Convertir le Trésor de la Langue Française en Ontolex-Lemon : un zeste de données liées

Sina Ahmadi, Mathieu Constant, Karën Fort, Bruno Guillaume, John P. Mccrae

Convertir le Trésor de la Langue Française en Ontolex-Lemon : un zeste de données liées

Sina Ahmadi[1*] Mathieu Constant[3]
Karën Fort[2,4] Bruno Guillaume[4] John P. McCrae[1]
(1) Insight Centre for Data Analytics, National University of Ireland Galway (2) Sorbonne Université
(3) ATILF, Université de Lorraine & CNRS (4) Université de Lorraine, CNRS, Inria, LORIA
sina.ahmadi@insight-centre.org, mathieu.constant@univ-lorraine.fr,
{karen.fort,bruno.guillaume}@loria.fr, john.mccrae@insight-centre.org

RÉSUMÉ

Nous présentons dans ce papier les travaux que nous avons réalisés pour convertir dans le modèle Ontolex-Lemon l'une des plus importantes ressources lexicographiques pour le français : le Trésor de la Langue Française. En effet, malgré l'utilisation généralisée de cette ressource, son format actuel, basé sur XML, ne respecte pas les standards les plus récents de la représentation des données lexicographiques, notamment ceux basés sur les données liées. Nos travaux mettent en lumière la nécessité d'établir des mécanismes permettant d'augmenter l'inter-opérabilité des ressources et des technologies pour créer et maintenir des ressources lexicographiques.

ABSTRACT

Revisiting the *Trésor de la Langue Française*

In this paper, we report our efforts to convert one of the most comprehensive lexicographic resources of French, the *Trésor de la Langue Française*, into the Ontolex-Lemon model. Despite the widespread usage of this resource, the original XML format seems to impede its integration in language technology tools. In order to breathe new life into this resource, we examine the usage and the conversion to more interoperable formats, primarily those based on the linguistic linked data, to provide this resource to a broader range of applications and users.

MOTS-CLÉS : Ressources lexicographiques, données linguistiques liées, TAL.

KEYWORDS: Lexical-semantic resources, linguistic linked data, natural language processing.

1 Introduction

Les ressources lexico-sémantiques sont des référentiels de connaissances qui présentent le vocabulaire d'une langue de manière descriptive, structurée ou conceptualisée. Parmi ces ressources, les dictionnaires sont les plus répandus et historiquement utilisés pour étudier les langues naturelles et les traiter grâce aux techniques de traitement automatique des langues (TAL) [1]. Par conséquent, ces dictionnaires jouent un rôle crucial dans plusieurs applications du TAL telles que la désambigüisation lexicale [2], l'étiquetage de rôles sémantiques [3] ou l'analyse syntaxique et lexicale [4]. Malgré le grand nombre de ressources issues des initiatives communautaires, comme le Wiktionnaire [1], les ressources créées par des experts restent fondamentales du fait de leur qualité et de leur degré d'élaboration.

∗. Le travail a été réalisé pendant une visite scientifique à l'ATILF.

1. https://fr.wiktionary.org

2 Ontolex-Lemon

Aux cours de ces dernières années, les standards basés sur les données liées et le Web sémantique ont changé l'éco-système de création, de représentation et de maintenance des ressources langagières, en particulier les dictionnaires [5, 6]. Les modèles de données tels qu'Ontolex-Lemon [7] définissent des ontologies en s'appuyant des ressources terminologiques et lexicales présentes sur le Web sémantique. Ces modèles permettent également d'augmenter l'inter-opérabilité et le multilinguisme des ressources en fournissant des mécanismes d'alignement de représentations sémantiques existantes sous forme d'ontologies [8, 9].

OntoLex-Lemon est un modèle basé sur LEMON – Lexicon Model for Ontologies [10] et fournit une base linguistique riche pour les ontologies, telles que la représentation des propriétés morphologiques et syntaxiques des entrées lexicales. Ce modèle s'inspire largement des modèles de données lexicaux précédents, en particulier LexInfo [11], LMF [12] et LIR [13], avec des améliorations telles qu'être purement en *Resource Description Framework* (RDF), ce qui le rend descriptif et modulaire et justifie sa promesse d'adaptabilité dans la gestion des ressources linguistiques. La figure 1 montre la conceptualisation de base de ce modèle qui est fondé sur le principe de référence sémantique où une entrée lexicale est définie par un individu, une classe ou une propriété définis dans l'ontologie.

