REPRÉSENTATION DE L'INFORMATION SÉMANTIQUE LEXICALE : LE MODÈLE *WORDNET* ET SON APPLICATION AU FRANÇAIS

Benoît Sagot

Publications linguistiques | « Revue française de linguistique appliquée »

2017/1 Vol. XXII | pages 131 à 146

ISSN 1386-1204

Article disponible en ligne à l'adresse :

https://www.cairn.info/revue-francaise-de-linguistique-appliquee-2017-1-page-131.htm

Rev. franç. de linguistique appliquée, *2017, XXII-1 (131-146)*

Représentation de l'information sémantique lexicale : le modèle *wordnet* et son application au français

Benoît Sagot, Inria / Paris

Résumé : *Le modèle wordnet est le plus répandu des modèles de représentation de la sémantique lexicale reposant sur un inventaire de sens* a priori. *A la suite du* Princeton WordNet *de l'anglais, des ressources de type wordnet ont été développées pour plusieurs dizaines de langues, dont le français, le plus souvent au moyen de techniques automatiques ou semi-automatiques. Dans cet article, nous revenons tout d'abord sur les caractéristiques et les limites du modèle wordnet. Nous dressons ensuite un panorama des méthodes utilisées pour le développement de wordnets, avant d'illustrer nos propres travaux dans ce domaine par le développement du* WOLF*, le WOrdnet Libre du Français.*

Abstract: *The wordnet model is the most widespread model for representing lexical semantics based on an a priori inventory of meanings. Following the* Princeton WordNet *of English, wordnet-type resources have been developed for dozens of languages, including French, most often using automatic or semi-automatic techniques. In this article, we first review the characteristics and limitations of the wordnet model. We then present a panorama of wordnet development methods. We finally illustrate our own work in this area on the development of* WOLF*, the free French wordnet.*

Mots-clefs : Sémantique lexicale, wordnet, *Wordnet Libre du Français* (WOLF)
Keywords: Lexical semantics, wordnet, *Wordnet Libre du Français* (WOLF)

1. Contexte et motivations

Depuis une quinzaine d'années, le rôle des ressources lexicales de niveau sémantique ou encyclopédique s'est considérablement accru au sein du domaine du traitement automatique des langues, divers travaux montrant l'intérêt de telles ressources pour améliorer les performances pour divers types de tâches. Par exemple, Gabrilovich et Markovitch (2006) ont prouvé que l'utilisation de connaissances encyclopédiques améliore la classification automatique de documents. De même, Nastase (2008) a mis en œuvre de telles connaissances pour améliorer le résumé automatique. Harabagiu & *al.* (2000) ont obtenu des améliorations dans un système de réponse à des questions en tirant parti des liens entre mots dans un réseau lexical sémantique, dont l'intérêt a également été montré pour des tâches comme la désambiguïsation lexicale (Cuadros & Rigau 2006) ou la traduction automatique (Carpuat & Wu 2007).

Un certain nombre d'architectures ont été proposées pour organiser et représenter les connaissances lexicales sémantiques, telles qu'ACQUILEX (Copestake & *al.* 1994), le *Roget's Thesaurus* (Kirkpatrick 1987), *MindNet* (Richardson & *al.* 1998), *ConceptNet* (Liu 2003) ou *Cyc* (Matuszek & *al.* 2006).

D'autres types de ressources, et notamment celles produites par les projets de la famille *FrameNet* (Baker & *al.* 1998), s'intéressent plus spécifiquement à la valence sémantique des

prédicats, et relèvent ainsi plutôt de la sémantique prédicative, à mi-chemin entre les niveaux sémantique et syntaxique[1].

Mais l'une des ressources sémantiques lexicales les plus connues et les plus utilisées dans les domaines du traitement automatique des langues et du web sémantique est le *Princeton WordNet* (PWN) (Fellbaum 1998) et ses équivalents pour d'autres langues. Parmi ces derniers, on peut citer les wordnets développés dans le cadre des projets *EuroWordNet* (Vossen 1999), *BalkaNet* (Tufiş 2000) ou *AsianWordnet* (Sornlertlamvanich 2010), ainsi que le *Open Multilingual Wordnet* (Bond & Paik 2012), qui normalise et fusionne tous les wordnets dont la redistribution est autorisée par leurs auteurs, et inclut à ce jour des wordnets pour 27 langues. Initialement, le PWN était pourtant développé dans un contexte psycho-lexicographique (Miller 1995), inspiré par des travaux sur les processus cognitifs d'accès au lexique.

Dans un wordnet, les lexèmes sont organisés en ensembles de synonymes, ou synsets, chaque synset représentant un sens. Un synset a un identifiant unique et contient donc un certain nombre de littéraux, qui sont approximativement des lemmes (simples ou composés), des termes voire des collocations, qui tous peuvent exprimer le sens représenté par le synset. Les synsets sont reliés entre eux par des relations sémantiques, la plus structurante étant la relation d'hypéronymie. Parmi les autres relations incluses dans le PWN on peut citer les relations de méronymie, d'holonymie ou d'antonymie. Par exemple, dans la version 3.1 du PWN, le synset nominal d'identifiant 02086723-n contient les littéraux {*dog*, *domestic dog*, *Canis familiaris*}. Le sens ainsi représenté est illustré par une définition (*a member of the genus Canis […] that has been domesticated by man since prehistoric times; occurs in many breeds*) et un exemple d'emploi (*the dog barked all night*). Ce synset a deux hypéronymes, les synsets 02085998-n {*canine*, *canid*} 'canidé' et 01320032-n {*domestic animal*, *domesticated animal*}. Il a un certain nombre d'hyponymes, dont par exemple les synsets 02089774-n {*hunting dog*} et 02113929-n {*Newfoundland*, *Newfoundland dog*}.

Les premiers wordnets, et notamment le PWN, ont été développés manuellement, afin de maximiser la pertinence linguistique et de minimiser le taux d'erreur. Cependant, pour la grande majorité des langues, un tel effort est bien trop coûteux en temps et en moyens humains pour pouvoir être reproduit. C'est la raison pour laquelle diverses approches semi-automatiques et totalement automatiques ont été proposées pour le développement de wordnet à partir de divers types de ressources, et notamment en s'appuyant sur la disponibilité préalable du PWN. Ces approches diffèrent à plusieurs niveaux : équilibre recherché entre précision et couverture (plus la couverture visée est grande, plus le taux d'erreur dans la ressource produite est élevé), degré de complexité des ressources utilisées (depuis de 'simples' lexiques bilingues jusqu'à des thésaurus complexes, depuis des corpus monolingues bruts jusqu'à des corpus multilingues parallèles), degré de complexité des algorithmes employés, etc.

[1] *FrameNet* est un projet toujours actif, démarré en 1997, qui s'appuie sur la sémantique des *frames*, développée par Fillmore (1968). Dans cette théorie, les mots n'ont de sens qu'en référence à un espace conceptuel structuré qui, dans le projet *FrameNet*, est mis en œuvre par des liens entre les *frames*, chacun d'entre eux étant une structure conceptuelle qui décrit un type particulier de situation, d'objet ou d'événement, ainsi que ses participants (actants) et ses propriétés. Ces liens hiérarchisent l'ensemble des *frames* mais sont de différentes natures. Les participants aux *frames*, ou *frame elements*, sont déterminés (en principe) au moyen de critères purement sémantiques, sans référence aux cadres argumentaux des lexicalisations de ces *frames*. Le *FrameNet* de l'anglais a été développé avec une approche très lexicographique. Depuis, d'autres *FrameNet* ont vu le jour, notamment pour l'espagnol, l'allemand, le suédois et le français. Dans la majorité des cas, une approche plus équilibrée entre annotation de corpus et développement du lexique a été employée.

