Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

Ressources lexicales pour la sémantique : WordNet, BabelNet, PropBank, FrameNet, DBpedia et SUMO

Ahana Chattopadhyay[1]
(1) Université de Lorraine, Nancy, France
ahana.chattopadhyay1@etu.univ-lorraine.fr

RÉSUMÉ

Cet article offre un aperçu concis des ressources lexicales ci-après, dans le cadre de la sémantique computationnelle : WordNet, BabelNet, PropBank, FrameNet, DBpedia et SUMO. L'accent est mis sur leur structure et leur application.

ABSTRACT

Lexical Resources for Semantics : WordNet, BabelNet, PropBank, FrameNet, DBpedia, and SUMO

This paper provides a short overview of the following lexical resources within Computational Semantics : WordNet, BabelNet, PropBank, FrameNet, DBpedia, and SUMO, with a focus on their structure and application.

MOTS-CLÉS : Ressources lexicales, sémantique, WordNet, BabelNet, PropBank, FrameNet, DBpedia, SUMO, traitement automatique du langage naturel.

KEYWORDS: Lexical resources, semantics, WordNet, BabelNet, PropBank, FrameNet, DBpedia, SUMO, natural language processing.

1 Introduction

La sémantique computationnelle repose sur un ensemble de ressources lexicales qui, bien que souvent évoquées, méritent une attention particulière quant à leur rôle et leur complémentarité. Dans ce travail, nous proposons une réflexion synthétique sur des outils incontournables tels que WordNet, BabelNet, PropBank, FrameNet, DBpedia et SUMO. Ces ressources fournissent des représentations structurées et formalisées du lexique qui permettent aux machines de décoder non seulement la forme des mots, mais aussi leurs multiples significations, leurs relations sémantiques complexes et leurs rôles dans la construction du sens des phrases.

2 Wordnet

Selon Miller *et al.* (1990), WordNet est un système de référence lexical en ligne, conçu en s'inspirant des théories psycholinguistiques sur la mémoire lexicale humaine. Il vise à améliorer la façon dont nous utilisons les informations lexicales grâce à l'informatique. Contrairement aux méthodes traditionnelles, comme l'ordre alphabétique, qui regroupent les mots par leur orthographe et éparpillent ceux qui ont des significations similaires, WordNet offre une alternative à cette organisation. Sa structure est basée sur des "synsets", qui sont des ensembles de synonymes regroupant les noms, les

Actes de l'atelier Avancement de l'AMR et de l'Analyse Sémantique 2025 (4AS), pages 17–23.

verbes, les adjectifs et les adverbes anglais. Chaque synset représente un concept lexical. Par exemple, “board” et “plank” sont dans le même synset car ils partagent un sens. Ces synsets sont reliés par des relations sémantiques telles que la synonymie (similarité de sens), l’antonymie (opposition), l’hyponymie/hypernymie (relation “est un”), et la méronymie/holonymie (relation partie-tout). “Érable” est un hyponyme de “arbre”, et “roue” est un méronyme de “voiture”. WordNet est donc organisé par le sens des mots, et non par leur forme. De plus, des versions multilingues de WordNet ont été développées dans plus de 100 langues, permettant l’exploitation de cette structure sémantique dans un cadre international et multilingue [1].

WordNet présente de multiples applications concrètes en Traitement Automatique des Langues (TAL) (Morato *et al.*, 2004), se concentrant principalement sur l’amélioration de la recherche d’informations et de la compréhension sémantique. Il est utilisé pour accroître la précision des moteurs de recherche par l’expansion de requêtes et la désambiguïsation du sens des mots, pour catégoriser et structurer des documents, et pour développer des systèmes capables d’extraire des informations de diverses sources, incluant le texte, l’audio et la vidéo. De plus, WordNet contribue à des tâches telles que l’enseignement des langues, la traduction et la construction de systèmes d’information personnalisés.

3 Babelnet

BabelNet (Navigli & Ponzetto, 2012) est un vaste réseau sémantique multilingue [2], généré de manière automatique. Son but est de fournir une ressource qui aide à comprendre comment les mots sont utilisés pour exprimer des significations. Cette connaissance lexicale est utile aussi bien aux personnes qui apprennent des langues qu’aux outils de traitement automatique des langues. Par exemple, dans la désambiguïsation des entités nommées, BabelNet permet de distinguer entre “Apple” la société et “apple” le fruit, en offrant un contexte et des relations sémantiques clairs pour chaque sens.

