HAL open science

État de l'art : Liage de ressources lexicales du français

Hee-Soo Choi

▶ **To cite this version:**

Hee-Soo Choi. État de l'art : Liage de ressources lexicales du français. RÉCITAL 2022 - 24e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues, Jun 2022, Avignon, France. hal-03675515v1

HAL Id: hal-03675515

https://hal.science/hal-03675515v1

Submitted on 23 May 2022 (v1), last revised 8 Jun 2022 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

État de l'art : Liage de ressources lexicales du français

Hee-Soo Choi[1, 2]
(1) ATILF, Universite de Lorraine, CNRS, F-54000 Nancy, France
(2) Inria, LORIA, Universite de Lorraine, CNRS, F-54000 Nancy, France
hee-soo.choi@loria.fr

RÉSUMÉ

Les ressources lexicales informatisées constituent des données indispensables à l'élaboration d'outils et de méthodes répondant aux différentes tâches de Traitement Automatique des Langues (TAL). Celles-ci sont hétérogènes dans leur taille, leur construction et leur niveau de description linguistique. Cette variété ouvre la porte à un regroupement des ressources ou à des tentatives de liage. Dans cet article, nous présentons un état de l'art sur les ressources lexicales du français. Plus précisément, nous abordons les différentes caractéristiques d'une ressource lexicale, les ressources construites à partir de liage ainsi que les approches employées à cette fin.

ABSTRACT

State of the art : Linking French Lexical Resources

Lexical resources are essential for the development of tools and methods for the various tasks of Natural Language Processing (NLP). These resources are heterogeneous in their size, their construction and their level of linguistic description. This variety opens the way to group or link these resources. In this article, we present a state of the art on French lexical resources. More precisely, we discuss the different features of a lexical resource, the resources based on linking, and the approaches used for this purpose.

MOTS-CLÉS : liage, alignement, mapping, ressources lexicales, état de l'art, français.

KEYWORDS: linking, alignment, mapping, lexical resources, state of the art, french.

1 Introduction

Les ressources lexicales informatisées ou *machine-readable* constituent des données indispensables pour développer des outils et des méthodes réalisant des tâches de Traitement Automatique des Langues (TAL). Du fait de l'influence des ressources sur les performances des systèmes, utiliser des données adéquates à une tâche se trouve être un des défis majeurs du domaine et nous incite à nous intéresser à l'objet même de ressource. La notion de ressource lexicale reste large dans la mesure où elle désigne des entités variées comme des dictionnaires, des lexiques ou des réseaux lexico-sémantiques mais qui ont pour point commun de décrire la langue. Cette variété ouvre alors la porte à un potentiel regroupement de ces ressources ou à des tentatives de liage. Par liage, nous désignons au sens large le processus de mise en correspondance sur le plan morphologique, syntaxique et sémantique d'entrées équivalentes provenant de deux (ou plus) ressources lexicales [1]. On peut

1. Nous sommes conscients que le terme de « liage » tend à faire référence aux *Linguistic Linked Open Data*. Un meilleur terme est encore en cours de recherche.

également retrouver les termes *mapping* et « alignement » dans la littérature (Gurevych *et al.*, 2012).

Si les premières ressources ont été créées manuellement par des linguistes et lexicographes, l'aspect chronophage de la tâche pousse à utiliser de plus en plus de techniques (semi-)automatiques et collaboratives qui soulèvent en contrepartie la question de la qualité des données. L'approche intermédiaire consiste alors à lier des ressources existantes afin d'augmenter la couverture tout en conservant la qualité des données issues de sources expertes.

Dans cet article, nous nous intéressons plus particulièrement aux ressources lexicales du français en présentant dans un premier temps, la notion de ressource lexicale et les critères à l'origine des différents types de ressource dont nous disposons. Dans un deuxième temps, nous abordons les ressources qui ont tiré profit de l'hétérogénéité des ressources existantes dans leur conception. Et enfin, nous terminons par décrire les techniques utilisées pour le liage de certaines ressources du français. Il est également important de souligner que nous focaliserons notre analyse sur les ressources du français libres d'utilisation.

2 Notion de ressource lexicale

Au sens large, une ressource lexicale décrit la langue en associant des mots ou des concepts à des informations de nature variée : des traductions vers une ou plusieurs langues dans les ressources multilingues, des explications à caractère linguistique comme l'origine ou les caractéristiques grammaticales ou à caractère conceptuel comme des liens thématiques ou des relations lexicales (Gala, 2013). Une ressource lexicale peut alors désigner des entités hétérogènes selon des critères souvent corrélés : son contenu linguistique, son type, sa taille et la manière dont elle a été construite.

Contenu linguistique et type de ressource Le niveau de description linguistique considéré, s'il a trait à la morphologie, la phonologie, la syntaxe ou la sémantique peut avoir une influence sur la représentation formelle de la ressource, donnant ainsi lieu à différents types de ressource. En effet, il est commun de retrouver des informations morphologiques et syntaxiques dans des dictionnaires ou des lexiques avec une structure de liste d'items. À titre d'exemple, on peut notamment citer Morphalou (ATILF, 2019), lexique des formes fléchies du français et Dicovalence (Van den Eynde & Mertens, 2006; Mertens, 2010), dictionnaire décrivant les cadres de valence de plus de 3 700 verbes simples (cf. Figure 1).

```
VAL$      commencer à: | AdjunctVerb
VTYPE$    adjunct_verb:2 adjunct_prep:à
VERB$     COMMENCER/commencer
NUM$      16850
EG$       ils commencent à chanter
TR_DU$    beginnen
TR_EN$    start (doing sth)
FRAME$
AUX$      avoir
```

FIGURE 1 – Exemple d'entrée de Dicovalence (extrait de Van den Eynde & Mertens (2010), page 10).