Dans cet article, nous présentons la méthodologie générale de développement de wordnets que nous avons mise en place et appliquée au développement de wordnets pour deux langues : le français et le slovène. Il s'agit donc d'une méthodologie indépendante de la langue, mais dépendante de la disponibilité de ressources telles que celles évoquées ci-dessus. Nous ne ferons toutefois état ici que de son application au français pour le développement du WOLF (*WOrdnet Libre du Français*), ressource librement disponible.

Avant de dresser un panorama des techniques de développement de wordnets utilisées par d'autres auteurs, il est peut-être nécessaire de justifier la pertinence du développement de nouveaux wordnets pour le français. Au moment du démarrage des travaux rapportés dans cet article (début 2008), le seul wordnet qui existait pour le français avait été développé dans le cadre du projet *EuroWordNet* (Vossen 1999). Dans la suite de cet article, nous dénoterons ce wordnet par le terme *French WordNet* (FWN). L'utilisation de cette ressource n'a jamais été très répandue, principalement pour des raisons liées à la licence qui lui était associée. De plus, il n'y a pas eu de suite en France au projet *EuroWordNet*, qui aurait pu travailler à l'extension et l'amélioration de cette ressource restreinte à un sous-ensemble des noms et des verbes, à l'exclusion des adjectifs et des adverbes (Jacquin & *al.* 2007). Depuis la création du WOLF, un autre wordnet a été développé en parallèle au moyen de ressources bilingues librement disponibles extraites de ressources *wiki* ainsi que d'un modèle de langue syntaxique du français. La première version, limitée aux synsets nominaux, est distribuée sous le nom de JAWS (Mouton & de Chalendar 2010). Une version ultérieure, obtenue grâce à une version améliorée de la méthode et couvrant toutes les parties du discours, est distribuée sous le nom de *WoNeF* (Pradet & *al.* 2014) et évaluée avec soin. Nous comparerons donc le WOLF avec le FWN, JAWS et WoNeF. On pourra noter que le WOLF est intégré à la plateforme *Open Multilingual Wordnet* (Bond & Paik 2012) et constitue par ce biais l'une des sources du wordnet multilingue *BabelNet* (Navigli & Ponzetto 2012).

2. Le développement automatique de wordnets : état de l'art

Les techniques automatiques de développement de wordnets se répartissent selon celle des deux approches principales qu'elles mettent en œuvre : l'approche par fusion et l'approche par extension (Vossen 1999). Dans le cas de l'approche par fusion, un wordnet pour une langue donnée est créé indépendamment des autres wordnets existants, en exploitant au mieux des ressources monolingues disponibles ; dans un second temps, le wordnet ainsi créé peut être aligné avec les wordnets disponibles pour d'autres langues (Rudnicka & *al.* 2012). Dans le cas de l'approche par extension, que nous avons utilisée, l'inventaire de sens du PWN est conservé (mêmes identifiants de synsets, mêmes relations entre synsets) et on cherche à peupler les synsets avec des littéraux de la langue cible, par exemple par désambiguïsation et traduction des littéraux anglais présents dans le PWN.

L'approche par extension repose donc sur l'approximation selon laquelle les concepts (les sens) et les relations entre eux sont indépendants de la langue, au moins pour une bonne part. C'est du reste la principale limite de cette approche : les wordnets produits sont biaisés par rapport au PWN, ce qui peut même, dans certains cas, rendre certains synsets ou certaines relations arbitraires (Orav & Vider 2004 ; Wong 2004). Par exemple, le PWN contient un synset {*performer, performing artist*}, défini comme étant un artiste réalisant un spectacle théâtral ou musical face à une audience. Mais il n'y a pas de mot en français qui dénote de façon globale les acteurs, chanteurs et autres artistes se produisant en spectacle. Dans un tel cas, il est toujours possible de laisser vide le synset en question dans la ressource produite. A l'inverse, certains sens raisonnablement répandus de la langue cible peuvent ne pas correspondre à un synset du PWN, par exemple parce qu'ils n'ont pas vraiment de réalité culturelle dans les pays de langue anglaise ou ne sont pas considérés comme suffisamment

importants. C'est ainsi le cas des sens ou concepts dénotés en français par {*raclette*}, {*Jacques Chirac*} ou {*Ecole Polytechnique*}. Parfois, c'est le découpage même en synsets qui ne correspond pas bien. Ainsi, les synsets {*lawyer*, *attorney*} (*a professional person authorized to practice law; conducts lawsuits or gives legal advice*) et {*advocate, counsel, counselor, counsellor, counselor-at-law, pleader*} (*a lawyer who pleads cases in court*) sont distingués selon des critères propres au système judiciaire anglo-saxon, voire américain, qui ne se superposent pas avec les distinctions françaises entre 'juriste', 'avocat' ou 'avoué'.

Malgré tout, ces problèmes sont plus que compensés par les avantages importants de l'approche par extension, qui est ainsi très utilisée pour le développement de wordnets, par exemple dans les projets *BalkaNet* (Tufiş 2000), *MultiWordnet* (Pianta & *al.* 2002) et *BabelNet* (Navigli & Ponzetto 2010, 2012). Le premier avantage est naturellement un coût de développement bien plus bas que pour l'approche par fusion. Le second avantage est que les wordnets produits sont alignés sur le PWN et donc également alignés entre eux, ce qui permet d'envisager leur utilisation dans des applications multilingues, telles que la traduction automatique ou l'extraction d'informations. Nous ne prétendons pas que l'approche par extension soit meilleure que l'approche par fusion, mais nous pensons qu'il s'agit du meilleur choix dans un contexte où, comme dans le nôtre, on souhaite développer à moindre coût une ressource à large couverture et de précision suffisante pour la majorité des applications possibles.

Les mises en œuvre de l'approche par extension varient selon le type de ressources qui sont disponibles pour la construction d'un wordnet dans une langue donnée. Les premiers travaux en ce sens utilisaient directement des dictionnaires électroniques bilingues, en essayant d'aligner les entrées des dictionnaires avec wordnet (Knight & Luk 1994 ; Yokoi 1995). Le problème qui apparaît immédiatement est la difficulté de la désambiguïsation des (sous-) entrées des dictionnaires bilingues. De plus, les dictionnaires bilingues ont souvent une couverture limitée et ne sont pas nécessairement disponibles pour toutes les paires de langues.

Une façon de surmonter ces difficultés est d'utiliser des lexiques bilingues ou multilingues extraits de corpus parallèles (Resnik & Yarowsky 1997 ; Fung 1995). L'hypothèse principale qui sous-tend ces travaux est que les différents sens d'un même mot qui est ambigu dans une langue sont souvent traduits par des mots différents dans une autre langue. De plus, si deux mots distincts, voire plus, sont traduits par le même mot dans une autre langue, ils sont souvent sémantiquement reliés, voire synonymes. Ceci permet de désambiguïser les mots polysémiques ou à l'inverse de créer des liens synonymiques. Les corpus parallèles ont été utilisés pour induire des synsets pour une nouvelle langue par divers auteurs (Dyvik 2002 ; Ide & *al.* 2002 ; Diab 2004).