BabelNet est structuré comme un graphe, où les nœuds représentent des concepts ou des entités nommées, et les arêtes indiquent les relations entre eux. Chaque concept est décrit en plusieurs langues, ce qui fait de BabelNet une ressource multilingue. BabelNet est construit en intégrant WordNet, un lexique computationnel de l’anglais, et Wikipédia, une encyclopédie multilingue, combinant ainsi richesse lexicale et couverture encyclopédique.

Parmi ses diverses applications, il est notamment utilisé pour la désambiguïsation des sens des mots (Navigli & Ponzetto, 2012), aidant à déterminer le sens précis d’un mot selon le contexte. Il a été utilisé pour perfectionner la traduction automatique statistique (TAS) en traitant les mots hors vocabulaire (OOV) (Du *et al.*, 2016) et pour faciliter le développement du web sémantique multilingue (Navigli, 2012) en reliant Wikipédia et WordNet pour fournir une ressource exhaustive de concepts et d’entités nommées dans de multiples langues.

4 Propbank

PropBank (Palmer *et al.*, 2005) est un projet monolingue en anglais, qui enrichit les structures syntaxiques du Penn Treebank en ajoutant une couche d’informations prédicat-argument, ou étiquettes de rôles sémantiques. Cette ressource fournit une représentation sémantique des phrases qui est organisée en couches où chaque verbe se voit attribuer des rôles numérotés (Arg0, Arg1, etc.) pour ses arguments, et des balises plus générales (ArgM) pour les arguments adjoints. Par exemple, dans la

1. Plus d’information sur : https://globalwordnet.org/resources/wordnets-in-the-world
2. Selon le site officiel (https://babelnet.org/about) BabelNet 5.3 couvre 600 langues.

phrase “John broke the window”, PropBank identifie “John” comme Arg0 (l’agent) et “the window” comme Arg1 (le patient). Ces rôles aident à clarifier qui a fait quoi à qui, ce qui est essentiel pour comprendre le sens d’une phrase. En revanche, Universal PropBank 2.0 (Jindal *et al.*, 2022) étend ce concept à 23 langues grâce à des méthodes de projection automatique et à des modèles neuronaux, en proposant des annotations à la fois de type “span” et “dépendance”. Conçu pour l’annotation sémantique multilingue, il permet d’aligner les structures sémantiques entre langues en utilisant l’anglais comme pivot.

PropBank a pour but de créer un corpus annoté à la main qui couvre un large éventail de phénomènes linguistiques. Ce corpus facilite le développement de systèmes de compréhension du langage plus robustes et indépendants du domaine. Il permet également d’étudier de manière quantitative comment et pourquoi les alternances syntaxiques se produisent (Palmer *et al.*, 2005). À l’aide de l’annotation des rôles sémantiques (SRL) (Jindal *et al.*, 2022), il est possible d’améliorer un large éventail d’applications en traitement automatique des langues, telles que l’inférence textuelle, le questionnement automatique, la traduction automatique et l’extraction d’informations.

5 Framenet

Le projet FrameNet (Baker *et al.*, 1998) se propose de créer des descriptions sémantiques de plusieurs milliers d’unités lexicales anglaises, en les appuyant sur des exemples annotés sémantiquement tirés de corpus d’anglais contemporain. La base de données FrameNet contient des descriptions des cadres sémantiques qui sous-tendent les significations des mots décrits, ainsi que la représentation de la valence (sémantique et syntaxique) de plusieurs milliers de mots et phrases. Chaque entrée inclut également une collection représentative d’attestations de corpus annotées, qui illustrent les liens observés entre les “éléments de cadre” et leurs réalisations syntaxiques. Par exemple, le cadre “TRANSPORTATION” inclut des éléments comme “MOVER”, “MEANS” et “PATH”, montrant comment les mots liés au transport sont compris sémantiquement. La version multilingue de FrameNet (Baker *et al.*, 2018) a été développée pour environ une douzaine de langues.