La sémantique est, quant à elle, davantage représentée sous forme de réseau. Pour le français, nous

pouvons citer les réseaux lexico-sémantiques, RL-fr, *Réseau Lexical du Français* (Lux-Pogodalla & Polguère, 2011), où les nœuds correspondent aux unités lexicales et les arêtes à des relations lexicales sémantiques ou combinatoires (cf. Figure 2) et JeuxDeMots (Lafourcade & Joubert, 2008; Lafourcade & Le Brun, 2020), gros réseau de plus de 5 000 000 de nœuds et 400 000 000 de relations, créé collaborativement à travers des jeux ludiques en ligne. Au-delà du français, le Princeton WordNet (PWN) (Fellbaum, 1998) constitue un des réseaux lexicaux majeurs de l'anglais. Élaboré par des psychologues cognitifs, les mots sont regroupés en ensembles de quasi-synonymes, les *synsets* qui expriment chacun un concept lexical. Les *synsets* sont ensuite reliés entre eux par des relations conceptuelles, sémantiques et lexicales [2]. Sous l'impulsion du PWN, d'autres *Wordnets* ont été construits dans différentes langues, dont le français avec le French WordNet issu du projet EuroWordNet (Vossen, 1998) et WOLF (Sagot & Fišer, 2008). Ce dernier a l'avantage d'être libre d'utilisation contrairement au French WordNet, qui n'a été que très peu exploité en raison de sa licence restrictive.

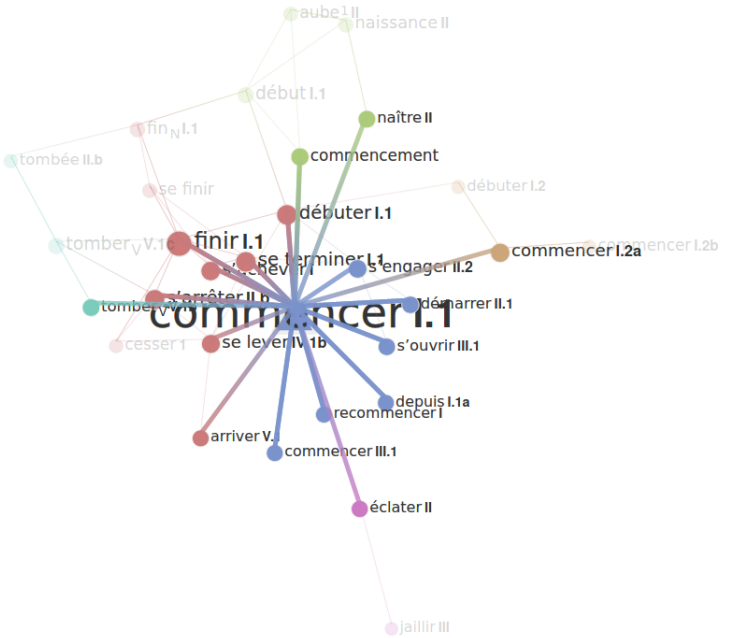

FIGURE 2 – Exemple d'entrée du RL-fr (représentation avec Spiderlex [3]).

La corrélation entre le contenu linguistique décrit dans la ressource et sa structure n'est cependant pas systématique puisqu'il est également possible de représenter des informations morphologiques en réseau à l'image de Morphonette (Hathout, 2010).

Méthodes de construction Si la construction de ressources lexicales repose sur une longue tradition d'un travail manuel effectué par des lexicographes, de nouvelles méthodes se sont développées avec l'évolution de l'informatique et du TAL.

2. https://wordnet.princeton.edu/
3. https://spiderlex.atilf.fr/fr/q/*commencer***I.1, le 11 mai 2022.

Nous pouvons distinguer trois types de méthodes de construction :

1. la construction par des « experts », en ce sens, des lexicographes *via* un travail manuel,
2. la construction par la foule, de manière collaborative,
3. la construction automatique en exploitant des ressources existantes.

Initialement, les ressources décrivant la langue sont largement représentées par les dictionnaires construits par les lexicographes dans un but descriptif et didactique. Les mots sont ainsi listés par ordre alphabétique et associés à des informations d'ordre morphologique, phonologique, syntaxique et sémantique à travers des exemples. Ce processus est toutefois long, coûteux, peu dynamique et d'une couverture limitée bien que la qualité des données est préservée.

Avec le développement de l'informatique, l'approche collaborative, appelée aussi peuplonomie ou myriadisation, permet de faire participer des personnes non expertes, c'est-à-dire sans formation linguistique ou lexicographique à proprement parler, mais intéressées par la langue. Le principe est de solliciter un grand nombre de personnes pour effectuer de petites tâches difficilement automatisables (Lafourcade & Joubert, 2013). La contribution de la foule peut être rémunérée ou non, comme par exemple dans Wikipedia, ressource phare issue d'une construction collaborative bénévole. Quant à l'approche collaborative rémunérée à la tâche, elle fait intervenir des problématiques éthiques qui ont été soulevées notamment dans le système Amazon Mechanical Turk (Fort *et al.*, 2011). Si la participation à l'élaboration de Wikipedia et ses ressources connexes est motivée par l'intérêt de la foule sur un sujet particulier, il est également possible de la faire participer à travers des jeux ayant des buts. C'est le cas de JeuxDeMots (Lafourcade & Joubert, 2008; Lafourcade & Le Brun, 2020), ressource lexico-sémantique créée à partir de divers jeux en ligne. L'avantage de cette méthode est d'obtenir une grande quantité de données de manière rapide, peu coûteuse et dynamique mais questionne la qualité des données recueillies.

Enfin, la dernière approche consiste à utiliser des ressources pré-existantes afin de créer une nouvelle ressource plus complète ou dans un autre formalisme, comme c'est le cas dans WOLF (Sagot & Fišer, 2008) ou d'enrichir une ressource pré-existante automatiquement comme dans la version 2 de Demonette (Namer *et al.*, 2019), une version de Demonette 1.2 (Hathout & Namer, 2014) enrichie avec le GLÀFF (Sajous *et al.*, 2013). Nous nous intéressons plus en détails aux ressources lexicales issues de ce mode de construction dans la section 3.