La troisième famille de travaux, plus récente, cherche à exploiter au mieux les ressources libres et collaboratives telles que *Wikipedia*. *Wikipedia* est une ressource encyclopédique libre disponible dans de nombreuses langues. Chaque article peut notamment être muni de catégories et être relié à des articles qui lui correspondent dans les *Wikipedia* d'autres langues. De nouveaux wordnets ont été induits en associant des pages de *Wikipedia* avec des synsets de wordnet (Suchanek & *al.* 2008), en utilisant des informations structurelles pour associer des catégories *Wikipedia* aux synsets (Ponzetto & Navigli 2009) ou en extrayant des mots-clés à partir des articles de *Wikipedia* (Reiter & *al.* 2008). Un modèle vectoriel permettant d'associer des pages *Wikipedia* à wordnet a été proposé par divers auteurs (Ruiz-Casado & *al.* 2005 ; Declerck & *al.* 2006). Les approches les plus abouties utilisent *Wikipedia* et d'autres ressources wiki, notamment *Wiktionary* (version française : le *Wiktionnaire*) pour produire des wordnets dans de multiples langues (de Melo & Weikum 2009 ; Navigli & Ponzetto 2012).

Au cours du développement du WOLF, nous avons cherché à tirer le meilleur parti des ressources librement disponibles, et notamment de corpus parallèles et de ressources de type wiki (*Wikipedia* et *Wiktionary*/*Wiktionnaire*), pour 'traduire' le PWN en français. Au cours de la première étape de développement, seuls les littéraux anglais monosémiques[2] ont été traduits au moyen de lexiques bilingues extraits de ressources wiki, et les lexèmes polysémiques ont été désambiguïsés et traduits grâce aux corpus parallèles. La seconde étape que nous décrirons ensuite a eu pour objectif d'étendre de façon significative la couverture du WOLF, en s'appuyant sur la première version désormais disponible pour contribuer à la désambiguïsation des lexiques bilingues extraits de ressources wiki, afin de pouvoir les utiliser également sur les lexèmes polysémiques. Ces deux premières étapes ne reposaient donc que sur la traduction du PWN, sans quasiment aucune information contextuelle sur les occurrences en corpus des mots concernés, sauf pour l'exploitation de corpus multilingues parallèles lors de la première étape. C'est principalement lors de la troisième étape majeure du développement de ces ressources que nous avons exploité ce type d'informations, afin d'identifier les intrus : l'analyse des similarités, ou plutôt des dissimilarités, entre les contextes d'apparition et les synsets 'proches' des synsets concernés nous a permis d'identifier des candidats-erreurs et d'augmenter le taux de précision des ressources. Nous allons désormais décrire rapidement ces différentes techniques, en suivant en parallèle leur utilisation pour le développement des versions successives du WOLF.

3. Approches pour le développement automatique et semi-automatique de wordnets : le cas du WOLF

La plupart des méthodes de développement de wordnets relevant de l'approche par extension peuvent être décomposées en deux étapes :

1. Une première étape consistant à construire, si l'on n'en dispose pas déjà, un lexique bilingue ou multilingue – l'utilité de disposer de lexiques impliquant d'autres langues en plus de la langue source (ici, l'anglais) et de la langue cible (ici, le français) trouvera sa justification ci-dessous.
2. Une deuxième étape consistant à rapprocher les entrées de ces lexiques bilingues ou multilingues, dont l'une des langues est la langue source (ici, l'anglais), avec les synsets du wordnet source (ici, le PWN), afin de construire le plus grand nombre possible de couples *(littéral de la langue cible, synset)* candidats ; l'idéal est de disposer d'une mesure de confiance sur ces candidats afin de pouvoir sélectionner les meilleurs d'entre eux et éliminer les autres.

3.1. Construction de lexiques bilingues ou multilingues

On peut classer en trois types les ressources desquelles il est possible d'extraire des lexiques bilingues ou multilingues : les ressources structurées (dictionnaires et lexiques généraux ou spécialisés), les ressources semi-structurées (articles *Wikipedia*) et les ressources non structurées (corpus alignés). Nous avons extrait des entrées bilingues ou multilingues de tous ces types de ressources.

[2] Par abus de langage, nous qualifions de 'monosémique' un littéral anglais dès lors qu'il n'apparaît que dans un seul synset du PWN. Il s'agit naturellement d'une approximation, qui repose sur l'hypothèse selon laquelle le PWN est complet. Cette approximation se justifie d'une part par la grande couverture de cette ressource et d'autre part par le fait qu'un littéral anglais qui est en réalité polysémique, s'il n'apparaît que dans un seul synset du PWN, a de fortes chances d'avoir ce sens unique pour sens très majoritaire : ainsi, il est plausible que les traductions que nous pourrions en trouver en français concernent toutes, ou presque toutes, ce sens majoritaire.

Les dictionnaires bilingues ou multilingues sont des ressources lexicales sémantiques souvent très précises, parfois de bonne couverture, et qui couvrent toutes les parties du discours. Ils proposent souvent des traductions pour chacun des sens des mots polysémiques. Toutefois, ils ne contiennent généralement que des informations non-contextualisées, qui ne permettent pas, lors de la deuxième étape, une mise en correspondance aisée avec les synsets du PWN. Parmi eux, les ressources de type *Wiktionary* (le *Wiktionary* anglais et le *Wiktionnaire* français, notamment) sont des dictionnaires collaboratifs disponibles dans de nombreuses langues, qui, dans le meilleur des cas, associent à un mot un ou plusieurs sens, et, pour chacun d'entre eux, une définition et des traductions dans d'autres langues. Ces ressources sont librement disponibles. Nous avons extrait du *Wiktionary* anglais et du *Wiktionnaire* français respectivement 62 826 et 59 659 entrées bilingues anglais-français. Nous avons également extrait des entrées bilingues de *Wikispecies*, une taxonomie collaborative des espèces vivantes qui inclut à la fois les noms latins standard et, pour les espèces les plus communes, leurs noms vernaculaires. Ceci nous a permis d'extraire 48 046 entrées bilingues français-anglais.

Les articles de l'encyclopédie collaborative en ligne *Wikipedia*, disponible pour de nombreuses langues, correspondent à un niveau de structuration intermédiaire entre dictionnaires électroniques et corpus. Nous avons extrait des entrées bilingues à partir des *Wikipedia* française, slovène et anglaise en considérant comme un littéral nominal le titre de chaque article puis en suivant les liens inter-wiki qui relient un article d'une *Wikipedia* à un autre, portant sur le même sujet ou concept, dans une autre *Wikipedia*. Le résultat est un ensemble de 286 822 couples *(littéral anglais, littéral français)* qui sont tous considérés comme nominaux (noms propres ou noms communs, qui ne sont pas différenciés dans le PWN).