FrameNet a été largement appliqué en linguistique computationnelle, en particulier dans les tâches nécessitant une compréhension sémantique profonde. Une application majeure est l’analyse sémantique basée sur les cadres, où les systèmes prédisent les prédicats évoquant des cadres et leurs rôles dans le texte afin de soutenir la compréhension du langage naturel à grande échelle (Das *et al.*, 2014). Ceci est crucial pour des tâches telles que l’extraction d’informations, où des connaissances structurées sont dérivées de textes non structurés, et pour la question-réponse (QA), où FrameNet aide à aligner les questions et les réponses en fonction de leurs structures sémantiques plutôt que de leurs formes de surface (Shen & Lapata, 2007). Les riches annotations prédicat-argument de FrameNet soutiennent également la reconnaissance de paraphrases, la reconnaissance de l’implication textuelle et la construction de ressources lexicales multilingues (Litkowski, 2010). Ces applications démontrent la polyvalence de FrameNet dans les systèmes de TAL monolingues et multilingues.

6 DBpedia

DBpedia (Auer *et al.*, 2007) est le fruit d’un effort communautaire visant à extraire des informations structurées de Wikipédia et à les rendre disponibles sur le web. Sa structure est basée sur un ensemble de données RDF [3] (Resource Description Framework) qui comprend des descriptions de plus de 1,95 million d’“objets”, incluant des personnes, des lieux, des albums de musique et des films. Par

3. Plus d’information sur : `https://www.w3.org/RDF/`

exemple, il contient des informations sur 80 000 personnes et 70 000 lieux. Pour illustrer la structure, les informations sont organisées de manière à ce que l'on puisse interroger et lier des données sur un sujet donné (comme une ville) à d'autres données sur des sujets connexes (comme la région dans laquelle se trouve la ville, ou les personnalités qui y sont nées), facilitant ainsi la navigation et la découverte de relations entre les données. DBpedia extrait des informations de 97 éditions linguistiques de Wikipédia et prend en charge l'alignement des infoboxes dans 23 langues (Mendes *et al.*, 2012).

DBpedia a été créé pour permettre aux utilisateurs de poser des requêtes complexes sur les données extraites de Wikipédia et de lier d'autres ensembles de données du web aux données de Wikipédia. DBpedia est utilisé dans plusieurs applications concrètes (Kobilarov *et al.*, 2009), notamment comme source de données pour les applications web, en fournissant des informations structurées et multilingues sur des millions de concepts. Il joue également un rôle central dans le Web de données en tant que hub d'interconnexion, permettant à des organisations comme la BBC de relier leurs données internes à des sources externes. Enfin, DBpedia est employé pour l'annotation sémantique de contenus, facilitant l'enrichissement automatique de documents grâce à des outils de reconnaissance d'entités comme OpenCalais et Zemanta. En outre, il est généralement utilisé pour la désambiguïsation, le questionnement automatique, l'extraction de relations et l'enrichissement de requêtes (Mendes *et al.*, 2012).

7 SUMO

SUMO (Suggested Upper Merged Ontology) [4] (Niles & Pease, 2001) est une ontologie de haut niveau conçue comme base pour le travail du groupe de travail Standard Upper Ontology de l'IEEE. Sa structure fournit des définitions pour les termes généraux et sert de fondation à des ontologies de domaine plus spécifiques. Par exemple, elle inclut des concepts de base comme 'Entity', qui se divise en 'Physical' (ce qui a une position dans l'espace/temps) et 'Abstract' (tout le reste), ainsi que des distinctions importantes au sein de 'Physical' comme 'Object' et 'Process'.

Selon Niles & Pease (2001), cette ontologie a été conçue pour donner aux ordinateurs un langage commun qui se rapproche de la richesse du langage humain, facilitant ainsi les applications logicielles avancées. Il est utilisé pour concevoir de nouvelles bases de connaissances, réutiliser et intégrer des bases de données existantes, et permettre l'interopérabilité entre les ontologies de domaine. Avec son ontologie normalisée, il contribue à résoudre les difficultés que rencontre actuellement le traitement automatique du langage et facilite une meilleure compatibilité entre les différents systèmes informatiques.

8 Ressources additionnelles

Wikidata (Vrandečić & Krötzsch, 2014) est une base de connaissances collaborative et multilingue développée par la Wikimedia Foundation pour centraliser les données structurées de Wikipédia et d'autres projets. Contrairement aux plus de 280 éditions linguistiques de Wikipédia, Wikidata repose sur une plateforme unique prenant en charge plus de 287 langues. Chaque entité est identifiée par un Q-ID et décrite à l'aide d'étiquettes multilingues et d'énoncés structurés. Elle permet la coexistence de données contradictoires en associant plusieurs affirmations à leurs sources respectives. Les données sont publiées sous licence CC0 dans des formats lisibles par machine (JSON, RDF). Wikidata est

4. Selon le site officiel (`https://www.ontologyportal.org/index.html`), SUMO est rédigé dans le langage SUO-KIF.

largement utilisée pour des tâches telles que le questionnement automatique, l'annotation sémantique ou la recherche d'information, et constitue l'un des projets Wikimedia les plus actifs avec des milliers de contributeurs.