3 Exploiter l'hétérogénéité des ressources lexicales

La richesse qui découle de l'hétérogénéité des ressources peut être exploitée dans la mesure où celles-ci présentent des données différentes et généralement complémentaires leur permettant ainsi de s'enrichir mutuellement. Nous présentons ici un inventaire partiel de ressources lexicales du français libres d'utilisation issues de ressources existantes (cf. Figure 3).

Utilisation de ressources pré-existantes Le Le*fff* (Sagot, 2010), le *Lexique des Formes Fléchies du Français*, est un large lexique flexionnel issu des tables du Lexique-Grammaire (Gross, 1975), du LVF (Dubois & Dubois-Charlier, 1997) et de Dicovalence (Van den Eynde & Mertens, 2006; Mertens, 2010) mais également de l'acquisition automatique avec validation manuelle de données sur corpus et d'informations syntaxiques.

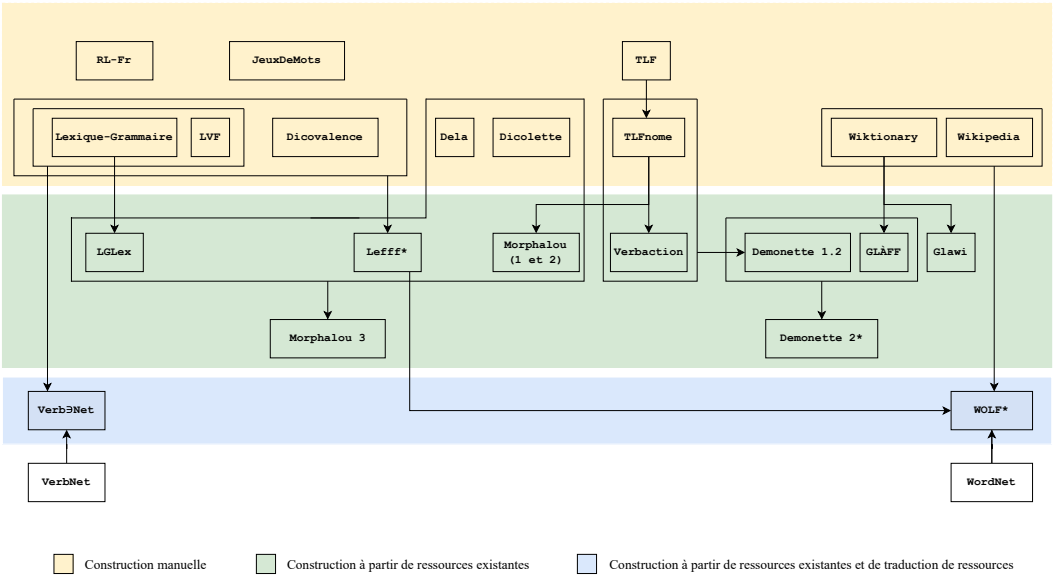

FIGURE 3 – Ressources lexicales du français créées à partir de ressources existantes (les ressources suivies d'un astérisque ont été créées avec d'autres ressources supplémentaires qui ne figurent pas sur ce schéma.).

Le Le*fff* a notamment été utilisé pour créer la troisième version de Morphalou (ATILF, 2019), également un lexique des formes fléchies du français qui résulte de la fusion de cinq ressources : Morphalou 2, Le*fff*, Dela, Dicolette et LGLex. Initialement, Morphalou est issu du TLFnome, la nomenclature du *Trésor de la Langue Française* et Morphalou 2 est une version contenant 30 000 lemmes supplémentaires provenant du TLFi. Si les deux ressources traitent de la morphologie, le Le*fff* couvre en plus des informations syntaxiques, tandis que Morphalou propose des transcriptions phonétiques d'une partie des lemmes et des formes fléchies.

Dans le même esprit, nous retrouvons le GLÀFF, *Un Gros Lexique À tout Faire du Français* (Sajous *et al.*, 2013), créé à partir du Wiktionnaire [4]. Outre les informations morphologiques et syntaxiques, il présente les transcriptions phonémiques et les fréquences de la forme et du lemme dans différents corpus.

Quant à Demonette 1.2 (Hathout & Namer, 2014), base lexicale morpho-sémantique du français organisée en réseau dérivationnel, elle provient du TLFnome et de Verbaction, lui-même créé à partir du TLFnome (Hathout *et al.*, 2002; Tanguy & Hathout, 2002). Demonette 2 a ensuite été construit à partir de Demonette 1.2, du GLÀFF et de ressources dérivationnelles développées et validées par des morphologues (Namer *et al.*, 2019).

Enfin, plus récemment, le graphe de connaissances multi-niveaux Holinet (Prost, 2022) a été créé en intégrant une couche lexicale et sémantique à partir de JeuxDeMots et une couche syntaxique à partir du French Treebank (Abeillé & Barrier, 2004).

4. https://fr.wiktionary.org/

Traduction de ressources Les ressources précédemment présentées ont été créées à partir de ressources construites manuellement, que ce soit par des experts ou par la communauté de manière collaborative. D'autres ressources ont été construites de la même manière en ajoutant des traductions de ressources dans une autre langue. C'est le cas de Verb∋Net (Danlos *et al.*, 2014), WOLF (Sagot & Fišer, 2008) et French FrameNet (Candito *et al.*, 2014; Djemaa, 2017).

Verb∋Net a été créé à partir de VerbNet (Kipper-Schuler, 2005), un lexique inspiré des classes de Levin (Levin, 1993) qui décrit les caractéristiques syntaxiques et sémantiques des verbes en anglais. Pour le français, les deux ressources Lexique-Grammaire et LVF ont été utilisées pour faire correspondre leurs classes à celles de VerbNet. Danlos *et al.* (2014) ont tenu à conserver au maximum la structure des 270 classes de VerbNet. Ainsi, pour préserver la qualité des données, la ressource a été majoritairement construite manuellement avec une phase de traduction des verbes anglais de VerbNet *via* des dictionnaires bilingues, Wiktionnaire et SCI-FRAN-EURADIC.