Le dernier type de ressources que nous avons utilisées dans nos expériences est non-structuré. Il s'agit de corpus bruts, mais qui sont alignés : ces textes ont été traduits (manuellement) dans différentes langues, et les liens entre équivalents de traductions sont disponibles au niveau des documents. Nous avons utilisé le corpus SEE-ERA.NET, un sous-corpus d'environ 1,5 million de mots du JRC-Acquis (Tufiş 2009) disponible en 8 langues dont l'anglais et le français, mais également le slovène, langue sur laquelle nous travaillions en parallèle, pour le développement du wordnet *sloWNet*. Outre ces trois langues, nous avons également utilisé les données en roumain, tchèque et bulgare, qui sont les langues pour lesquelles nous disposions de wordnets alignés avec le PWN, tous issus du projet *BalkaNet* mentionné précédemment. Nous avons utilisé différents outils pour étiqueter morpho-syntaxiquement et lemmatiser ces textes, avant de les aligner au niveau des phrases puis des mots à l'aide de l'outil *Uplug* (Tiedemann 2003). Après un seuillage sur les mesures de confiance et les nombres d'occurrences de chaque lien d'alignement, nous avons extrait à partir des corpus alignés multilingues un ensemble de lexiques multilingues (par exemple en extrayant les équivalents de traduction dans toutes les autres langues de chaque occurrence de chaque mot français)[3]. Le résultat est un ensemble de cinq lexiques multilingues qui contiennent entre 49 356 entrées (lexique français-roumain-tchèque-bulgare-anglais) et 59 019 entrées (français-tchèque-bulgare-anglais). Le tableau 1 illustre ce processus en montrant certaines entrées du lexique français-tchèque-bulgare-anglais. On peut constater que ces entrées ne sont pas toutes correctes. Les erreurs qui apparaissent peuvent avoir différentes origines : erreurs d'étiquetage morphosyntaxique, erreur de lemmatisation, erreurs d'alignement. Cependant, la plupart de ces erreurs sont éliminées à l'étape suivante de désambiguïsation.

[3] L'alignement ayant été réalisé au niveau des mots simples, on notera que nous n'avons pas pu, par cette méthode, extraire d'entrées bilingues comportant des mots ou des termes composés.

Tableau 1. *Extrait du lexique français-tchèque-bulgare-anglais tiré du corpus parallèle*.

Fréquence	*Partie du discours*	*Français*	*Tchèque*	*Bulgare*	*Anglais*
18	n	droit	právo	законодателство	law
56	n	droit	právo	право	law
4	n	loi	právo	закон	law
4	n	loi	právo	законодателство	law
6	n	loi	právo	право	law
33	n	loi	zákon	закон	law
8	n	loi	zákon	закона	law
19	n	législation	právo	законодателство	law
7	n	législation	právo	право	law
4	n	législation	předpis	законодателство	law

Lors d'un travail ultérieur (Hanoka & Sagot 2012), nous avons extrait un ensemble de paires de littéraux en relation de traduction à partir d'un grand nombre de ressources de types variés, y compris une quinzaine d'éditions du *Wiktionary* (dont le *Wiktionary* anglais et le *Wiktionnaire* français), les corpus OPUS, les lexiques bilingues du projet *Apertium* et plusieurs wordnets librement disponibles. Ces entrées bilingues, couvrant un millier de langues, est représenté sous la forme d'un graphe de traductions. Ce graphe, nommé YaMTG et librement disponible (Hanoka & Sagot 2014), a servi dans sa version 1.0 lors d'une étape d'extension du WOLF non décrite dans cet article (on pourra se reporter à Hanoka (2015)).

3.2. Désambiguïsation des entrées bilingues ou multilingues

La création ou l'extension d'un wordnet au moyen d'une approche qui préserve le même inventaire de synsets que le PWN peut être vue comme une tâche consistant à produire des couples *(littéral, synset)*, que nous appelons des *candidats*. Cet objectif peut être atteint pour la construction d'un wordnet français en attribuant un synset à chaque entrée d'un lexique bilingue anglais–français tel que ceux extraits précédemment. Mais la situation est bien différente selon les cas. Les lexiques multilingues extraits de corpus alignés sont les seuls à être issus d'occurrences en contexte, lequel contexte est partiellement représenté par les langues tierces, c'est-à-dire autres que la langue cible et la langue source (cf. section 3.2.1.). A l'inverse, dans le cas des littéraux monosémiques, la désambiguïsation (l'attribution au candidat d'un synset du PWN) est triviale. Le cas des candidats extraits de dictionnaires et contenant un lexème polysémique est plus complexe (cf. section 3.2.3.).

3.2.1. Désambiguïsation des entrées multilingues issues de corpus alignés

La désambiguïsation des entrées multilingues telles qu'illustrées au tableau 1 peut être réalisée par intersection sémantique, de la façon suivante (Sagot & Fišer 2008). Les langues sélectionnées, outre le français (ou le slovène), sont celles pour lesquelles nous disposions de wordnets, issus du projet *BalkaNet* (Tufiş 2000), alignés sur le PWN. Pour chaque entrée multilingue, nous avons extrait de ces wordnets l'ensemble des synsets associés à chaque littéral de chaque langue concernée (cf. tableau 2). On peut alors considérer, si l'entrée est valide, que le sens véhiculé par les différents littéraux qui composent cette entrée, lorsqu'ils

sont en lien de traduction, est représenté par l'intersection des synsets contenant chacun des littéraux. Sur l'exemple du tableau 2, il n'y a qu'un seul synset qui est commun au tchèque *právo*, au bulgare *право* et à l'anglais *law* : ce synset représente donc le sens véhiculé par ces trois mots ainsi que par le mot français *droit* dans toutes les occurrences où ils sont en relation de traduction quadrilingue. On peut donc proposer un candidat associant le littéral français *droit* à cet unique synset. Dans ce cas, le candidat proposé est correct, puisqu'il associe *droit* à un synset dont la glose dans le PWN est '*the branch of philosophy concerned with the law and the principles that lead courts to make the decisions they do*'.

Tableau 2. *Illustration du processus de désambiguïsation des entrées multilingues extraites de corpus alignés*.
(Les identifiants de synsets proviennent de la version 2.0 du PWN, tout comme les versions d'origine des wordnets *BalkaNet*).

Tchèque	**Bulgare**	**Anglais**	**Français**
právo	право	law	droit
6129345-n	4893549-n	577416-n	→ 5791721-n
5559593-n	4888072-n	5529208-n	
5791721-n	7928837-n	5531141-n	
4617988-n	577416-n	5791721-n	
7928837-n	5791721-n	6129345-n	
	1000872-n	7712371-n	
	4881053-n	7928837-n	
	4617988-n		

Nous avons ainsi produit plusieurs milliers de candidats à partir de chacun des cinq lexiques multilingues. Pour le français, nous avons ainsi construit entre 1 338 candidats (à partir du lexique français-roumain-tchèque-bulgare) et 5 073 candidats (à partir du lexique français-roumain-anglais). Naturellement, les candidats proposés uniquement à partir d'un seul lexique trilingue sont moins sûrs que ceux proposés par plusieurs lexiques, dont le lexique complet impliquant cinq langues. Des erreurs sont à attendre dans tous les cas, qui sont la conséquence des erreurs mentionnées plus haut (étiquetage morphosyntaxique, lemmatisation, alignement) mais également des ambiguïtés résiduelles (par exemple si l'intersection entre ensembles de synsets ne donne pas un seul synset, comme dans nos exemples, mais deux).

