WordNet en XML-HTML

Guy Lapalme
RALI-DIRO, Université de Montréal
C.P. 6128, Succ. Centre-Ville,
Montréal, Québec, Canada H3C 3J7
lapalme@iro.umontreal.ca

Résumé. Nous présentons une version XML des informations du WordNet de Princeton qui conserve toute l'information originale, mais l'organise dans un format plus pratique pour la consultation et l'accès par programme. Ces fichiers XML ont permis de générer un ensemble de fichiers HTML permettant d'explorer les synsets avec un simple navigateur internet. Une application de démonstration Java illustre la facilité d'accès à l'information en XML pour d'autres applications de TAL.

Abstract. This paper describes an XML version of the original Princeton WordNet which keeps all of the original information but in a more effective format for browsing and program access. These XML files were used to generate a set of HTML files to enable a easy and fast browsing of the synsets. A Java application was developed as a demonstration of the access to the XML format from other NLP applications.

Mots-clés : WordNet, XML, Feuilles de transformation XSLT, synset.

Keywords: WordNet, XML, XSLT, StyleSheet Transformation, synset.

1 Se balader dans WordNet

WordNet was designed for use under program control.
George Miller (Miller, 1995, p. 39)

WordNet [1] est le réseau lexical le plus connu pour l'anglais et il sert de modèle de référence pour pratiquement toutes les autres langues. La version 3 de la base de données regroupe 155 287 entrées (c.-à-d. mots ou expressions en anglais) dans 117 659 ensembles de synonymes (appelés *synsets*) eux-mêmes reliés par des relations sémantiques telles la synonymie, l'antonymie, l'hyponymie, l'implication et quelques autres. Cette organisation est inspirée d'une certaine perception de l'organisation des mots dans le cerveau où les mots sont liés entre eux. Il est donc possible de passer d'un concept à un autre à l'aide d'associations d'idées matérialisées par les relations sémantiques indiquées entre les groupes de mots synonymes.

En plus de la grande couverture et de la qualité linguistique des informations qu'on retrouve WordNet, une des raisons de sa popularité est le fait que les auteurs ont décidé dès le départ de rendre leurs données librement disponibles à la communauté dans un format machine bien organisé. En plus des données, les concepteurs de WordNet fournissent plusieurs programmes (pour les plateformes Unix/Linux et Windows) pour fouiller dans la structure interne qui est de niveau relativement bas avec un codage assez particulier. Le fait que ces programmes étaient écrits dans un dialecte de C relativement portable a facilité la création d'interfaces de programmation (API) dans plusieurs langages de programmation [2]. On peut accéder à WordNet sur le web ou avec des applications spécialisées. Le contenu de WordNet a également été transformé vers d'autres formalismes comme Prolog [3], RDF [4] ou OWL [5]. Il y a également une version *Linked Data* [6].

1. http://wordnet.princeton.edu
2. http://wordnet.princeton.edu/wordnet/related-projects/
3. http://wordnet.princeton.edu/wordnet/man/prologdb.5WN.html
4. http://semanticweb.cs.vu.nl/lod/wn30/
5. http://www.w3.org/TR/wordnet-rdf/
6. http://datahub.io/dataset/w3c-wordnet

Même si le format interne *orienté bytes* est pratique pour les machines, il est malaisé pour un humain de s'y retrouver. En fait, ces fichiers sont produits par programme à partir de *fichiers de lexicographes* dont le format est également assez cryptique.

1.1 Version XML

Comme XML est maintenant un format de données bien répandu, nous avons pensé qu'il serait intéressant de produire des versions XML de la base de données WordNet avec des schémas pour les valider. Un avantage important de XML est le fait que les données peuvent se décrire elles-mêmes si des noms d'étiquettes appropriés sont choisis. Ceci permet à un humain de s'y retrouver plus facilement que la liste de caractères ou codes hexadécimaux utilisés dans la distribution de WordNet. Comme pratiquement tous les langages de programmation disposent d'analyseurs efficaces de XML, nous pensons que ce format sera également plus pratique pour les machines que les formats originaux.

C'est pourquoi nous avons développé une feuille de transformation XSLT pour transformer les fichiers `data` et `code` du répertoire `dict` de la distribution de WordNet en un ensemble de fichiers XML équivalents et validés avec un Schema XML. La transformation conserve en XML toute l'information des fichiers originaux de WordNet.