Dans le même esprit, WOLF, *WOrdNet Libre du Français*, a été créé sur la base de PWN (Fellbaum, 1998), de ressources multilingues et de lexiques bilingues extraits de Wikipedia [5] et Wiktionary [6]. Dans WOLF, Sagot & Fišer (2008) adoptent deux approches : une approche de traduction et une approche d'alignement. L'approche par traduction consiste à traduire les entrées monosémiques du PWN en utilisant Wikipedia et ses ressources associées (Wiktionnary anglais et français, Wikispecies) et le thesaurus multilingue Eurovoc. Quant à l'approche d'alignement, elle permet de traiter les entrées polysémiques en les alignant sur un corpus parallèle en cinq langues (le français, l'anglais, le bulgare, le roumain et le tchèque) de JRC-Acquis (Steinberger *et al.*, 2006). L'idée est d'extraire des informations sémantiques des traductions pour ensuite lever l'ambiguïté des entrées grâce aux Wordnets de chaque langue. Quantitativement, cette méthode a permis au WOLF de couvrir davantage de *synsets* (32 351) que le FrenchWordNet (22 121). Le traitement des adverbes a également fait l'objet d'un enrichissement à partir de DicoSyn [7], un dictionnaire de synonymes, et du Le*fff* (Sagot *et al.*, 2009).

Quant au French FrameNet, il consiste en une version française du FrameNet (Baker *et al.*, 1998), une base de données lexicale reposant sur la sémantique des cadres. Le French FrameNet vise à associer les entrées lexicales aux cadres de FrameNet et à fournir une couche d'annotations sémantiques sur le French Treebank (Abeillé & Barrier, 2004) et le corpus Sequoia (Candito & Seddah, 2012).

En outre, nous pouvons également citer BabelNet (Navigli & Ponzetto, 2012), réseau sémantique multilingue contenant du français, créé automatiquement à partir de Wikipedia et PWN. Inspiré des WordNets, BabelNet regroupe des *Babel synsets* en plusieurs langues auxquels sont associés des définitions. Des techniques de traduction automatique sont également appliquées pour les entrées dans les langues peu dotées.

Dans la partie suivante, nous nous concentrons sur les aspects à prendre en compte dans l'automatisation des liages de ressources, notamment la question de la représentation des données. Ensuite, nous développerons les techniques de liage utilisées pour certaines ressources du français.

5. https://fr.wikipedia.org/
6. https://www.wiktionary.org/
7. https://crisco2.unicaen.fr/des/

4 Vers une automatisation du liage

4.1 La question de la représentation des données

L'une des difficultés majeures dans le liage de ressources repose sur l'hétérogénéité de la représentation des données qui fait intervenir à la fois le format au sens strict mais aussi les formalismes linguistiques. Avec le développement du TAL, standardiser le format des ressources lexicales est devenue essentiel pour favoriser les interactions entre les différentes ressources.

C'est dans cet objectif que le Lexical Markup Framework (LMF) (Francopoulo *et al.*, 2006), standard ISO/TC37 pour les lexiques du TAL, a été créé. Le modèle LMF se concentre sur la représentation linguistique des données lexicales et non sur la représentation du monde. Il est composé d'un *core package*, le noyau qui décrit les informations de l'entrée lexicale, et de plusieurs extensions à ce noyau. Le LMF se veut adaptable à des ressources monolingues et multilingues traitant de l'écrit comme de l'oral et couvrant aussi bien la morphologie, la syntaxe et la sémantique. De nombreuses ressources ont été adaptées au modèle LMF comme Morphalou, les WordNets (Henrich & Hinrichs, 2010) et UBY (Gurevych *et al.*, 2012). Dans le Le*fff*, Sagot (2010) développe un formalisme lexical appelé Alexina, *Architecture pour les LEXiques INformatiques et leur Acquisition* qui permet la modélisation et l'acquisition lexicale couvrant à la fois les niveaux morphologique et syntaxique. Le modèle est indépendant de la langue et compatible avec le standard LMF. Cela a donc permis la création d'autres lexiques Alexina dans d'autres langues comme par exemple en espagnol avec le Le*ff*e (Molinero *et al.*, 2009).

L'interopérabilité entre les ressources constitue également un des défis du mouvement LLOD [8], *Linguistic Linked Open Data* qui vise à représenter les ressources en se basant les principes du *Linked Data* (Chiarcos *et al.*, 2013). Les ressources sont représentées sous des formats ouverts comme le RDF (Resource Description Framework) et leurs entrées lexicales sont désignées par des URIs (Uniform Resource Identifier) permettant ainsi l'accès à des informations issues d'autres ressources auxquelles elles sont liées. Différents types de ressources y sont représentés, notamment des dictionnaires, des lexiques, des ontologies et des terminologies. Face à des difficultés d'interopérabilité au niveau sémantique entre plusieurs ressources en raison du format RDF, le modèle *lemon* a été créé en s'appuyant sur une interface ontologie-lexique et permet alors de séparer l'expression d'un concept en langage naturel et sa description formelle dans l'ontologie (McCrae *et al.*, 2012). Le modèle a ensuite donné lieu à des améliorations sous Ontolex-Lemon (McCrae *et al.*, 2017). Dans les ressources du français, JeuxDeMots a notamment été converti au format Ontolex-lemon *via* des liens avec BabelNet (Navigli & Ponzetto, 2012), lui permettant ainsi d'intégrer les LLOD (Tchechmedjiev *et al.*, 2017).