Lancée en 2007 et acquise par Google en 2010, Freebase (Bollacker *et al.*, 2008; Pellissier Tanon *et al.*, 2016) était une vaste base de données collaborative représentant des connaissances générales sous forme de graphes. Elle modélisait les faits sous forme de triplets (sujet–prédicat–objet) avec un schéma ouvert, et proposait un langage de requête propre (MQL). À son apogée, Freebase contenait plus de 3 milliards de faits sur 50 millions d'entités. Google a finalement décidé de fermer Freebase et de migrer ses données vers Wikidata, en raison du caractère multilingue et communautaire de ce dernier. Cette migration a nécessité un alignement complexe des schémas, avec l'aide d'outils comme le Primary Sources Tool.

VerbNet [5] est une ressource lexicale regroupant environ 5 800 verbes anglais répartis en plus de 270 classes en fonction des similarités syntaxiques et sémantiques. S'inspirant de la classification de Beth Levin, VerbNet associe à chaque classe des cadres syntaxiques, des rôles thématiques et des représentations logiques des événements. Ces classes comprennent également des contraintes de sélection et des prédicats sémantiques. VerbNet est essentiel dans des tâches telles que l'étiquetage sémantique, la modélisation d'événements ou la désambiguïsation des sens des verbes.

SemLink (Stowe *et al.*, 2021) relie différentes ressources lexicales (PropBank, VerbNet, FrameNet, WordNet) en fournissant des mappages entre leurs rôles et sens verbaux. Il comprend un corpus annoté et des outils pour assurer la mise à jour continue. SemLink 2.0 propose des mises à jour automatiques et manuelles, des représentations vectorielles de sens et des APIs. En unifiant ces ressources, SemLink facilite la création de corpus d'entraînement, le développement d'analyseurs sémantiques et l'interopérabilité des annotations.

9 Conclusion

Les ressources lexicales examinées jouent un rôle fondamental dans le domaine de la sémantique computationnelle et du traitement automatique des langues. WordNet et BabelNet offrent des informations indispensables sur le sens et les relations entre les mots, ce qui facilite la compréhension du langage par les machines. PropBank et FrameNet apportent des précisions sur les rôles sémantiques, contribuant à l'analyse de la structure des phrases. DBpedia et SUMO fournissent des connaissances structurées et des ontologies, ce qui aide à organiser et à accéder à l'information de manière efficace. Ces outils sont essentiels pour permettre aux systèmes informatiques de traiter le langage de manière plus intelligente, en allant au-delà de la simple reconnaissance des mots pour atteindre une interprétation plus riche du sens. Par ailleurs, des ressources additionnelles telles que Wikidata, VerbNet et SemLink complètent cet ensemble d'outils, en offrant des fonctionnalités variées et en encourageant l'interopérabilité des données.

Remerciements

L'idée de rédiger cet article a été fortement inspirée par le cours "Ressources lexicales", dispensé il y a quelques mois dans le cadre de notre programme de Master en Traitement Automatique des Langues, par Bruno Guillaume. Je le remercie pour l'inspiration que ce cours m'a apportée.

5. D'après le guide officiel accessible à l'adresse suivante : https://verbs.colorado.edu/verb-index/VerbNet_Guidelines.pdf.

Références

AUER S., BIZER C., KOBILAROV G., LEHMANN J., CYGANIAK R. & IVES Z. (2007). Dbpedia : a nucleus for a web of open data. In *Proceedings of the 6th International The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference*, ISWC'07/ASWC'07, p. 722–735, Berlin, Heidelberg : Springer-Verlag.

BAKER C. F., ELLSWORTH M., PETRUCK M. R. L. & SWAYAMDIPTA S. (2018). Frame semantics across languages : Towards a multilingual FrameNet. In D. SCOTT, M. WALKER & P. FUNG, Éds., *Proceedings of the 27th International Conference on Computational Linguistics : Tutorial Abstracts*, p. 9–12, Santa Fe, New Mexico, USA : Association for Computational Linguistics.

