Convertir le Trésor de la Langue Française en Ontolex-Lemon

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

Objectifs

❶ Nous présentons 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 informatisé** (TLFi).

❷ 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.

❸ Nous mettons en place **un point d'accès SPARQL** (*SPARQL endpoint*) pour lancer des requêtes sur le TLFi.

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). Par conséquent, ces dictionnaires jouent un rôle crucial dans plusieurs applications du TAL. Malgré le grand nombre de ressources issues des initiatives communautaires, comme le Wiktionnaire (`https://fr.wiktionary.org`), les ressources créées par des experts restent fondamentales du fait de leur qualité et de leur degré d'élaboration.

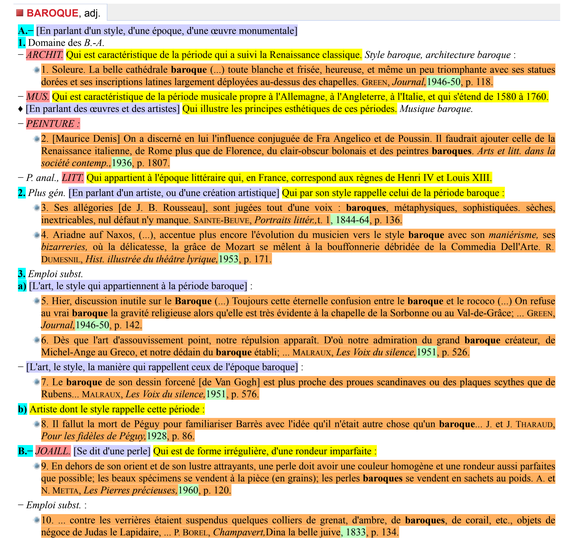
Aux cours de ces dernières années, les **standards basés sur les données liées et le Web sémantique** ont changé l'éco-système de création, de représentation et de maintenance des ressources langagières, en particulier les dictionnaires. Les modèles de données tels qu'**Ontolex-Lemon** définissent des 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.

TLF*i*

Le TLFi `https://www.atilf.fr/ressources/tlfi/` 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. Il 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**, comme indiqués dans l'image suivante de l'entrée baroque :


BAROQUE, adj.

A.— [En parlant d'un style, d'une époque, d'une œuvre monumentale]
1. Domaine des *B.-A.*
– *ARCHIT.* Qui est caractéristique de la période qui a suivi la Renaissance classique. *Style baroque, architecture baroque* :
• 1. Soleure. La belle cathédrale **baroque** (...) toute blanche et frisée, heureuse, et même un peu triomphante avec ses statues dorées et ses inscriptions latines largement déployées au-dessus des chapelles. Green, *Journal*, 1946-50, p. 118.
– *MUS.* Qui est caractéristique de la période musicale propre à l'Allemagne, à l'Angleterre, à l'Italie, et qui s'étend de 1580 à 1760.
♦ [En parlant des œuvres et des artistes] Qui illustre les principes esthétiques de ces périodes. *Musique baroque.*
– *PEINTURE* :
• 2. [Maurice Denis] On a discerné en lui l'influence conjuguée de Fra Angelico et de Poussin. Il faudrait ajouter celle de la Renaissance italienne, de Rome plus que de Florence, du clair-obscur bolonais et des peintres **baroques**. *Arts et litt. dans la société contemp.*, 1936, p. 1807.
– *P. anal.* *LITT.* Qui appartient à l'époque littéraire qui, en France, correspond aux règnes de Henri IV et Louis XIII.
2. *Plus gén.* [En parlant d'un artiste, ou d'une création artistique] Qui par son style rappelle celui de la période baroque :
• 3. Ses allégories [de J. B. Rousseau], sont jugées tout d'une voix : **baroques**, métaphysiques, sophistiquées, sèches, inextricables, nul défaut n'y manque. Sainte-Beuve, *Portraits littér.*, t. I, 1844-64, p. 136.
• 4. Ariadne auf Naxos, (...), accentue plus encore l'évolution du musicien vers le style **baroque** avec son *maniérisme*, ses *bizarreries*, où la délicatesse, la grâce de Mozart se mêlent à la bouffonnerie débridée de la Commedia Dell'Arte. R. Dumesnil, *Hist. illustrée du théâtre lyrique*, 1953, p. 171.
3. *Emploi subst.*
a) [L'art, le style qui appartiennent à la période baroque] :
• 5. Hier, discussion inutile sur le **Baroque** (...) Toujours cette éternelle confusion entre le **baroque** et le rococo (...) On refuse au vrai **baroque** la gravité religieuse alors qu'elle est très évidente à la chapelle de la Sorbonne ou au Val-de-Grâce; ... Green, *Journal*, 1946-50, p. 142.
• 6. Dès que l'art d'assouvissement point, notre répulsion apparaît. D'où notre admiration du grand **baroque** créateur, de Michel-Ange au Greco, et notre dédain du **baroque** établi; ... Malraux, *Les Voix du silence*, 1951, p. 526.
– [L'art, le style, la manière qui rappellent ceux de l'époque baroque] :
• 7. Le **baroque** de son dessin forcené [de Van Gogh] est plus proche des proues scandinaves ou des plaques scythes que de Rubens... Malraux, *Les Voix du silence*, 1951, p. 576.
b) Artiste dont le style rappelle cette période :
• 8. Il fallut la mort de Péguy pour familiariser Barrès avec l'idée qu'il n'était autre chose qu'un **baroque**... J. et J. Tharaud, *Pour les fidèles de Péguy*, 1928, p. 86.
B.— *JOAILL.* [Se dit d'une perle] Qui est de forme irrégulière, d'une rondeur imparfaite :
• 9. En dehors de son orient et de son lustre attrayants, une perle doit avoir une couleur homogène et une rondeur aussi parfaites que possible; les beaux spécimens se vendent à la pièce (en grains); les perles **baroques** se vendent en sachets au poids. A. et N. Metta, *Les Pierres précieuses*, 1960, p. 120.
– *Emploi subst.* :
• 10. ... contre les verrières étaient suspendus quelques colliers de grenat, d'ambre, de **baroques**, de corail, etc., objets de négoce de Judas le Lapidaire, ... P. Borel, *Champavert*, Dina la belle juive, 1833, p. 134.

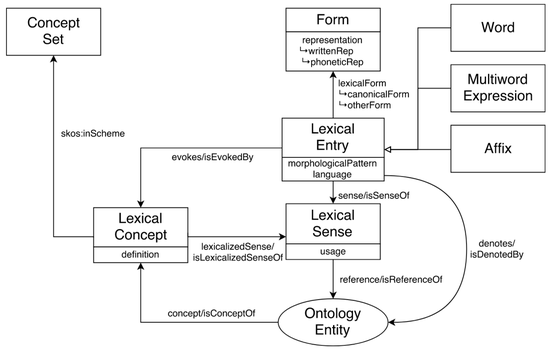
En outre, les sens de chaque entrée peuvent être **représentés dans une hiérarchie** où les sens peuvent avoir des sous-sens pour montrer un sens plus strict. De ce fait, la structure de **chaque entrée lexicale présente une complexité** qui fait obstacle à son intégration dans des applications en TAL utilisant les standards actuels.

Malgré l'utilisation généralisée du TLFi, 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.

Ontolex-Lemon

Le **Web sémantique** représente un moyen efficace de représentation des données et permet aux ordinateurs et aux utilisateurs de récupérer et de partager efficacement des informations. OntoLex-Lemon est un modèle basé sur **Lemon – Lexicon Model for Ontologies** [1] 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 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 suivante 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.



Conversion

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.

```
<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 .
```

Actuellement, **32 073** entrées du TLFi sont converties en Ontolex-Lemon.

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.

Références

[1] 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.

Le point d'accès SPARQL

Lancer vos requêtes SPARQL sur le TLFi à `https://tlfi-sparql.atilf.fr/`.