Dans la version originale de WordNet, chaque synset possède un numéro d'identification (un nombre hexadécimal correspondant au nombre d'octets depuis le début du fichier) qui est utilisé pour accéder efficacement à chaque synset. Dans la version XML, ce nombre est conservé, mais il est précédé par une lettre indiquant la partie du discours afin de créer un identificateur unique de type `xsd:ID` pour les synsets. Les applications peuvent ensuite utiliser ces identificateurs pour accéder directement aux synsets. Ce choix arbitraire, comme le serait tout identificateur de ce type, permet une correspondance facile avec les synsets dans les fichiers originaux ; un choix analogue est fait dans d'autres versions XML de WordNet (certaines sont présentées en section 3) qui gardent aussi des références à ces identificateurs dans leur noms de synset.

Chaque ligne d'un fichier `data` correspond à un élément XML définissant un synset. Il y a quatre fichiers fichiers XML `data.{adj,adv,noun,verb}.xml` correspondant à chaque fichier original `data.{adj,adv,noun,verb}` du répertoire `dict`. Ces quatre fichiers sont ensuite *inclus* dans un fichier maître pour former un seul fichier XML pour des traitements par programme ou avec une autre feuille de style. Le fichier maître est également validable avec un schéma XML.

Chaque synset est représenté par un élément XML avec le contenu suivant, décrit formellement avec un schéma XML :
- deux attributs : `id`, un identificateur unique basé sur le numéro original du synset, et `type` qui indique la partie du discours (nom, verbe, adjectif ou adverbe) ;
- des éléments enfants : `word` indique les mots en anglais des membres du synset ; `pointer` fait des références à d'autres synsets, son contenu indique le type de référence tel que *hyperonyme*, *hyponyme*, *semblable à*, ... ; `frame` écrit des contextes d'utilisation ; `def` donne des définitions ; `example` donne des exemples de mots du synset.

La partie gauche de la figure 1 présente ces éléments sous forme d'une page HTML : les éléments `word` sont en gras suivis de leur type ; suivent dans le tableau, le `type` et le champ `def` ; les liens *sortant* du synset sont regroupés par type. Le bas de la figure donne la forme XML qui est en correspondance directe avec les données originales. La version HTML indique plutôt les mots qui forment le synset plutôt que leur identificateur. Les liens entrant dans ce synset sont également indiqués de la même manière, ils sont calculés par la feuille de style qui produit ces pages web, car cette information n'apparaît pas dans les données initiales.

Afin de faciliter l'accès au synset à partir de mots anglais, la distribution de WordNet fournit aussi des fichiers index (un par partie du discours) donnant la liste des synsets où apparaissent un mot ou une expression en anglais. La feuille de style produit également les index XML correspondants également validés avec un schéma XML. Un traitement semblable a été appliqué pour les listes d'exceptions.

1.1.1 Naviguer dans les synsets

À partir du format XML décrit dans la section précédente, une autre feuille de style XSLT a été écrite pour générer plus de 117 000 fichiers HTML (un par synset, un exemple est donné dans la partie gauche de la figure 1). En suivant les liens entrant et sortant des synsets avec un simple fureteur internet, on peut explorer rapidement et facilement les autres synsets. Le texte affiché pour chaque lien donne la liste des mots du synset cible et une infobulle donne sa définition. Ce mode d'exploration est beaucoup plus pratique que le simple numéro de synset normalement disponible. Le réseau compte plus

de 377 000 liens entre les synsets. Au bas de chaque page HTML, il est possible de consulter la version XML du synset en cliquant sur la dernière ligne.

La feuille de style produit également une page d'accueil [7] permettant de trouver l'ensemble des mots débutant par une lettre ainsi qu'un champ de texte qui permet d'accéder directement à une entrée spécifique. Aucun traitement morphologique n'est effectué (voir toutefois la section 1.1.2), mais le système offre des suggestions dynamiques en fonction du début du mot déjà tapé. Tous ces mécanismes de recherche sont effectués en Javascript dans le navigateur internet, sans aucun traitement particulier du côté du serveur.