3.2.2. Désambiguïsation des entrées issues de dictionnaires : le cas des entrées monosémiques et la construction de la première version du WOLF

Les entrées bilingues obtenues à partir de ressources lexicales (*Wiktionary*, *Wikipedia*) ne sont pas contextualisées, mais sont de première importance car elles sont bien plus nombreuses et précises que celles extraites de corpus alignés. Comme indiqué ci-dessus les candidats obtenus *via* un littéral anglais monosémique peuvent être trivialement associés à un synset du PWN. Par exemple, le littéral anglais *battle of Gettysburg* est monosémique, puisqu'il n'apparaît que dans un seul synset, le synset 01282108-n. Or nous avons extrait de *Wikipedia* la traduction française *bataille de Gettysburg* pour ce littéral. Nous pouvons donc produire le candidat français (*bataille de Gettysburg*, 01282108-n). C'est en combinant ces candidats avec ceux construits à partir de corpus alignés selon la méthode décrite ci-dessus qu'a été

construite la première version du WOLF (version 0.1.4 (Sagot et Fišer 2008)). Cette version contient 32 251 synsets non vides (c'est-à-dire contenant au moins un littéral français), à comparer aux 22 121 synsets non vides obtenus en projetant sur le même inventaire de synsets le wordnet français du projet *EuroWordNet* (Vossen 1999), qui de plus, contrairement au WOLF, ne contient que des synsets verbaux et nominaux.

3.2.3. Désambiguïsation des entrées issues de dictionnaires : le cas des entrées polysémiques et l'extension initiale du WOLF

Le second ensemble de candidats est bien plus bruité : il s'agit des candidats que l'on obtient à partir des littéraux anglais polysémiques en associant toutes leurs traductions disponibles à tous leurs synsets. Par exemple, le littéral anglais *dog* appartient à huit synsets différents. Le lien de traduction anglais-français (*dog*, *chien*) donnera donc lieu à autant de candidats français impliquant le littéral français *chien*, dont certains sont erronés.

Nous avons donc mis en place une technique de désambiguïsation des candidats (Sagot & Fišer 2011, 2012a, 2014), technique qui repose toutefois sur l'existence d'un wordnet dans la langue cible. En conséquence, cette méthode n'a pas été appliquée pour le développement des premières versions du WOLF (versions 0.1.x), mais a été utilisée pour le développement de sa version 0.2 (cf. ci-dessous). L'idée est d'entraîner un classifieur qui utilise différents traits associés à chaque candidat *(littéral, synset)* pour tenter de déterminer si le candidat est valide ou non. Parmi ces traits, nous avons fait usage de la similarité distributionnelle[4] entre le littéral et le synset formant le candidat, la distribution du synset étant approchée par celle des littéraux qu'il contient déjà ainsi que ceux qui sont contenus dans les synsets qui lui sont proches[5]. Pour plus de détails sur la notion de synset proche et sur les autres types de traits utilisés, on pourra se reporter à (Fišer & Sagot 2015). Nous donnons au classifieur – en l'espèce, le classifieur à maximum d'entropie *megam* (Daumé III 2004) – les données d'entraînement suivantes : parmi tous les candidats que nous avons construits, tous ceux qui sont déjà présents dans la version alors disponible du WOLF sont considérés comme des exemples positifs, et tous les autres candidats comme des exemples négatifs. C'est là une double approximation : d'une part, les candidats déjà présents dans le WOLF ne sont pas tous valides, d'autre part, les candidats qui n'y sont pas encore ne sont pas tous erronés – en réalité, notre objectif est ici précisément d'identifier parmi eux ceux qui sont corrects. Nous examinons alors les résultats du classifieur sur ses propres données d'entraînement : chaque candidat reçoit alors un score compris entre 0 (candidat certainement erroné) et 1 (candidat certainement valide). Nous avons fixé empiriquement un seuil identique de 0,1 : tout candidat dont le score est supérieur ou égal à ce seuil est conservé. Ainsi, la similarité sémantique entre *chien* et le synset {*andiron, firedog, dog, dog-iron*} n'est que de 0,035, alors qu'elle est de 0,331 avec le synset {*dog, domestic dog, Canis familiaris*}. Malgré l'utilisation hétérodoxe du concept de classifieur dans cette approche, les résultats que nous avons obtenus en montrent la validité, à la fois quantitativement et qualitativement, lors de son application pour l'extension du WOLF alors en version 0.6[6].

[4] Telle que mesurée par le programme *SemanticVectors* (Widdows & Ferraro 2008).

[5] Par exemple, le synset {*andiron, firedog, dog, dog-iron*} du PWN, qui est vide dans la version initiale du WOLF, est représenté par un sac de mots dont voici un extrait : *appareil, mécanisme, barre, rayon, support, balustre, dispositif*.

[6] La différence entre la version 0.4 mentionnée ci-dessus et la version 0.6 mentionnée ici a consisté en un travail spécifique sur les synsets adverbiaux, travail qui a tiré parti de relations de morphologie dérivationnelle et de la base de synonymes *DicoSyn* (Sagot & *al.* 2009). Son impact quantitatif sur la ressource est limité, puisqu'il ne concerne que les synsets adverbiaux.

En effet, sur le plan quantitatif, nous avons alors construit 177 980 candidats *(littéral, synset)* à partir de nos lexiques bilingues parmi lesquels la méthode a permis de retenir 55 159 candidats, dont 15 313 (28 %) étaient déjà présents dans le WOLF 0.6. Autrement dit, la méthode a proposé 39 823 nouveaux candidats. Ces candidats ont permis d'attribuer au moins un littéral à 13 899 synsets qui étaient précédemment vides, d'où 43 % de synsets non vides en plus et 65 % de couples *(littéral, synset)* en plus.

Sur le plan qualitatif, une évaluation manuelle de 400 des 177 980 candidats proposés pour WOLF, choisis aléatoirement, montre une précision moyenne de 52 %, qui monte à 81 % si l'on se restreint aux candidats faisant partie des 55 159 candidats retenus. A l'inverse, la précision parmi les candidats éliminés tombe à 40 %. On note même que les trois derniers quartiles (si l'on classe les candidats en fonction du score que leur attribue le classifieur), qui correspondent à peu près aux candidats rejetés, correspondent à des précisions moyennes en chute libre : successivement 63 %, 41 % et 20 %, ce qui montre la pertinence des scores associés aux candidats.

Pour une évaluation globale de la ressource ainsi obtenue, le WOLF 0.2, on pourra se reporter à la section 5.

4. Identification automatique et suppression manuelle des entrées erronées

Malgré les bons résultats obtenus à ce stade, les techniques utilisées, comme toutes les techniques de population automatique de wordnets, n'est évidemment pas parfaite et produit des erreurs. De plus, d'autres techniques ont permis d'étendre encore la version 0.2 du WOLF, notamment par l'exploitation de relations de dérivation morpho-sémantique (Gábor & *al.* 2012 ; Apidianaki & Sagot 2014) et l'exploitation du graphe de traduction fortement multilingue *YaMTG* mentionné précédemment (Hanoka & Sagot 2012, 2014 ; Hanoka 2015).

Des efforts de validation manuelle ont été engagés. Une validation manuelle partielle des synsets verbaux BCS, et notamment de la quasi-totalité des BCS 1 verbaux, a été réalisée (correction et complétion) : 825 couples *(littéral, synset)* ont été ainsi ajoutés, 4 933 ont été supprimés et 5 204 ont été confirmés. De plus, des techniques d'identification d'erreurs plus simples que celles décrites dans cette section ont été appliquées, afin de vérifier à partir du lexique morphologique et syntaxique Le*fff* (Sagot 2010) que les littéraux connus du Le*fff* l'étaient avec la partie du discours correspondant à leur synset. Une validation manuelle des littéraux ainsi identifiés comme suspects a permis de supprimer 4 155 couples *(littéral, synset)* incorrects.

Il restait néanmoins dans le WOLF un nombre significatif de couples *(littéral, synset)* erronés. Nous avons donc mis en place une technique d'identification d''intrus' dans les synsets, technique indépendante de la langue qui repose sur l'utilisation de corpus afin de désambiguïser au mieux les littéraux dans leurs contextes d'apparition (Sagot & Fišer 2012b). Il s'agit en réalité d'une adaptation et amélioration de la méthode décrite ci-dessus pour la désambiguïsation d'entrées bilingues polysémiques. En effet, nous cherchons ici à évaluer non pas des couples *(littéral, synset)* candidats à l'ajout dans le wordnet, mais des couples *(littéral, synset)* qui y sont déjà, afin de détecter parmi eux des 'intrus'. Mais l'enjeu sous-jacent reste le même : pouvoir associer à des couples *(littéral, synset)* un score qui permette de distinguer les couples corrects des couples incorrects.

L'amélioration apportée ici consiste à prendre en compte la polysémie de façon plus complète dans notre façon d'exploiter le corpus de façon distributionnelle. En effet, nous ne nous contentons pas de calculer la distribution globale de chaque littéral, mais nous partons de chacune de ses occurrences. Plus précisément, cette approche, que nous n'avons appliquée à ce jour qu'aux synsets nominaux, se décompose en trois étapes :

1. Pour chaque couple *(littéral, synset)* du wordnet à nettoyer dont le littéral est attesté dans un corpus (monolingue) de taille importante, évaluation de la similarité sémantique entre chaque occurrence du littéral dans ce corpus, représentée par les mots constituant son contexte, et le synset, représenté par les littéraux de ce synset et de ses synsets proches[7].

2. Calcul d'un score global pour chaque couple *(littéral, synset)* à partir des similarités sémantiques locales calculées précédemment pour chacune des occurrences du littéral.

3. Identification des candidats intrus, c'est-à-dire des couples *(littéral, synset)* dont le score global ainsi calculé est en dessous d'un certain seuil. Le tableau 3 indique quelques exemples de candidats ainsi extraits.

Tableau 3. *Exemples de candidats intrus trouvés dans la version 0.2 du* WOLF *par la méthode décrite ici. Lorsque l'évaluation manuelle est OK, cela indique qu'il s'agit bien d'un intrus, autrement dit, que le couple (*littéral, synset*) est incorrect*.

littéral	*synset (PWN 3.0)*	*littéraux anglais du synset*	*score (×10³)*	*évaluation manuelle*
abord	8307589	meeting, group meeting	0,013	OK
activité	14006945	activeness, action, activity	0,014	NO
activité	5833022	business	0,011	OK
adresse	35189	achievement, accomplishment	0,017	OK
agence	3015254	chest, chest of drawers, bureau, dresser	0,015	OK
besogne	6545137	deed of conveyance, title	0,012	OK
bout	8566028	terminal, end	0,019	NO
bureau	13945102	office, power	0,006	OK
cadre	10069645	executive director, executive	0,017	OK
cadre	10014939	managing director, manager, director	0,014	OK

Enfin, l'objectif étant une amélioration significative de la ressource, nous avons procédé à un tri manuel entre candidats intrus pour que soient retirés du wordnet les couples *(littéral, synset)* effectivement erronés. Nous avons pour cela utilisé l'outil *sloWCrow*d, outil d'annotation collaborative dédié développé par Tavčar & *al.* (2012). Grâce à un jeu de référence de 100 candidats annotés manuellement au préalable, les annotateurs volontaires qui ont participé à cette tâche ont pu être évalués. Nous avons éliminé les annotations, en réalité peu nombreuses, effectuées par des utilisateurs dont la précision était trop basse. *In fine*, 9 validateurs sur les 17 ayant validé au moins un synset ont été retenus, 7 des 8 non retenus étant des utilisateurs factices opérant des tests ou des démonstrations, le dernier étant un utilisateur dont le taux de précision était seulement de 75 %. Les utilisateurs dont les validations ont été retenues ont à ce jour annoté de 101 à 1 644 couples, avec une précision évaluée entre 80 % à 96 %, pour un total de 6 400 validations individuelles permettant de confirmer ou d'éliminer 1 540 couples *(littéral, synset)* distincts. L'effort doit donc être poursuivi avant que la version 1.0 de WOLF, qui intègrera les résultats de cette campagne, ne soit distribuée.

[7] Au sens déjà évoqué précédemment et tel que décrit dans Sagot & Fišer (2012a), Fišer & Sagot (2015).

5. Données quantitatives et évaluation du WOLF

5.1. Données quantitatives

Le tableau 4 comporte des données quantitatives sur trois versions successives du WOLF et sur d'autres ressources comparables. Les catégories BCS (pour *Basic Concept Sets*) reflètent une classification des synsets développée dans le cadre du projet *BalkaNet* qui va des synsets de base (BCS1) aux synsets les moins importants (Hors BCS) – la notion de fréquence est prise en compte dans cette classification mais de façon non exclusive.

Tableau 4. *Données quantitatives (nombre de synsets non vides) sur la version initiale du* WOLF *(version 0.1.4), obtenue après la phase d'extension (version 0.2) et sa version la plus récente (version 1.0b4). Sont également indiquées les données correspondantes pour le* PWN*, le wordnet français développé dans le cadre du projet* EuroWordNe*t (colonne* FWN*), le wordnet nominal* JAWS *et son successeur* WoNeF.

	PWN	*WOLF*			*FWN*	*JAWS*	*WoNeF*
	2.0	0.1.4	0.2	1.0b4			
Total	115 424	32 351	46 449	56 479	22 121	34 367	53 442
BCS1	1 218	869	1 067	1 107	1 211	760	816
BCS2	3 471	1 665	2 519	2 900	3 022	1 729	2 097
BCS3	3 827	1 796	2 585	2 963	2 304	1 706	1 906
Hors BCS	106 908	27 492	40 278	49 509	15 584	30 172	48 623
N	79 689	28 187	36 933	42 427	17 381	34 367	37 355
V	13 508	1 546	4 105	5 870	4 740	0	3 845
Adj	18 563	1 422	4 282	6 691	0	0	10 238
Adv	3 664	667	1 125	1 487	0	0	2 002

5.2. Evaluation de la qualité du WOLF

L'évaluation de ressources lexicales est toujours une tâche délicate, et cela s'applique à un wordnet comme le WOLF. De façon générale, on peut identifier trois grands paradigmes d'évaluation des ressources lexicales : évaluation comparative par rapport à d'autres ressources comparables, évaluation manuelle de la précision, voire de la couverture, et évaluation orientée-tâche. A ce jour, le WOLF n'a pas encore fait l'objet d'une évaluation orientée-tâche. Nous proposons donc ici une évaluation comparative par rapport à des ressources de référence, jointe à une évaluation manuelle de couples *(littéral, synset)* choisis aléatoirement, dont nous avons déduit, ensemble, une estimation de la qualité générale du WOLF. En effet, une simple évaluation par rapport à d'autres ressources ne permettrait pas de différencier, parmi les couples *(littéral, synset)* présents dans la ressource à évaluer mais absents de la ressource de référence, ceux qui sont effectivement erronés de ceux qui sont corrects et sont donc manquants dans la ressource de référence.

Le WOLF 1.0 n'étant pas encore disponible, comme expliqué précédemment, les évaluations données dans cette partie concernent le résultat de l'étape d'extension initiale, c'est-à-dire le WOLF dans sa version 0.2.

Nous avons donc procédé à une double évaluation, partiellement automatique et partiellement manuelle, de la façon suivante. Tout d'abord, tout couple *(littéral, synset)*

présent dans le WOLF et dans le FWN est considéré comme correct (rappelons que le FWN ne contient que des synsets nominaux et verbaux). Nous avons ensuite sélectionné aléatoirement 100 couples *(littéral, synset)* nominaux du WOLF qui ne sont pas dans le FWN alors que le synset correspondant n'y est pas vide. Nous avons fait de même avec 100 couples verbaux, et avons évalué manuellement ces 200 couples. Enfin, nous avons sélectionné aléatoirement et évalué manuellement 100 couples impliquant un synset qui est vide dans le FWN. Parmi ces derniers, pas moins de 92 étaient corrects, ce qui peut s'expliquer par le fait que les synsets concernés, dont la majorité sont nominaux, sont souvent rares ou spécifiques, avec des littéraux souvent monosémiques et donc faciles à traduire. Pour chacune de ces catégories, nous avons estimé le nombre de couples *(littéral, synset)* corrects en multipliant le nombre de couples par la précision obtenue. Les résultats nous permettent d'estimer à environ 65 700 le nombre de couples *(littéral, synset)* corrects dans le WOLF 0.2 sur un total de 76 436 couples, soit une précision de l'ordre de 86 %. On peut comparer ce score avec ceux de *WoNeF* : la précision des 88 736 couples présents dans cette ressource a été évaluée avec soin par Pradet & *al.* (2014) comme étant de 68,9 %. Une version orientée-précision de *WoNeF* est également distribuée. Elle atteint 91,5 % de précision, mais ne contient que 15 625 couples.

6. Travaux en cours et perspectives

Le travail de validation manuelle des candidats intrus se poursuit. C'est donc à court terme que devrait être publiée la version 1.0 de WOLF, qui en intègrera les résultats. C'est d'autant plus souhaitable que les taux d'erreurs constatés sur les wordnets dans lesquels nous avons cherché à détecter des intrus sont bien plus élevés dans les synsets de base que dans les synsets les plus spécifiques, comme le montre la précision de 92 % concernant les couples *(littéral, synset)* obtenus sur les synsets non couverts par le FWN, qui sont donc les moins fréquents. Or ces synsets de base auront vraisemblablement tendance à jouer un plus grand rôle dans toute intégration de ces ressources au sein d'outils tels que des systèmes de traduction automatique, d'extraction d'information ou d'annotation sémantique profonde.

En parallèle aux ressources lexico-sémantiques comme les wordnets, l'importance des représentations sémantiques sous forme de vecteurs a pris récemment une importance considérable, renouvelant ainsi les approches sémantiques distributionnelles par des outils dont certains s'appuient sur la notion de *word embedding* et dont le plus connu est *word2vec* (Mikolov & *al.* 2013). Nous prévoyons, dans un avenir proche, de faire évoluer nos techniques d'extension et de nettoyage de synsets afin qu'ils tirent parti de ces avancées. Il sera ainsi intéressant, pour améliorer le WOLF, de comparer l'adéquation des informations fournies par *SemanticVectors*, utilisé jusqu'ici, avec celles produites par *word2vec*, *GloVe*, ou l'une de leurs extensions. Le calcul de représentations vectorielles pour les synsets et les couples *(littéral, synset*) (c'est-à-dire les lexèmes), par exemple pour développer des systèmes de désambiguïsation lexicale, sera également possible, et a déjà été étudié pour d'autres langues (Rothe & Schütze 2015 ; Perianin & *al.* 2016).

Enfin, nous aimerions explorer l'utilisation de ressources telles que wordnet dans le contexte de l'informatisation de certains travaux en linguistique historique. Il est évident que les similarités sémantiques telles qu'elles peuvent être extraites d'un wordnet ne recouvrent qu'une partie des changements sémantiques attestés dans la diachronie des langues. Il n'en reste pas moins qu'une ressource de type wordnet constitue un point de départ utile en vue du développement de modélisations informatiques de l'histoire du lexique, par exemple du lexique français, qui prenne en compte les changements phonétiques, morphologiques et sémantiques.

Benoît Sagot
Inria (équipe ALMAnaCH)
2 rue Simone Iff, CS 42 112, 75589 Paris Cedex 12
<benoit.sagot@inria.fr>

Références

Apidianaki, M. & B. Sagot (2014). Data-driven synset induction and disambiguation for wordnet development. *Language Resources and Evaluation*, 48-4, 655-677.

Bond, F. & K. Paik (2012). A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, Matsue, Japon, 64-71

Carpuat, M. & D. Wu (2007). Improving statistical machine translation using word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, 61-72.

Copestake, A., A. Sanfilippo, T. Briscoe & V. de Paiva. (1993). The ACQUILEX LKB: An Introduction. In T. Briscoe, A. Copestake et V. de Paiva (eds), *Inheritance, Defaults and the Lexicon*, Cambridge, Cambridge University Press, 148-163.

Cuadros, M. & G. Rigau (2006). Quality assessment of large scale knowledge resources. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP '06)*, Sydney, Australie, 534-541.

Daumé III, H. (2004). Notes on CG and LM-BFGS optimization of logistic regression. <https://www.umiacs.umd.edu/~hal/docs/daume04cg-bfgs.pdf>.

Declerck, T., A.G. Pérez, O. Vela, Z. Gantner & D. Manzano-Macho (2006). Multilingual lexical semantic resources for ontology translation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2006)*, Gênes, Italie.

Diab, M (2004). The feasibility of bootstrapping an Arabic WordNet leveraging parallel corpora and an English WordNet. In *Proceedings of the Arabic Language Technologies and Resources*.

Dyvik, H (2002). Translations as semantic mirrors: from parallel corpus to wordnet. In *Post-proceedings of the ICAME 2002 Conference* (revised version), Göteborg, Suède.

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

Fillmore, C. (1968). The case for case. In E. Bach et R.T. Harms (eds), *Universals in Linguistic Theory*, New-York, Holt, Rinehart and Winston, 1-88.

Fišer, D. & B. Sagot (2015). Constructing a poor man's wordnet in a resource-rich world. *Language Resources and Evaluation*, 1-35.

Fung, P (1995). A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. In *Proceedings of the 33rd annual meeting of the Association for Computational Linguistics (ACL '95)*, Cambridge (MA), 236-243.

Gabrilovich, E. & S. Markovitch (2006). Overcoming the brittleness bottleneck using *wikipedia*: enhancing text categorization with encyclopedic knowledge. In *Proceedings of the 21st national conference on Artificial intelligence (AAAI'06)*, Boston (MA), 1301-1306.

Gábor, K., M. Apidianaki, B. Sagot & E. Villemonte de La Clergerie (2012). Boosting the Coverage of a Semantic Lexicon by Automatically Extracted Event Nominalizations. In *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC)*, Istanbul, Turquie.

Hanoka, V (2015). *Extraction et complétion de terminologies multilingues*, Thèse de doctorat, Université Paris Diderot.

Hanoka, V. & B. Sagot (2012). Wordnet creation and extension made simple: A multilingual lexicon-based approach using *wiki* resources. In *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC)*, Istanbul, Turquie.

Hanoka, V. & B. Sagot (2014). YaMTG: An Open-Source Heavily Multilingual Translation Graph Extracted from Wiktionaries and Parallel Corpora. In *Proceedings of the international conference on Language Resources and Evaluation (LREC)*, Reykjavik, Islande.

Harabagiu, S., D. Moldovan, M. Pasca & *al*. (2000). Falcon: Boosting knowledge for answer engines, In *Proceedings of TREC-9*, 479-488.

Ide, N., T. Erjavec & D. Tufiş (2002). Sense discrimination with parallel corpora. In *Proceedings of the ACL'02 workshop on Word sense disambiguation: recent successes and future directions (WSD '02)*, Philadelphie (Pennsylvanie), 61-66.

Jacquin, C., E. Desmontils & L. Monceaux (2007). French EuroWordNet Lexical Database Improvements. In *Proceedings of CicLing'07* (LNCS 4394).

Kirkpatrick, B. (1987). *Roget's Thesaurus of English Words and Phrases*, Londres, Penguin.

Knight, K. & S.K. Luk. (1994). Building a large-scale knowledge base for machine translation. In *Proceedings of the 12th national conference on Artificial intelligence (AAAI '94)*, Seattle, (Washington), 773-778.

Liu, H. (2003). Unpacking meaning from words: A context-centered approach to computational lexicon design. In *Proceedings of the 4th International and Interdisciplinary Conference on Modeling and Using Context (Context 2003)*, 218-232.

Matuszek, C., J. Cabral, M. Witbrock & J. Deoliveira (2006). An introduction to the syntax and content of Cyc. In *Proceedings of the 2006 AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering*, 44-49.

de Melo, G. & G. Weikum (2009). Towards a universal wordnet by learning from combined evidence. In *Proceedings of the 18th ACM conference on Information and knowledge management (CIKM '09)*, Hong Kong, Chine, 513-522.

Mikolov, T., I. Sutskever, K. Chen, G. Corrado & J. Dean (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 3111-3119.

Miller, G.A (1995). Wordnet: A lexical database for English. *Communications of the ACM*, 38-11, 39-41.

Mouton, C. & G. de Chalendar (2010). JAWS: Just Another WordNet Subset, In *Actes de la conference sur le Traitement Automatique des Langues Naturelles (TALN 2010)*, Montréal (Québec).

Nastase, V (2008). Topic-driven multi-document summarization with encyclopedic knowledge and spreading activation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '08)*, Honolulu (Hawaii), 763-772.

Navigli, R. & S.P. Ponzetto (2012). *BabelNet*: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193, 217-250.

Orav, H. & K. Vider (2004). Concerning the difference between a conception and its application in the case of the estonian wordnet. In *Proceedings of the 2nd International Conference of the Global WordNet Association (GWC-2004)*, Brno, République tchèque, 285-290.

Perianin, T., H. Senuma & A. Aizawa (2016). Exploiting Synonymy and Hypernymy to Learn Efficient Meaning Representations. In *Proceedings of the 18th International Conference on Asia-Pacific Digital Libraries (ICADL 2016)*, Tsukuba, Japan, 137-143.

Pianta, E., L. Bentivogli & C. Girardi (2004). Fighting arbitrariness in wordnet-like lexical databases: a natural language motivated remedy. In *Proceedings of the 1st International Conference of the Global WordNet Association (GWC-2002)*, Mysore, Inde.

Pradet, Q., G. de Chalendar & J. Desormeaux Baguenier. (2014). WoNeF, an improved, expanded and evaluated automatic French translation of wordnet. In *Proceedings of the Seventh Global Wordnet Conference (GWC 2014)*, 32-39.

Reiter, N., M. Hartung & A. Frank (2008). A Resource-Poor Approach for Linking Ontology Classes to *Wikipedia* Articles. In J. Bos et R. Delmonte (eds), *Semantics in Text Processing. STEP 2008 Conference Proceedings, Research in Computational Semantics*, vol. 1, College Publications, 381-387.

Resnik, P. & D. Yarowsky (1997). A perspective on word sense disambiguation methods and their evaluation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, Washington, D.C., 79-86.

Richardson, S.D., W.B. Dolan & L. Vanderwende (1998). *Mindnet*: Acquiring and structuring semantic information from text. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, Montréal (Québec), 1098-1102.

Rothe, S. & H. Schütze (2015). Autoextend: Extending word embeddings to embeddings for synsets and lexemes. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, Beijing, Chine, 1793-1803.

Rudnicka, E., M. Maziarz, M. Piasecki & S. Szpakowicz (2012). A strategy of mapping Polish Wordnet onto Princeton Wordnet. In *Proceedings of COLING 2012: Posters*, Mumbai, Inde, 1039-1048.

Ruiz-Casado, M., E. Alfonseca & P. Castells (2005). Automatic assignment of *wikipedia* encyclopedic entries to wordnet synsets, In *Proceedings of Advances in Web Intelligence*.

Sagot, B. (2010). The Le*fff*, a freely available, accurate and large-coverage lexicon for French. In *Proceedings of the 7th international conference on Language Resources and Evaluation (LREC'2010)*, La Valette, Malte.

Sagot, B. & D. Fišer (2008). Building a free French wordnet from multilingual resources. In *Proceedings of Ontolex 2008*, Marrakech, Maroc.

Sagot, B. & D. Fišer (2011). Extending wordnets by learning from multiple resources. In *Proceedings of the 5th Language and Technology Conference (LTC 2011),* Poznań, Pologne.

Sagot, B. & D. Fišer. (2012a). Automatic Extension of WOLF. In *Proceedings of the 6th International Global Wordnet Conference (GWC 2012)*, Matsue, Japon.

Sagot, B. & D. Fišer. (2012b). Cleaning noisy wordnets. In *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC)*, Istanbul, Turquie.

Sagot, B., K. Fort & F. Venant (2009). Extension et couplage de ressources syntaxiques et sémantiques sur les adverbes. *Linguisticæ Investigationes*, 32-2.

Sornlertlamvanich, V. (2010). *Asian WordNet*: Development and service in collaborative approach. In *Proceedings of the 5th International Conference of the Global WordNet Association (GWC 2010)*, Mumbai, Inde.

Suchanek, F.M., G. Kasneci & G. Weikum (2008). *Yago*: A large ontology from *Wikipedia* and *WordNet*. *Journal of Web Semantics*, 6-3, 203-217.

Tavčar, A., D. Fišer & T. Erjavec (2012). Slowcrowd: orodje za popravljanje wordneta z izkoriščanjem moči množic. In *Proceedings of the 8th Language Technologies Conference (part of IS 2012)*, vol. C, Ljubljana, Slovénie, 197-202.

Tiedemann, J. (2003). *Recycling Translations-Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing*. Thèse de doctorat, Uppsala Universitet.

Tufiş, D. (2000). *BalkaNet*. Design and Development of a Multilingual Balkan *WordNet*. *Romanian Journal of Information Science and Technology*, 7-1/2.

Tufiş, D., S. Koeva, T. Erjavec, M. Gavrilidou & C. Krstev (2009). Building language resources and translation models for machine translation focused on south Slavic and Balkan languages. In J. Machačová & K. Rohsmann (eds), *Scientific results of the SEE-ERA.NET Pilot Joint Call*, 37-48.

Vossen, P. (ed.) (1999). EuroWordNet*: a multilingual database with lexical semantic networks for European Languages*. Dordrecht, Kluwer Academic Publisher.

Widdows, D. & K. Ferraro (2008). Semantic vectors: a scalable open source package and online technology management application. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Maroc.

Wong, S.H.S (2004). Fighting arbitrariness in wordnet-like lexical databases: a natural language motivated remedy. In *Proceedings of the 2nd International Conference of the Global WordNet Association (GWC 2004)*, Brno, République tchèque, 234-241.

Yokoi, T. (1995). The EDR electronic dictionary, *Communications of the ACM*, 38-11, 42-44.