3 Convertir le TLFi en Ontolex-Lemon

Le Trésor de la Langue Française est une des plus importantes ressources lexicographiques du français. Il contient 100 000 entrées, 270 000 définitions et 430 000 exemples du XIVème au XXème siècle [14]. La version informatisée de ce dictionnaire, appelée le TLF informatisé (TLFi) [2], est disponible sous format XML avec une DTD associée. La micro-structure de ce dictionnaire est enrichie par plusieurs types d'informations, notamment des sens et des définitions, des exemples d'usage, des étymologies, des indications d'emplois et de domaine général, ainsi que des locutions. En outre, les sens de chaque entrée peuvent être représentés dans une hiérarchie où les sens peuvent avoir des sous-sens pour montrer un sens plus strict. De ce fait, la structure de chaque entrée lexicale présente une complexité qui fait obstacle à son intégration dans des applications en TAL utilisant les standards actuels.

Pour faciliter l'utilisation du TLFi, nous l'avons donc converti au modèle Ontolex-Lemon. Afin de réaliser cette tâche, la structure XML du dictionnaire est parcourue et les éléments essentiels en sont extraits. Étant donné la complexité des données TLFi en XML et le manque d'uniformité dans la structure, l'extraction des données se concentre actuellement uniquement sur les lemmes, les catégories grammaticales, les sens et les définitions. Ces informations sont ensuite converties en format RDF en Ontolex-Lemon. Actuellement, 32 073 entrées du TLFi sont converties. L'annexe 2 présente le résultat de cette conversion pour l'entrée « baroque » (adjectif).

4 Conclusion et travaux futurs

Les standards actuels de données liées permettent d'augmenter l'inter-opérabilité et l'accessibilité des données langagières. Par conséquent, une version du TLFi en Ontolex-Lemon pourrait en permettre une meilleure intégration au sein des applications de TAL. La ressource est cependant très riche et nous travaillons à en extraire davantage de données pour pouvoir la convertir entièrement en Ontolex-Lemon dans un futur proche.

2. https://www.atilf.fr/ressources/tlfi/

Références

[1] Eric Laporte. Dictionaries for language processing. Readability and organization of information. PPGEL/UFES, 2013.

[2] Rada Mihalcea. Knowledge-based methods for WSD. In *Word sense disambiguation*, pages 107–131. Springer, 2007.

[3] Jie Zhou and Wei Xu. End-to-end learning of semantic role labeling using recurrent neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, pages 1127–1137, 2015.

[4] Matthieu Constant and Joakim Nivre. A transition-based system for joint lexical and syntactic analysis. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 161–171, 2016.

[5] Philipp Cimiano, Christian Chiarcos, John P. McCrae, and Jorge Gracia. *Linguistic Linked Data - Representation, Generation and Applications*, pages 3–9. Springer International Publishing, Cham, 2020.

[6] Rinke Hoekstra, Albert Meroño-Peñuela, Kathrin Dentler, Auke Rijpma, Richard Zijdeman, and Ivo Zandhuis. An ecosystem for linked humanities data. In *European Semantic Web Conference*, pages 425–440. Springer, 2016.

[7] John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. The Ontolex-Lemon model : development and applications. In *Proceedings of eLex 2017 conference*, pages 19–21, 2017.

[8] Andon Tchechmedjiev. *Interopérabilité sémantique multilingue des ressources lexicales en données lexicales liées ouvertes*. PhD thesis, Université Grenoble Alpes, 2016.

[9] Sabine Tittel and Christian Chiarcos. Historical lexicography of old french and linked open data : Transforming the resources of the dictionnaire étymologique de l'ancien français with ontolex-lemon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). GLOBALEX Workshop (GLOBALEX-2018), Miyazaki, Japan*, pages 58–66, 2018.

[10] John McCrae, Dennis Spohr, and Philipp Cimiano. Linking lexical resources and ontologies on the semantic web with lemon. In *Extended Semantic Web Conference*, pages 245–259. Springer, 2011.