4.2 Les approches de liage

Face à la diversité des ressources lexicales et leur complémentarité, l'ambition d'un liage automatique est de plus en plus envisagé en TAL. Toutefois, l'hétérogénéité des ressources mais également les ambiguïtés de la langue impliquent généralement une validation manuelle du liage automatique effectué. On parle alors de méthodes semi-automatiques. En outre, le niveau de description linguistique joue un rôle dans le type de liage. En effet, si les informations morphologiques et syntaxiques peuvent

8. https://linguistic-lod.org/

être déduites selon des règles, les informations sémantiques font généralement intervenir des mesures de similarité et des algorithmes de désambiguïsation (Gurevych *et al.*, 2016).

Approche par règles Un exemple de ressource morpho-syntaxique créée à base de règles est le Le*fff*. Le modèle lexical Alexina est basé sur deux niveaux de représentation qui sépare la description du lexique de son utilisation. Le lexique intensionnel décrit pour chaque entrée lexicale son lemme et des informations syntaxiques tandis que le lexique extensionnel, généré automatiquement à partir de règles lexicales appliquées sur le lexique intensionnel, associe chaque forme fléchie à une entrée et ses informations morphologiques. Les ressources utilisées sont préalablement converties au format Alexina, puis fusionnées automatiquement avec le Le*fff* après validation manuelle (Sagot, 2010).

Pour évaluer la ressource, deux expériences ont été faites en utilisant un outil de TAL :

1. une évaluation d'un étiqueteur morpho-syntaxique avec/sans le Le*fff*,
2. une évaluation d'un analyseur syntaxique avec le Le*fff*/avec un autre lexique.

Denis & Sagot (2009) ont comparé les performances d'un étiqueteur morpho-syntaxique avec et sans les informations lexicales extraites du Le*fff*. L'étiqueteur entraîné uniquement sur le French Treebank (Abeillé & Barrier, 2004) donne une exactitude de 97% et 86,1% sur les mots inconnus du corpus d'entraînement. En ajoutant le Le*fff* au modèle, l'exactitude augmente à 97,7% et à 90,1% pour les mots inconnus du corpus d'entraînement.

Une deuxième expérience a été effectuée avec l'analyseur syntaxique FRMG qui se base initialement sur le Le*fff* (Thomasset & Villemonte De La Clergerie, 2005). Les tables du Lexique-Grammaire ayant été converties suivant le formalisme Alexina, les auteurs ont remplacé les entrées verbales du Le*fff* par celles du Lexique-Grammaire. Ils obtiennent des résultats légèrement meilleurs avec le Le*fff* avec des f-mesures de 59,9% avec le Le*fff* et de 56,6% avec les tables du Lexique-Grammaire.

Guillaume *et al.* (2014) proposent une autre technique semi-automatique de liage en utilisant un analyseur syntaxique par règles, Leopar (Perrier & Guillaume, 2013). L'objectif est d'enrichir le lexique FRILEX, une version de Dicovalence convertie au format Leopar, avec le LVF. LVF et Dicovalence étant basés sur des théories linguistiques différentes, un liage direct s'avère impossible. Les auteurs prennent l'exemple du verbe « compter » qui présente 18 sens dans LVF et 17 dans Dicovalence. L'idée est alors de se baser sur les exemples fournis par le LVF et Dicovalence. Un des sens de « compter » de Dicovalence est utilisé dans les règles de Leopar qui est appliqué sur des exemples de LVF. Si l'exemple est analysé, on considère qu'il y a une compatibilité syntaxique entre le sens du LVF et le sens de Dicovalence. Dans le cas contraire, cela signifie que les sens ne correspondent pas. Pour évaluer le liage effectué, un expert a lié manuellement les sens de LVF et Dicovalence. Pour 14 sens de « compter », le sens correct de Dicovalence est un des sens donnés par le liage automatique. Pour trois sens, le sens correct n'a pas été pris en compte dans la conversion de Dicovalence en FRILEX et pour le dernier sens, Leopar n'a pas donné le bon résultat.

Approche par alignement L'alignement automatique de sens (*Word Sense Alignment* en anglais) est défini comme une liste de paires de sens de deux ressources, où les éléments de chaque paire représentent une signification équivalente (Matuschek, 2015). Cette approche soulève des problématiques liées aux méthodes pour déterminer automatiquement deux sens équivalents, d'autant plus lorsque les ressources ne présentent pas le même niveau de granularité dans sa description des données. Si un humain peut déterminer par intuition deux sens équivalents, les systèmes automatiques nécessitent

de caractériser la proximité sémantique avec des mesures de similarité en utilisant des informations porteuses de sens comme des exemples ou des définitions (Tchechmedjiev, 2016).

Pour le français, nous pouvons citer deux exemples de ressources qui présentent des techniques d'alignement : JeuxDeMots (Lafourcade & Joubert, 2008; Lafourcade & Le Brun, 2020) et DBnary (Sérasset, 2015).

Afin d'intégrer JeuxDeMots au LLOD, Tchechmedjiev *et al.* (2017) choisissent de le lier à BabelNet. Le choix s'est porté sur cette ressource dans la mesure où JeuxDeMots ne présente pas de définition contrairement à BabelNet. La première étape avant le liage est de convertir JeuxDeMots au format Ontolex-lemon dont les éléments de base sont les entrées lexicales et les sens lexicaux. Comme le montre la figure 4, les nœuds sources d'une relation de raffinement (*REFINE*) deviennent des entrées lexicales et les nœuds qui en sont cibles deviennent des sens lexicaux. Pour les relations d'association (*ASSOC*) dont la source sont des nœuds de raffinement comme « frégate > navire », le concept lexical qui correspond au nœud de raffinement est lié aux entrées lexicales des mots correspondants par la propriété ontolex :evokes/ontolex :isEvokedBy, dans ce cas-ci « navire », « bâteau », « marine ».

