte les plus complexes traités dans la TEI, la structure de leurs entrées étant très variable, non seulement d'un dictionnaire à l'autre mais également au sein d'un même dictionnaire. Les auteurs écrivent que des éléments d'information donnés peuvent aller n'importe où dans certains dictionnaires et que, dans les grands dictionnaires complexes tels que l'*Oxford English Dictionary*, les exceptions et les organisations inhabituelles du contenu sont courantes. En conséquence, il est probablement impossible, selon Ide et Véronis, de définir une structure fixe pour de tels documents. Pour gérer cette situation, un nouvel élément (entryFree) a été ajouté à la DTD en cours de développement, permettant de combiner tous les composants du dictionnaire dans n'importe quel ordre. On ne parlait pas à l'époque d'interopérabilité. Il s'agissait de concevoir un standard pour encoder des ressources de nature similaire, tout en sachant qu'un standard ne peut inclure toutes les particularités des ressources existantes et à venir.

Un problème soulevé par Nastase et Strapparava (2015), qui travaillent à l'alignement de Wiktionary et WordNet, s'applique également au processus de mise en conformité d'une ressource particulière avec une norme donnée. De la même manière que l'alignement de ressources contenant des informations de types différents entraîne la perte d'une partie de leur contenu, une tentative d'encoder dans un format standard la connaissance singulière issue d'une ressource telle que Wiktionary entraînera la suppression des informations lexicales « non orthodoxes ». Or dans Wiktionary, l'exception est la règle.

Lors du développement de DBnary, Sérasset (2015) a décidé d'encoder plusieurs éditions de langue de Wiktionary dans le modèle LEMON, bien qu'il ait jugé celui-ci insuffisant pour représenter des données lexicales contenues dans les dictionnaires. LEMON, écrit-il, suppose que toutes les données soient bien formées et entièrement spécifiées, ce qui n'est pas le cas de Wiktionary – ni, selon Ide et Véronis (1995), celui des dictionnaires « réguliers ». En conséquence, Sérasset a dû étendre le modèle LEMON pour encoder les données de Wiktionary. C'est aussi le cas pour la plupart des modèles fondés sur LMF : ils nécessitent la modification du format ou un travail d'instanciation d'envergure, comme UBY-LMF, que Meyer (2013) propose pour encoder plusieurs ressources en adoptant un format commun. Meyer (2013, p. 136) commente le modèle de Sérasset (2012) qui, contrairement au sien, « représente des relations sémantiques entre mots graphiques, ce qui génère des proximités entre unités qui n'ont pas grand-chose à voir dans la réalité ». Meyer propose, lui, un modèle qui représente les relations sémantiques entre les sens particuliers des mots. La sémantique lexicale donne bien évidemment raison à Meyer. Mais ce que Meyer ne dit pas, c'est que l'option de Sérasset est un non-

choix qui consiste à représenter l'information telle qu'elle figure dans Wiktionary. Meyer produit quant à lui un modèle pour représenter les relations présentes dans une ressource qu'il a désambiguïsée automatiquement. Contrairement à ce qu'il écrit[13], il ne s'agit donc pas de représenter fidèlement les données de Wiktionary. Par exemple, la manière dont les définitions et les étymologies sont reliées ne correspond pas à ce que l'on trouve dans le dictionnaire et la richesse des informations contenues dans la microstructure est perdue (par ex. les étymons, la langue d'origine et le processus de formation des mots sont donnés en texte brut dans UBY-LMF, sans formatage particulier : l'encodage présent dans le wikicode est perdu). Rien n'est prévu non plus pour les prononciations (les ressources avec lesquelles Wiktionary est aligné en sont dépourvues).

Ainsi, le format standard mis en avant dans la présentation d'une ressource complexe résulte souvent soit de l'abandon d'une partie des données à encoder, soit d'une modification du standard. Aucune de ces options ne nous convient car nous souhaitons que le contenu de nos DE soit aussi proche que possible de celui des différentes éditions de langue de Wiktionary. Au lieu de tordre le contenu de Wiktionary pour l'adapter à une norme donnée – ou de tordre n'importe quelle norme afin de l'accommoder au contenu de Wiktionary –, nous avons décidé de concevoir une structure *ad hoc* permettant de modéliser la macro- et la microstructure du Wiktionnaire, de Wikizionario et du Wiktionary anglais dans GLAWI, GLAWIT et ENGLAWI. En plus de se conformer au contenu de Wiktionary, la structure des ressources française, italienne et anglaise est quasi-identique (tout en encodant la spécificité de chaque édition de langue), ce qui facilite l'adaptation d'un outil conçu pour une ressource donnée à une autre (contrairement à l'adaptation d'un extracteur wiki d'une langue à une autre, dont la difficulté est souvent sous-estimée, comme mentionné à la Section 4). Nous mettons d'ailleurs librement à disposition G-Peso[14] (Sajous, Calderone & Hathout, 2020), une série de scripts prêts à l'emploi et modifiables, permettant d'extraire des informations choisies à partir de nos DE.