BAKER C. F., FILLMORE C. J. & LOWE J. B. (1998). The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98/COLING '98, p. 86–90, USA : Association for Computational Linguistics. DOI : 10.3115/980845.980860.

BOLLACKER K., EVANS C., PARITOSH P., STURGE T. & TAYLOR J. (2008). Freebase : a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, p. 1247–1250, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/1376616.1376746.

DAS D., CHEN D., MARTINS A. F. T., SCHNEIDER N. & SMITH N. A. (2014). Frame-semantic parsing. *Computational Linguistics*, **40**(1), 9–56. DOI : 10.1162/COLI_a_00163.

DU J., WAY A. & ZYDRON A. (2016). Using babelnet to improve oov coverage in smt. In *International Conference on Language Resources and Evaluation*.

JINDAL I., RADEMAKER A., ULEWICZ M., LINH H., NGUYEN H., TRAN K.-N., ZHU H. & LI Y. (2022). Universal Proposition Bank 2.0. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 1700–1711, Marseille, France : European Language Resources Association.

KOBILAROV G., BIZER C., AUER S. & LEHMANN J. (2009). Dbpedia - a linked data hub and data source for web applications and enterprises. In *Proceedings of Developers Track of 18th International World Wide Web Conference (WWW 2009), April 20th-24th, Madrid, Spain*.

LITKOWSKI K. (2010). Hans c. boas (ed.). multilingual framenets in computational lexicography : Methods and applications. *International Journal of Lexicography - INT J LEXICOGR*, **23**, 105–109. DOI : 10.1093/ijl/ecp034.

MENDES P., JAKOB M. & BIZER C. (2012). DBpedia : A multilingual cross-domain knowledge base. In N. CALZOLARI, K. CHOUKRI, T. DECLERCK, M. U. DOĞAN, B. MAEGAARD, J. MARIANI, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, p. 1813–1817, Istanbul, Turkey : European Language Resources Association (ELRA).

MILLER G. A., BECKWITH R., FELLBAUM C., GROSS D. & MILLER K. J. (1990). Introduction to wordnet : An on-line lexical database*. *International Journal of Lexicography*, **3**(4), 235–244. DOI : 10.1093/ijl/3.4.235.

MORATO J., MARZAL M., LLORENS J. & MOREIRO J. (2004). Wordnet applications. *Proceedings of the 2nd Global Wordnet Conference*, **2004**.

NAVIGLI R. (2012). Babelnet goes to the (multilingual) semantic web. In *Proceedings of the 3rd International Conference on Multilingual Semantic Web - Volume 936*, MSW3'12, p. 1–10, Aachen, DEU : CEUR-WS.org.

NAVIGLI R. & PONZETTO S. P. (2012). Babelnet : The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, **193**, 217–250. DOI : https ://doi.org/10.1016/j.artint.2012.07.001.

NILES I. & PEASE A. (2001). Towards a standard upper ontology. In *Proceedings of the International Conference on Formal Ontology in Information Systems - Volume 2001*, FOIS '01, p. 2–9, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/505168.505170.

PALMER M., GILDEA D. & KINGSBURY P. (2005). The proposition bank : An annotated corpus of semantic roles. *Comput. Linguist.*, **31**(1), 71–106. DOI : 10.1162/0891201053630264.

PELLISSIER TANON T., VRANDEČIĆ D., SCHAFFERT S., STEINER T. & PINTSCHER L. (2016). From freebase to wikidata : The great migration. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, p. 1419–1428, Republic and Canton of Geneva, CHE : International World Wide Web Conferences Steering Committee. DOI : 10.1145/2872427.2874809.

SHEN D. & LAPATA M. (2007). Using semantic roles to improve question answering. In J. EISNER, Éd., *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, p. 12–21, Prague, Czech Republic : Association for Computational Linguistics.

STOWE K., PRECIADO J., CONGER K., BROWN S. W., KAZEMINEJAD G., GUNG J. & PALMER M. (2021). SemLink 2.0 : Chasing lexical resources. In S. ZARRIESS, J. BOS, R. VAN NOORD & L. ABZIANIDZE, Éds., *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, p. 222–227, Groningen, The Netherlands (online) : Association for Computational Linguistics.

VRANDEČIĆ D. & KRÖTZSCH M. (2014). Wikidata : a free collaborative knowledgebase. *Commun. ACM*, **57**(10), 78–85. DOI : 10.1145/2629489.