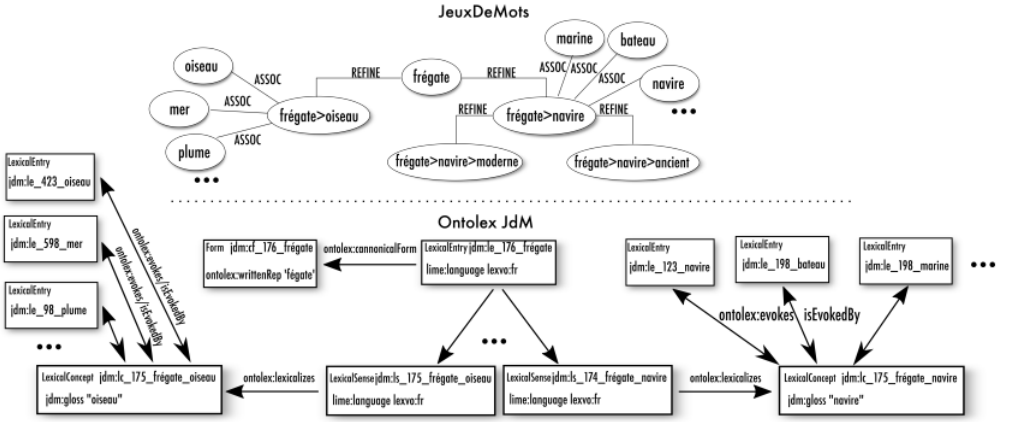

FIGURE 4 – Un exemple de la conversion des nœuds de termes et des raffinements vers le modèle Ontolex pour le terme « frégate » (Figure 1 de Tchechmedjiev *et al.* (2017)).

Ensuite, pour aligner Ontolex-JeuxDeMots à BabelNet, un algorithme d'alignement JdMBabelizer, inspiré de Pilehvar & Navigli (2014) est utilisé. Ce dernier se base sur un critère de décision calculé à partir d'une similarité pondérée de Lesk, où les poids des relations de JeuxDeMots et les fréquences relatives normalisées des mots des définitions de BabelNet sont pris en compte.

Dans Tchechmedjiev (2016), l'auteur s'intéresse à un alignement automatique de sens entre plusieurs langues dans DBnary (Sérasset, 2015). DBnary est une ressource multilingue de données lexicales extraites de Wiktionary présentant un ensemble de dictionnaires monolingues dont les entrées lexicales, appelés vocables, sont associés à des traductions bilingues. Les relations de traduction étant rattachées aux vocables et non aux sens (cf. Figure 5), Tchechmedjiev (2016) compare les gloses associées aux sens avec des mesures de similarité sémantique.

Pour certaines langues, notamment les langues agglutinantes comme l'allemand ou le finnois, les mesures s'avèrent peu adaptées dans la mesure où elle repose sur des comparaisons exactes de mots.

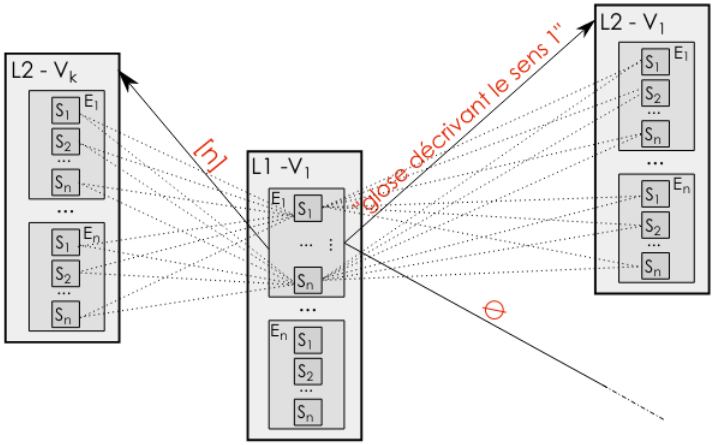

FIGURE 5 – Les données de DBnary post-extraction sans pré-traitement des relations de traductions (Figure 4.1 de Tchechmedjiev (2016)).

La comparaison est alors adoucie en reconnaissant des similarités partielles entre des mots, grâce à des similarités de chaînes. Les différentes méthodes appliquées permettent d'obtenir un alignement bilingue d'un sens d'une langue A à un sens d'une langue B.

5 Conclusion

À travers cet article, nous avons établi un état de l'art sur les ressources lexicales du français libres d'utilisation pour aller vers un possible liage de ces ressources qui permettrait de profiter des spécificités de chacune dans des tâches de Traitement Automatique des Langues. Dans un premier temps, nous avons présenté les différentes caractéristiques d'une ressource lexicale. Cette dernière est définie comme un objet qui décrit la langue construit initialement selon une longue tradition de travail manuel, par des experts. Les approches se diversifient avec le développement de l'informatique et du TAL, permettant ainsi de s'affranchir du processus long et coûteux de la construction manuelle. Les méthodes automatiques ou par la foule nécessitent toutefois un contrôle et une évaluation sur la qualité des données qui est généralement moindre que celle des ressources expertes.

Au-delà des méthodes de construction, les ressources lexicales peuvent prendre plusieurs formes selon la nature des informations décrites, si elles sont d'ordre syntaxique, morphologique ou sémantique. On retrouve alors différents objets tels que des dictionnaires, des réseaux lexico-sémantiques ou des bases de données. La diversité des informations décrites amène à un regroupement ou un liage afin d'obtenir une couverture plus grande, que ce soit en largeur, qu'en profondeur. Nous avons ainsi présenté plusieurs ressources issues de ressources existantes et les méthodes automatiques utilisées.

Bien qu'une intersection entre ces ressources est certainement présente, le liage automatique reste une tâche délicate qui nécessite une évaluation et une validation humaine. En outre, la notion de ressource lexicale reste relativement large et pourrait englober d'autres entités si l'on confronte deux dimensions : sa construction et son utilisation. La question se pose particulièrement pour les

plongements lexicaux. En effet, ces objets constituent des informations sémantiques de la langue, représentées non pas symboliquement mais par le biais de vecteurs, et sont de plus en plus présents dans les systèmes de TAL. Si leur utilisation dans ces systèmes est apparentée à celle des ressources lexicales, leur construction et notamment leur opacité peuvent l'éloigner de la définition primaire de ressource lexicale. Leur statut hybride ouvre toutefois la porte à de nouvelles perspectives de liage dans la mesure où ils pourraient être utilisés dans le but de pallier l'hétérogénéité des ressources lexicales.