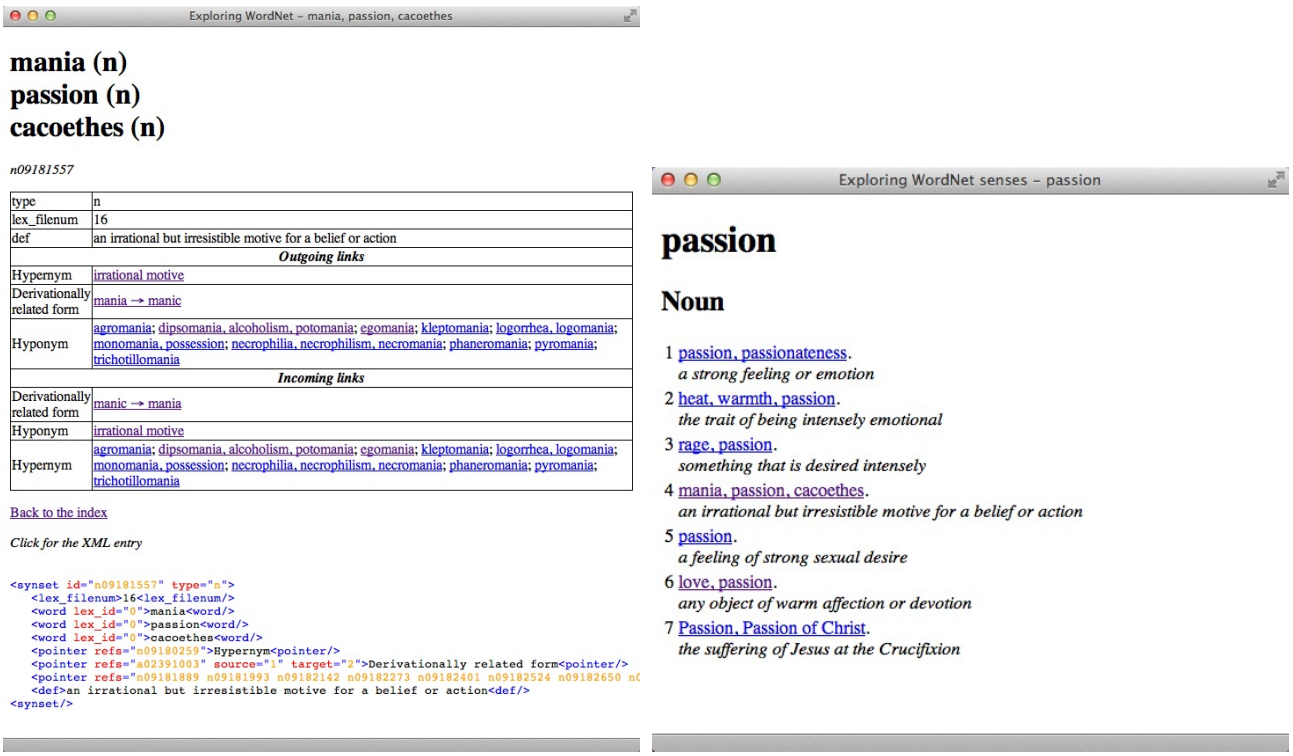

FIGURE 1 – À gauche : affichage du synset 091917557 correspondant aux noms : `mania`, `passion`, `cacoethes`. À droite : Affichage des sens associés au mot `passion`, c.-à-d. les synsets dans lesquels `passion` apparaît.

1.1.2 Naviguer dans les sens d'un mot

Il est également possible de trouver tous les sens d'un mot de WordNet à partir d'un mot ou d'une expression [8], comme le permet un formulaire sur le site de WordNet. La partie droite de la figure 1 montre une sortie de ce type de recherche pour le mot `passion`. L'analyse de la morphologie est effectuée sur les mots de la requête avec une version Javascript du programme Morphy [9] du site de WordNet.

1.2 Programmation avec les fichiers XML

Afin de démontrer la possibilité d'utilisation par programme de ces fichiers XML, nous avons développé `Jwn`, une version Java du programme `wn` [10] très souvent utilisé pour accéder aux données originales de WordNet. La grande majorité des options de `wn` ont été implantées en utilisant la même syntaxe d'appel de ligne de commande, si ce n'est qu'aucun traitement morphologique n'est appliqué sur la chaîne d'entrée. Ce programme doit plutôt être considéré comme un exemple d'utilisation que comme un outil de production.

7. `http://rali.iro.umontreal.ca/WordNet-XML/HTML/start.html`
8. `http://rali.iro.umontreal.ca/WordNet-XML/HTML/start-senses.html`
9. `http://wordnetcode.princeton.edu/tools/morphy.tgz`
10. `http://wordnet.princeton.edu/man/wn.1WN.html`

2 Accès aux fichiers

On peut récupérer l'ensemble de ces fichiers (librement sans contrainte, dans le sens exprimé par le synset `a01064167`) sur le site Web du RALI à

`http://rali.iro.umontreal.ca/rali/?q=en/wordnet-browsing#data`

3 Travaux connexes

XML a déjà été utilisé pour l'organisation de plusieurs dérivés de WordNet pour d'autres langues que l'anglais. Notamment pour l'allemand, GermaNet (Henrich & Hinrichs, 2010) [11] a été balisé à l'aide du standard Lexical Markup Framework qui est un formalisme XML très complet pour les dictionnaires. Bond et Foster (Bond & Foster, 2013) présentent le *Open Multilingual WordNet* (Bond, 2014) une base de données qui réunit avec une licence libre des versions de WordNet dans 22 langues différentes ainsi qu'un module d'interrogation. Les données sont reliées aux synsets originaux. Le format d'échange est celui de lignes dans lesquelles les informations sont séparées par des tabulations. Il est également possible d'obtenir une version RDF-XML ou LMF-XML pour chaque langue, y compris l'anglais. Ce dernier fichier de près de 100 Meg combine toute l'information du WordNet de Princeton, mais s'éloigne de la structure originale des informations. Notre but était de produire une version XML relativement légère, mais qui colle le plus possible à la version originale, en grande partie parce que cet exercice nous a servi à mieux appréhender WordNet.

En français, Benoît Sagot a développé Wolf (WordNet Libre du Français) (Sagot & Fišer, 2008) dont le but était d'associer des mots français aux synsets de l'anglais à l'aide d'une intégration automatique de plusieurs sources d'informations lexicales dans laquelle on retrouve parfois des combinaisons de sens assez surprenantes. Jusqu'à tout récemment, Wolf était présenté dans un format *à la* XML malheureusement *mal formé* au sens XML du terme et sans validation. La dernière version (Sagot, 2014) est grandement améliorée et validée avec une DTD ce qui en facilite l'analyse par programme, mais plus de la moitié des synsets n'ont pas d'équivalent français. Le lien entre les mots des synsets anglais et français est fait par des identificateurs de synsets du WordNet original. Une version XML du WordNet original est d'autant plus intéressante, car elle permet de combiner les informations des deux sources avec un même formalisme.

4 Conclusion

Nous avons décrit une version XML du WordNet original et son utilisation pour créer un réseau de pages HTML permettant une exploration rapide de WordNet avec un navigateur sans nécessiter d'application spécifique. Il est également possible d'utiliser le fichier XML pour l'intégrer à des applications de TAL.

Références

BOND F. (2014). Open Multilingual Wordnet Documentation. http ://compling.hss.ntu.edu.sg/omw/doc.html.

BOND F. & FOSTER R. (2013). Linking and Extending an Open Multilingual Wordnet. In *51st Annual Meeting of the Association for Computational Linguistics : ACL-2013*, p. 1353–1362, Sofia, Bulgaria.

HENRICH V. & HINRICHS E. (2010). Standardizing wordnets in the iso standard lmf : Wordnet-lmf for germanet. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, p. 456–464, Beijing.

MILLER G. A. (1995). WordNet : A lexical database for english. *Communications of the ACM*, **38**(11), 39–41.

SAGOT B. (2014). Les lexiques morphologiques et syntaxiques alexina et le wordnet libre du français. In N. GALA & M. ZOCK, Eds., *Construction de ressources lexicales pour le traitement automatique des langues*, p. 217–254. John Benjamins Publishing Company.

SAGOT B. & FIŠER D. (2008). Construction d'un wordnet libre du français à partir de ressources multilingues. In *TALN 2008*, Avignon.

11. `http://www.sfs.uni-tuebingen.de/lsd/index.shtml`