6. Conclusion et perspectives

Nous exploitons depuis une décennie les éditions française, italienne et anglaise de Wiktionary et avons produit pour chacune de ces trois langues un DE et un lexique flexionnel : GLAWI et GLÀFF, GLAWIT et GLAFF-IT, ENGLAWI et ENGLAFF. Nous avons pour cela conçu des extracteurs finement calibrés pour chaque langue, qui nous ont permis de développer des ressources fidèles au contenu de Wiktionary, tout en le structurant et en le rendant exploitable. Ces ressources prêtes à l'emploi ont l'avantage sur les autres ressources existantes non seulement au plan qualitatif (cf. Section 4) mais également quantitatif. Concernant leurs nomenclatures, nous avons montré pour le

[13] « Nous modélisons Wiktionary conformément au standard ISO-LMF, que nous adaptons aux spécificités des dictionnaires collaboratifs » (Meyer, 2013, p. IV).

[14] http://redac.univ-tlse2.fr/tools/g-peto/

français leur supériorité en termes de couverture interlexique et leur meilleure capacité à couvrir divers corpus dans cette langue (Sajous et al., 2014). Nous avons également montré que nos méthodes d'extraction produisent à la fois plus de lemmes et des paradigmes flexionnels plus complets pour le français et l'anglais que les approches utilisées pour créer des ressources telles qu'UniMorph et DBnary (Sajous, Calderone & Hathout, 2020). Ces ressources ont été exploitées dans des tâches de TAL, pour dériver d'autres ressources et enrichir des bases lexicales, mais également pour des observations linguistiques et métalexicographiques[15].

La diversité des travaux menés grâce aux DE et aux lexiques que nous avons conçus à partir de Wiktionary, ainsi que leur comparaison à d'autres ressources, montrent la pertinence du recours au modèle wiki pour la constitution de RL et de l'approche d'extraction que nous avons mise en œuvre. Nous nous garderons pour autant d'enjoliver une situation qui n'est pas idéale. D'une part, ce que nous gagnons en précision grâce à la finesse de nos extractions, nous le perdons en rapidité d'actualisation : les changements incessants du wikicode rendent la maintenance de nos extracteurs longue et fastidieuse. D'autre part, Wiktionary est une ressource pour le moins perfectible : bien qu'affichant une grande richesse lexicale et une nomenclature hors norme, la qualité et la densité des relations sémantiques (ainsi que leur ancrage au niveau des mots graphiques), ou de leurs relations morphologiques, sont problématiques. Pour le français, le Wiktionnaire ne pallie pas le manque d'un WordNet de qualité, pas plus que ne peuvent le faire, isolément, les approches comme la construction automatique de ressources, l'AD ou le micro-travail. Une solution hybride faisant intervenir plusieurs méthodes différentes de conception de RL est peut-être encore à inventer. La pluralité est certainement également à encourager dans la collaboration entre acteurs du domaine : dans leur synthèse sur l'extraction de connaissances lexicales à partir de DE, Ide et Véronis (1993) concluaient sur une interaction souhaitable entre linguistes, lexicographes et TAListes. Pour la lexicographie anglo-saxonne, on peut dire que l'essai est transformé. Kilgarriff (2013) s'interrogeait sur le rôle futur, imprédictible, du lexicographe, mais pressentait que la frontière entre le dictionnaire et le corpus deviendrait moins nette, le lexicographe opérant à l'interface. Quelques années plus tard, Rundell (2017) voit une opportunité dans le recours aux foules : avec un encadrement adéquat, les amateurs peuvent, selon lui, apporter des contributions très utiles. Il évoque sa vision du processus de conception de dictionnaire comme la répartition du travail entre trois participants : les lexicographes, les machines et les amateurs volontaires. Selon Rundell, chacun de ces trois acteurs possédant ses propres atouts, l'enjeu est de trouver quelle est l'option la plus efficace pour accomplir chaque tâche. Du côté de la lexicographie française, les effets durables de l'absence d'investissement et du rejet idéologique de la linguistique de corpus par des lexicographes encore influents sont encore douloureusement perceptibles. Le manque d'outils qui découle d'une demande inexistante dans ce domaine ainsi que le manque de porosité entre les maisons d'édition et le monde académique laissent entrevoir peu de

[15] Voir Sajous, Calderone & Hathout (2020) pour une recension non exhaustive.

perspectives. Dans le domaine des RL en revanche, des jalons ont été posés : le Wiktionnaire contient des définitions et des relations lexicales, rendues exploitables grâce à GLAWI. Les relations lexicales sont néanmoins en nombre insuffisant et relient des mots graphiques. Un alignement entre les sens du Wiktionnaire et des occurrences en corpus pourrait se faire par recours au micro-travail, comme l'a montré Rumshisky (2011). Le micro-travail pourrait également être utilisé pour la construction de synsets par clusterisation (Rumshisky et al., 2012). Reste à déterminer des conditions éthiques satisfaisantes et s'assurer que les résultats des expériences préalables sont transposables aux internautes de langue française : à notre connaissance, aucune étude d'envergure ne permet d'anticiper une participation massive de locuteurs du français à des micro-tâches de nature linguistique. Enfin, on peut envisager de coupler les approches par micro-travail à celles de Sagot et Fišer (2008) qui exploitent la structure du Princeton WordNet. C'est un vaste programme qui suppose le recours à plusieurs types d'expertise (donc la collaboration entre plusieurs acteurs), la disponibilité d'un corpus approprié, et la volonté d'établir de nouvelles priorités. La première étant celle de favoriser la production de contenu, et son partage, avant tout.

Bibliographie

Bański, P., Bowers, J., & Erjavec, T. (2017). TEI-Lex0 Guidelines for the Encoding of Dictionary Information on Written and Spoken Forms. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek & V. Baisa (Eds.), *Proceedings of eLex 2017* (pp. 485-494). Leiden.

Bederson, B. B., & Quinn, A. J. (2011). Web Workers Unite! Addressing Challenges of Online Laborers. In *Proceedings of CHI 2011* (pp. 97-106). Vancouver. https://doi.org/10.1145/1979742.1979606

Brabham, D. C. (2013). *Crowdsourcing*. Cambridge: MIT Press.

Calderone, B., Sajous, F., & Hathout, N. (2016). GLAW-IT: A free large Italian dictionary encoded in a fine-grained XML format. In *Proceedings of SLE'2016* (pp. 43-45). Naples. ISBN: 978-1-4503-4447-0.

Calderone, B., Pascoli, M., Sajous, F., & Hathout, N. (2017). Hybrid Method for Stress Prediction Applied to GLAFF-IT, a Large-scale Italian Lexicon. In J. Gracia, F. Bond, J. P. McCrae, P. Buitelaar, C. Chiarcos, & S. Hellmann (Eds.), *Language, Data, and Knowledge* (pp. 26-41). Cham: Springer International Publishing.

Chamberlain, J., Kruschwitz, U., & Poesio, M. (2009). Constructing An Anaphorically Annotated Corpus With Non-Experts: Assessing The Quality Of Collaborative Annotations. In *Proceedings of the ACL-IJCNLP Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources* (pp. 57-62). Singapore.

Čibej, J., Fišer, D., & Kosem, I. (2015). The role of crowdsourcing in lexicography. In *Proceedings of eLex 2015* (pp. 70-83). Herstmonceux Castle.

Corbin, P. (1998). La lexicographie française est-elle en panne ? In *Cicle Conferències 96-97, Lèxic, corpus i diccionaris* (p. 83-112). Barcelona.

Corbin, P. (2008). Quel avenir pour la lexicographie française ? In *Actes du CMLF 2008* (p. 1227-1250). Paris. https://doi.org/10.1051/cmlf08352

Corbin, P., & Gasiglia, N. (2020). Les dictionnaires monolingues généraux du français « actuel » gratuits en ligne (début 2019). In *Actes du CMLF 2020*. Montpellier. https://doi.org/10.1051/shsconf/20207805008

Creese, S., McGillivray, B., Nesi, H., Rundell, M., & Sule, K. (2018). Everything You Always Wanted to Know about Dictionaries (But Were Afraid to Ask): A Massive Online Course. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (Eds.), *Proceedings of EURALEX 2018* (pp. 59-66). Ljubljana.

De Smedt, T., Marfia, F., Matteucci, M., & Daelemans, W. (2014). Using Wiktionary to Build an Italian Part-of-Speech Tagger. In E. Métais, M. Roche, & M. Teisseire (Eds.), *Proceedings of NLDB 2014*, LNCS, volume 8455. Cham: Springer.

Eckard, E., Barque, L., Nasr, A., & Sagot, B. (2012). Dictionary-Ontology Cross-Enrichment. Using TLFi and WOLF to enrich one another. In M. Zock & R. Reinhard (Eds.), *Proceedings of CogALex 2012* (pp. 81-93). Mumbai.

Estellés-Arolas, E., & González-Ladrón-de Guevara, F. (2012). Toward an integrated crowdsourcing definition. *Journal of Information Science*, *38*(2), 189-200.

Estellés-Arolas, E., Navarro-Giner, R., & González-Ladrón-de Guevara, F. (2015). Crowdsourcing Fundamentals: Definition and Typology. In F. J. Garrigos-Simon, I. Gil-Pechuán, & S. Estelles-Miguel (Eds.), *Advances in Crowdsourcing* (pp. 33-48). Springer International Publishing Switzerland.

Fabre, C., Hathout, N., Ho-Dac, L.-M., Morlane-Hondère, F., Muller, P., Sajous, F., Tanguy, L., & Van de Cruys, T. (2014). Présentation de l'atelier SemDis 2014 : sémantique distributionnelle pour la substitution lexicale et l'exploration de corpus spécialisés. In *Actes de l'atelier SemDis à TALN 2014* (p. 196-205). Marseille.

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

Francopoulo, G. (Ed.) (2013). *Lexical Markup Framework*. London: ISTE/Wiley.

Gurevych, I., Eckle-Kohler, J., & Matuschek, M. (2016). *Linked Lexical Knowledge Bases: Foundations and Applications*. Morgan & Claypool Publishers.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of COLING 1992* (pp. 539-545). Nantes.

Howe, J. (2006). The rise of crowdsourcing. *Wired*, 14.06.

Ide, N., & Véronis, J. (1993). Extracting Knowledge Bases from Machine-Readable Dictionaries: Have We Wasted Our Time. In *Proceedings of the KB&KS'93 Workshop* (pp. 257-266). Tokyo.

Ide, N., & Véronis, J. (1995). Encoding dictionaries. *Computers and the Humanities*, *29*(2), 167-179.

Jacquin, C., Desmontils, E., & Monceaux, L. (2007). French EuroWordNet Lexical Database Improvements. In *Proceedings of CICLing 2007* (pp. 12-22). Mexico City. ISBN: 978-3-540-70938-1.

Jurgens, D., & Navigli, R. (2014). It's All Fun and Games until Someone Annotates: Video Games with a Purpose for Linguistic Annotation. *Transactions of the ACL*, *2*, 449-163. http://dx.doi.org/10.1162/tacl_a_00195

Kilgarriff, A. (2013). Using corpora [and the web] as data sources for dictionaries. In H. Jackson (Ed.), *The Bloomsbury Companion to Lexicography* (pp. 77-96). London: Bloomsbury.

Kirov, C., Sylak-Glassman, J., Que, R., & Yarowsky, D. (2016). Very-large scale parsing and normalization of Wiktionary morphological paradigms. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis (Eds.), *Proceedings of LREC 2016* (pp. 3121-3126). Portorož.

Klementiev, A., Irvine, A., Callison-Burch, C., & Yarowsky, D. (2012). Toward Statistical Machine Translation without Parallel Corpora. In Daelemans, W. (Ed.), *Proceedings of EACL'2012* (pp. 130-140). Avignon.

Lafourcade, M. (2007). Making People Play for Lexical Acquisition with the JeuxDeMots prototype. In *Proceedings of SNLP 2007, 7th Symposium on Natural Language Processing*. Pattaya, Thailande.

Lenci, A. (2018). Distributional Models of Word Meaning. *Annual Review of Linguistics*, *4*, 151-171. https://doi.org/10.1146/annurev-linguistics-030514-125254

Lew, R. (2014). User-generated content (UGC) in online English dictionaries. *OPAL*, *4*, 8-26.

Liebeck, M., & Conrad, S. (2015). IWNLP: Inverse Wiktionary for Natural Language Processing. In C. Zong & M. Strube (Eds.), *Proceedings of IJCNLP 2015*, vol. 2 (pp. 414-418). Beijing.

Metheniti, E., & Neumann, G. (2018). Wikinflection: Massive Semi-Supervised Generation of Multilingual Inflectional Corpus from Wiktionary. In D. Haug, S. Oepen, L. Øvrelid, M. Candito & J. Hajič (Eds.), *Proceedings of TLT 2018* (pp. 147-161). Oslo.

Meyer, C. M. (2013). *Wiktionary: The Metalexicographic and the Natural Language Processing Perspective*. PhD thesis, Technische Universität Darmstadt, Darmstadt.

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, *38*(11), 39-41. https://doi.org/10.1145/219717.219748

Murray, G. C., & Green, R. (2004). Lexical Knowledge and Human Disagreement on a WSD Task. *Computer Speech & Language*, *18*(3), 209-222. https://doi.org/10.1016/j.csl.2004.05.001

Nastase, V., & Strapparava, C. (2015). knoWitiary: A Machine Readable Incarnation of Wiktionary. *International Journal of Computational Linguistics and Applications*, *6*(2), 61-82.

Navarro, E., Sajous, F., Gaume, B., Prévot, L., Hsieh, S., Kuo, I., Magistry, P., & Huang, C.-R. (2009). Wiktionary and NLP: Improving synonymy networks. In I. Gurevych & T. Zesch (Eds.), *Proceedings of the 2009 ACL-IJCNLP Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources* (pp. 19-27). Singapore.

Navigli, R., & Ponzetto, S. P. (2010). BabelNet: Building a Very Large Multilingual Semantic Network. In J. Hajič, S. Carberry, S. Clark & J. Nivre (Eds.), *Proceedings of the 48th ACL Meeting* (pp. 216-225). Uppsala.

Pierrel, J.-M. (2013). Structuration et usage de ressources lexicales institutionnelles sur le français. *Linguisticae investigationes Supplementa*, *30*, 119-152. https://doi.org/10.1075/lis.30.04pie

Plank, B., & Klerke, S. (2019). Lexical Resources for Low-Resource PoS Tagging in Neural Times. In M. Hartmann, B. Plank (Eds.), *Proceedings of NoDaLiDa 2019* (pp. 25-34). Turku.

Poesio, M., Chamberlain, J., Kruschwitz, U., Robaldo, L., & Luca, D. (2015). Phrase Detectives: Utilizing Collective Intelligence for Internet-Scale Language Resource Creation. In Q. Yang & M. Wooldridge (Eds.), *Proceedings of IJCAI 2015* (pp. 4202-4206). Buenos Aires.

Rojc, M., & Kačič, Z. (2007). Time and Space-efficient Architecture for a Corpus-based Text-to-speech Synthesis System. *Speech Communication*, *49*(3), 230-249.

Romary, L., Khemakhem, M., Khan, F., Bowers, J., Calzolari, N., George, M., Pet, M., & Bański, P. (2019). LMF Reloaded. In M. Gürlek, A. N. Çiçekler & Y. Taşdemir (Eds.), *Proceedings of AsiaLex 2019* (pp. 533-539). Istanbul.

Rumshisky, A. (2011). Crowdsourcing Word Sense Definition. In N. Ide, A. Meyers, S. Pradhan & K. Tomanek (Eds.), *Proceedings of LAW V* (pp. 74-81). Portland.

Rumshisky, A., Botchan, N., Kushkuley, S., & Pustejovsky, J. (2012). Word Sense Inventories by Non-experts. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (Eds.), *Proceedings of LREC 2012* (pp. 4055-4059). Istanbul.

Rundell, M. (2017). Dictionaries and crowdsourcing, wikis, and user-generated content. In P. Hanks & G.-M. de Schryver (Eds.), *International Handbook of Modern Lexis and Lexicography*. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-45369-4_26-1

Sagot, B., & Fišer, D. (2008). Construction d'un wordnet libre du français à partir de ressources multilingues. In *Actes de TALN 2008*. Avignon.

Sajous, F., Hathout, N., & Calderone, B. (2013). GLÀFF, un Gros Lexique À tout Faire du Français. In *Actes de TALN 2013* (pp. 285-298). Les Sables d'Olonne.

Sajous, F., Hathout, N., & Calderone, B. (2014). Ne jetons pas le Wiktionnaire avec l'oripeau du web ! Études et réalisations fondées sur le dictionnaire collaboratif. In F. Neveu, P. Blumenthal, L. Hriba,

A. Gerstenberg, J. Meinschaefer et S. Prévost (dir.), *Actes de CMLF 2014* (p. 663-680). Berlin. https://doi.org/10.1051/shsconf/20140801216

Sajous, F., & Hathout, N. (2015). GLAWI, a free XML-encoded Machine-Readable Dictionary built from the French Wiktionary. In *Proceedings of eLex 2015* (pp. 405-426). Herstmonceux Castle.

Sajous, F., Josselin-Leray, A., & Hathout, N. (2018). The Complementarity of Crowdsourced Dictionaries and Professional Dictionaries viewed through the Filter of Neology. *Lexis*, *12*. https://doi.org/10.4000/lexis.2322

Sajous, F., Hathout, N., & Josselin-Leray, A. (2019). Du vin et devin dans le Wiktionnaire : neutralité de point de vue ou neutralité *et* point de vue ? *Études de linguistique appliquée*, *194*(2), 147-164.

Sajous, F., Calderone, B., & Hathout, N. (2020). ENGLAWI: From Human-to Machine-Readable Wiktionary. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis (Eds.), *Proceedings LREC 2020* (pp. 3016-3026). Marseille.

Sajous, F., Josselin-Leray, A., & Hathout, N. (2020). Les domaines de spécialité dans les dictionnaires généraux : le lexique de l'informatique analysé par les foules et par les professionnels... de la lexicographie. *Neologica*, *14*, 83-107.

Schlippe, T., Ochs, S., & Schultz, T. (2010). Wiktionary as a source for automatic pronunciation extraction. In T. Kobayashi, K. Hirose & S. Nakamura (Eds.), *Proceedings of INTERSPEECH 2010* (pp. 2290-2293). Makuhari.

Segonne, V., Candito, M., & Crabbé, B. (2019). Using Wiktionary as a resource for WSD: the case of French verbs. In S. Dobnik, S. Chatzikyriakidis & V. Demberg (Eds.). *Proceedings of IWCS 2019* (pp. 259-270). Gothenburg. http://dx.doi.org/10.18653/v1/W19-0422

Sekine, S. (2010). We Desperately Need Linguistic Resources! –Based on the Users' Point of View. *FLaReNet Forum*, Barcelona.

Sérasset, G. (2012). Dbnary: Wiktionary as a LMF based Multilingual RDF network. In *Proceedings of LREC 2012* (pp. 2466–2472). Istanbul.

Sérasset, G. (2015). DBnary: Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF. *Semantic Web 6*(4), 355-361.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In M. Lapata, H. T. Ng (Eds.), *Proceedings of EMNLP 2008* (pp. 254-263). Honolulu.

Tanguy, L., Urieli, A., Calderone, B., Hathout, N., & Sajous, F. (2011). A multitude of linguistically-rich features for authorship attribution. In V. Petras, P. Forner & P. D. Clough (Eds.), *Notebook for PAN at CLEF 2011*. Amsterdam.

Tiedemann, J., & Ljubešić, N. (2012). Efficient Discrimination Between Closely Related Languages. In M. Zock & R. Reinhard (Eds.), *Proceedings of COLING 2012* (pp. 2619-2634). Mumbai.

Vossen, P. (Ed.) (1998). *EuroWordNet: a multilingual database with lexical semantic networks*. Norwell: Kluwer Academic Publishers.

Weale, T., Brew, C., & Fosler-Lussier, E. (2009). Using the Wiktionary Graph Structure for Synonym Detection. In I. Gurevych & T. Zesch (Eds.), *Proceedings of the ACL-IJCNLP Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources* (pp. 28-31). Singapore.

Zesch, T., & Gurevych, I. (2010). Wisdom of Crowds versus Wisdom of Linguists – Measuring the Semantic Relatedness of Words. *Journal of Natural Language Engineering 16*(1), 25-59.

Zesch, T., Müller, C., & Gurevych, I. (2008a). Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis & D. Tapias (Eds.), *Proceedings of LREC 2008* (pp. 1648-1652). Marrakech.

Zesch, T., Müller, C., & Gurevych, I. (2008b). Using Wiktionary for Computing Semantic Relatedness. In *Proceedings of AAAI 2008* (pp. 861-866). Chicago.