Remerciements

Je remercie vivement les relecteurs pour leurs remarques avisées qui ont permis l'amélioration de cet article. Je tiens également à remercier mes directeurs de thèse, Mathieu Constant, Karën Fort et Bruno Guillaume pour nos discussions et leurs retours, sans lesquels cet article n'aurait vu le jour.

Références

ABEILLÉ A. & BARRIER N. (2004). Enriching a French treebank. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal : European Language Resources Association (ELRA).

ATILF (2019). Morphalou. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

BAKER C. F., FILLMORE C. J. & LOWE J. B. (1998). The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, p. 86–90, Montreal, Quebec, Canada : Association for Computational Linguistics. DOI : 10.3115/980845.980860.

CANDITO M., AMSILI P., BARQUE L., BENAMARA F., DE CHALENDAR G., DJEMAA M., HAAS P., HUYGHE R., MATHIEU Y. Y., MULLER P., SAGOT B. & VIEU L. (2014). Developing a French FrameNet : Methodology and First results. In *LREC - The 9th edition of the Language Resources and Evaluation Conference*, Reykjavik, Iceland. HAL : hal-01022385.

CANDITO M. & SEDDAH D. (2012). Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *TALN 2012 - 19e conférence sur le Traitement Automatique des Langues Naturelles*, Grenoble, France. HAL : hal-00698938.

CHIARCOS C., CIMIANO P., DECLERCK T. & MCCRAE J. P. (2013). Linguistic linked open data (LLOD). introduction and overview. In *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013) : Representing and linking lexicons, terminologies and other language data*, p. i – xi, Pisa, Italy : Association for Computational Linguistics.

DANLOS L., NAKAMURA T. & PRADET Q. (2014). Vers la création d'un verbenet du français. In *Atelier FondamenTAL, TALN 2014*.

DENIS P. & SAGOT B. (2009). Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort. In *Pacific Asia Conference on Language, Information and Computation*, Hong Kong, China. HAL : inria-00514366.

DJEMAA M. (2017). *Stratégie domaine par domaine pour la création d'un FrameNet du français : annotations en corpus de cadres et rôles sémantiques*. Theses, Université Sorbonne Paris Cité. HAL : tel-01661689.

DUBOIS J. & DUBOIS-CHARLIER F. (1997). *Les verbes français.* Larousse-Bordas, Paris, France.

FELLBAUM C. (1998). *WordNet : An Electronic Lexical Database*. Bradford Books.

FORT K., ADDA G. & BRETONNEL COHEN K. (2011). Amazon Mechanical Turk : Gold Mine or Coal Mine ? *Computational Linguistics*, p. 413–420. DOI : 10.1162/COLI_a_00057, HAL : hal-00569450.

FRANCOPOULO G., BEL N., GEORGE M., CALZOLARI N., MONACHINI M., PET M. & SORIA C. (2006). Lexical markup framework (LMF) for NLP multilingual resources. In *International Committee on Computational Linguistic and the Association for Computational Linguistics - COLING / ACL 2006*, Sydney/Australia : coling acl. HAL : inria-00121483.

GALA N. (2013). Ressources lexicales mono- et multilingues : une évolution historique au fil des pratiques et des usages. In *Ressources lexicales : contenu, évaluation, utilisation, évaluation.*, volume 30 de *Lingvisticæ Investigationes Supplementa*, p. 1–42. John Benjamins Publishing. HAL : hal-03203895.

GROSS M. (1975). *Méthodes en syntaxe*. Hermann, Paris, France.

GUILLAUME B., FORT K., PERRIER G. & BEDARIDE P. (2014). Mapping the Lexique des Verbes du Français (Lexicon of French Verbs) to a NLP Lexicon using Examples. In *International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland. HAL : hal-00969184.

GUREVYCH I., ECKLE-KOHLER J., HARTMANN S., MATUSCHEK M., MEYER C. M. & WIRTH C. (2012). UBY - a large-scale unified lexical-semantic resource based on LMF. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, p. 580–590, Avignon, France : Association for Computational Linguistics.

GUREVYCH I., ECKLE-KOHLER J. & MATUSCHEK M. (2016). Linked lexical knowledge bases : Foundations and applications. *Synthesis Lectures on Human Language Technologies*, **9**, 1–146. DOI : 10.2200/S00717ED1V01Y201605HLT034.

HATHOUT N. (2010). Morphonette : a morphological network of French. working paper or preprint.

HATHOUT N. & NAMER F. (2014). Démonette, a french derivational morpho-semantic network. *Linguistic Issues in Language Technology*, **11**(5), 125–168.

HATHOUT N., NAMER F. & DAL G. (2002). An Experimental Constructional Database : The MorTAL Project. In P. BOUCHER, Éd., *Many Morphologies*, p. 178–209. Somerville, Mass. : Cascadilla.

HENRICH V. & HINRICHS E. (2010). Standardizing wordnets in the ISO standard LMF : Wordnet-LMF for GermaNet. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, p. 456–464, Beijing, China : Coling 2010 Organizing Committee.

KIPPER-SCHULER K. (2005). *VerbNet : A broad-coverage, comprehensive verb lexicon*. Thèse de doctorat, University of Pennsylvania.

LAFOURCADE M. & JOUBERT A. (2008). JeuxDeMots : un prototype ludique pour l'émergence de relations entre termes. In *Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, Lyon, France.

LAFOURCADE M. & JOUBERT A. (2013). Bénéfices et limites de l'acquisition lexicale dans l'expérience jeuxdemots. In *Ressources lexicales : contenu, évaluation, utilisation, évaluation.*, volume 30 de *Lingvisticæ Investigationes Supplementa*, p. 187–216. John Benjamins Publishing.