[11] Philipp Cimiano, Paul Buitelaar, John McCrae, and Michael Sintek. Lexinfo : A declarative model for the lexicon-ontology interface. *Web Semantics : Science, Services and Agents on the World Wide Web*, 9(1) :29–51, 2011.

[12] Gil Francopoulo, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet, and Claudia Soria. Lexical markup framework (lmf). In *International Conference on Language Resources and Evaluation-LREC 2006*, page 5, 2006.

[13] Elena Montiel-Ponsoda, Guadalupe Aguado De Cea, Asunción Gómez-Pérez, and Wim Peters. Modelling multilinguality in ontologies. *Coling 2008 : Companion volume : Posters*, pages 67–70, 2008.

[14] Ruth Radermacher. *Le Trésor de la langue française : une analyse lexicographique*. PhD thesis, Strasbourg 2, 2004.

A Le TLFi en Ontolex-Lemon (extrait)

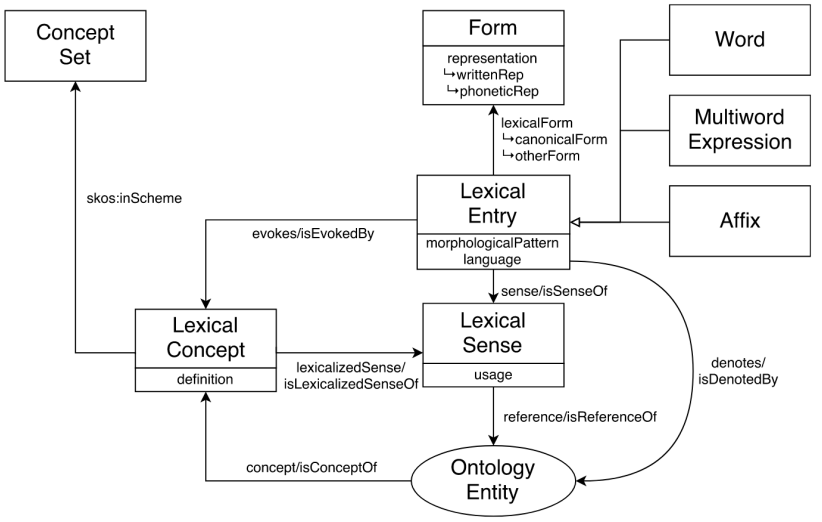

FIGURE 1 – Le modèle de données de base d'Ontolex-Lemon [7]

```
<https://www.cnrtl.fr/definition/11597> a ontolex:LexicalEntry ;
   rdfs:label "baroque"@fr ;
   lexinfo:partOfSpeech lexinfo:adjective ;
   ontolex:sense <https://www.cnrtl.fr/definition/11597#A._1.UND-1>,
       <https://www.cnrtl.fr/definition/11597#A._1.UND-2>,
       <https://www.cnrtl.fr/definition/11597#A._1.UND-3>,
       <https://www.cnrtl.fr/definition/11597#A._3.UND-6>,
       <https://www.cnrtl.fr/definition/11597#UND-9>.
   <https://www.cnrtl.fr/definition/11597#A._1.UND-1> skos:definition
       "Qui est caractéristique de la période qui a suivi la
       Renaissance classique."@fr .
   <https://www.cnrtl.fr/definition/11597#A._1.UND-2> skos:definition
       "Qui est caractéristique de la période musicale propre à
       l'Allemagne, àl'Angleterre, àl'Italie, et qui s'étend de 1580 à
       1760."@fr .
   <https://www.cnrtl.fr/definition/11597#A._1.UND-3> skos:definition
       "Qui appartient àl'époque littéraire qui, en France, correspond
       aux règnes de Henri IV et Louis XIII."@fr .
   <https://www.cnrtl.fr/definition/11597#A._3.UND-6> skos:definition
       "Artiste dont le style rappelle cette période"@fr .
   <https://www.cnrtl.fr/definition/11597#UND-9> skos:definition "Qui
       est de forme irrégulière, d'une rondeur imparfaite"@fr .
```

FIGURE 2 – La conversion de l'entrée « baroque » (adjectif) du TLFi en Ontolex-Lemon. L'entrée originale est accessible à `https://www.cnrtl.fr/definition/baroque`