LAFOURCADE M. & LE BRUN N. (2020). Jeuxdemots : Un réseau lexico-sémantique pour le français, issu de jeux et d'inférences. *Revue Lexique*, **27**, 47–86.

LEVIN B. (1993). *English Verb Classes and Alternations : A Preliminary Investigation*. University of Chicago Press.

LUX-POGODALLA V. & POLGUÈRE A. (2011). Construction of a French Lexical Network : Methodological Issues. In *First InternationalWorkshop on Lexical Resources, WoLeR 2011*, p. 54–61, Ljubljana, Slovenia. HAL : hal-00686467.

MATUSCHEK M. (2015). *Word sense alignment of lexical resources*. Thèse, Darmstadt University of Technology.

MCCRAE J., CEA G., BUITELAAR P., CIMIANO P., DECLERCK T., GOMEZ-PEREZ A., GRACIA J., HOLLINK L., MONTIEL-PONSODA E., SPOHR D. & WUNNER T. (2012). Interchanging lexical resources on the semantic web. *Language Resources and Evaluation*, **46**, 701–719.

MCCRAE J. P., GIL J. B., GRÀCIA J., BITELAAR P. & CIMIANO P. (2017). The ontolex-lemon model : Development and applications. In *Electronic lexicography in the 21st century : Lexicography from scratch. Proceedings of eLex 2017* : Electronic lexicography in the 21st century (eLex).

MERTENS P. (2010). Restrictions de sélection et réalisations syntagmatiques dans DICOVALENCE Conversion vers un format utilisable en TAL. In *Conference Traitement Automatique des Langues Naturelles (TALN)*, Montréal, Canada.

MOLINERO M. A., SAGOT B. & NICOLAS L. (2009). A morphological and syntactic wide-coverage lexicon for Spanish : The Leffe. In *RANLP 2009 - Recent Advances in Natural Language Processing*, Borovets, Bulgaria. HAL : inria-00616693.

NAMER F., BARQUE L., BONAMI O., HAAS P., HATHOUT N. & TRIBOUT D. (2019). Demonette2 - Une base de données dérivationnelle du français à grande échelle : premiers résultats. In E. MORIN, S. ROSSET & P. ZWEIGENBAUM, Éds., *Conférence sur le Traitement Automatique des Langues Naturelles (TALN) - PFIA*, p. 233–244, Toulouse, France : ATALA. HAL : hal-02567772.

NAVIGLI R. & PONZETTO S. P. (2012). Babelnet : The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.*, **193**, 217–250.

PERRIER G. & GUILLAUME B. (2013). Leopar : an Interaction Grammar Parser. ESSLLI 2013 - Workshop on High-level Methodologies for Grammar Engineering. HAL : hal-00920728.

PILEHVAR M. T. & NAVIGLI R. (2014). A robust approach to aligning heterogeneous lexical resources. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 468–478, Baltimore, Maryland : Association for Computational Linguistics. DOI : 10.3115/v1/P14-1044.

PROST J.-P. (2022). Integrating a Phrase Structure Corpus Grammar and a Lexical-Semantic Network : the HOLINET Knowledge Graph. In *Proceedings of LREC 2022 (to appear)*.

SAGOT B. (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *7th international conference on Language Resources and Evaluation (LREC 2010)*, La Valette, Malte. HAL : inria-00521242.

SAGOT B. & FIŠER D. (2008). Building a free French wordnet from multilingual resources. In *OntoLex*, Marrakech, Morocco.

SAGOT B., FORT K. & VENANT F. (2009). Extension et couplage de ressources syntaxiques et sémantiques sur les adverbes. *Lingvisticae Investigationes*, **32**(2), 305–315. DOI : 10.1075/li.32.2.12sag, HAL : hal-00446914.

SAJOUS F., HATHOUT N. & CALDERONE B. (2013). GLÁFF, un Gros Lexique Á tout Faire du Français. In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, p. 285–298, Les Sables d'Olonne, France.

SÉRASSET G. (2015). DBnary : Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF. *Semantic Web – Interoperability, Usability, Applicability*, **6**(4), 355–361. DOI : 10.3233/SW-140147, HAL : hal-00953638.

STEINBERGER R., POULIQUEN B., WIDIGER A., IGNAT C., ERJAVEC T., TUFIŞ D. & VARGA D. (2006). The JRC-Acquis : A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy.

TANGUY L. & HATHOUT N. (2002). Webaffix : un outil d'acquisition morphologique dérivationnelle à partir du web. In J.-M. PIERREL, Éd., *Actes de la 9eConférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN-2002)*, p. 245–254, Nancy : ATALA.

TCHECHMEDJIEV A. (2016). *Interopérabilité Sémantique Multi-lingue des Ressources Lexicales en Données Liées Ouvertes*. Thèse, Université Grenoble Alpes. HAL : tel-01681358.

TCHECHMEDJIEV A., MANDON T., LAFOURCADE M., LAURENT A. & TODOROV K. (2017). Ontolex JeuxDeMots and Its Alignment to the Linguistic Linked Open Data Cloud. In *ISWC : International Semantic Web Conference*, volume LNCS, p. 678–693, Vienne, Austria. DOI : 10.1007/978-3-319-68288-4_40, HAL : lirmm-01615473.

THOMASSET F. & VILLEMONTE DE LA CLERGERIE É. (2005). Comment obtenir plus des méta-grammaires. In *Actes de la 12ème conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*, p. 1–10, Dourdan, France : ATALA.

VAN DEN EYNDE K. & MERTENS P. (2006). Le dictionnaire de valence dicovalence : manuel d'utilisation.

VAN DEN EYNDE K. & MERTENS P. (2010). Le dictionnaire de valence dicovalence : manuel d'utilisation version 2.0.

VOSSEN P. (1998). *EuroWordNet : A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers.