Construction d'un dictionnaire multilingue de biodiversité à partir de dires d'experts

Mamadou Dieye*, Mohamed Rafik Doulache*, Mustapha Floussi*, Julie Chabalier , Isabelle Mougenot *,***, Mathieu Roche *,*****,**

**Université Montpellier 2*
*** Natural Solutions, Marseille*
**** Espace-Dev, IRD, Montpellier*
***** LIRMM, CNRS, Montpellier*

RÉSUMÉ. *Nous présentons les premières phases d'un travail orienté vers la construction d'un thésaurus trilingue dédié à la biodiversité en méditerranée. Notre démarche se veut résolument synthétique et exploite à posteriori les dires d'experts accessibles depuis les publications scientifiques faisant référence dans le domaine. Les termes clés, qui sont à la base du socle conceptuel nécessaire à une meilleure compréhension de la biodiversité en méditerranée, sont acquis au travers d'une approche rigoureuse de fouille de texte qui est détaillée ici. Les premiers résultats obtenus sont rapportés et évalués. Les experts du domaine seront ensuite sollicités pour aider à articuler les termes au travers de structures hiérarchiques et ainsi retranscrire leur savoir sous la forme d'un thésaurus dédié à la biodiversité dans le bassin méditerranéen.*

MOTS-CLÉS : *thésaurus, biodiversité, fouille de texte*

ABSTRACT. *We describe the first steps towards developing a methodology for building a trilingual thesaurus that is dedicated to biodiversity in the Mediterranean region. A key property of our synthetic approach is to discover expert domain knowledge by means of literature-based information extraction techniques. Thus, a highly relevant set of terms, which provides the foundation for a better understanding of biodiversity in a Mediterranean context, is acquired through a text mining method that is outlined in the manuscript. The first results are reported and evaluated. Experts in ecology and biodiversity will be invited in the near future to help improve the design of the terms through hierarchical taxonomies and thus inject their in-depth knowledge in form of a thesaurus dedicated to biodiversity in the Mediterranean region.*

KEYWORDS: *thesaurus, biodiversity, text mining*

1. Introduction

Un projet de construction d'un glossaire trilingue (français, anglais, arabe) des termes de l'écologie, financé par la Banque Mondiale et le Conservatoire du Littoral vient de débuter. Un premier outil de type Wiki a d'ores et déjà permis à des scientifiques du domaine de dégager une cinquantaine de termes clés du domaine. Le travail que nous proposons consiste à extraire des termes consacrés à l'écologie au travers d'une approche de fouille de texte. L'organisation entre les termes du glossaire est également au centre de nos préoccupations dans le cadre de ce projet global. Les experts du domaine seront alors sollicités afin de matérialiser leur savoir sous forme de relations entre termes clés de la biodiversité au travers de liens de généralisation/spécialisation, de voisinage ou encore de synonymie. A cet effet, un environnement dédié viendra faciliter la construction collaborative d'un thésaurus trilingue à partir du glossaire en cours d'acquisition.

Cet article décrit la première phase de ce projet qui consiste à proposer un processus d'extraction de la terminologie à partir d'un corpus anglais afin de constituer un glossaire de la biodiversité. Dans ce contexte, une approche de fouille de texte composée de quatre étapes (conversion des fichiers PDF, normalisation, étiquetage morphosyntaxique, extraction de la terminologie) est appliquée (cf. Figure 1). Chaque étape est décrite en section 2. Une fois la terminologie acquise, un transfert lexical à partir des termes sources (en anglais) vers des termes cibles (en français) sera effectué. Dans ce cadre, un processus de fouille du Web est proposé. Celui-ci est résumé en section 3.

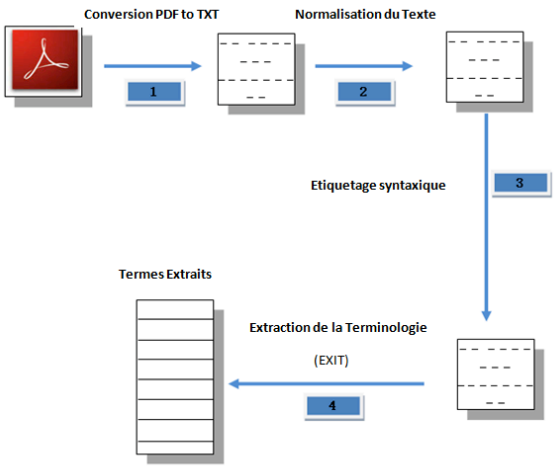

Figure 1 : Processus de Fouille de Texte

2. Processus de Fouille de Texte

Le processus de fouille de texte afin d'extraire la terminologie à partir d'un fichier PDF est décrit dans les sous-sections suivantes.

2.1 Prétraitement des données (étapes 1 et 2)

La première étape du processus consiste à convertir les fichiers PDF en données textuelles. Pour cette tâche, plusieurs outils peuvent être appliqués (cf. Tableau 1). Notre travail d'analyse a permis de mettre en relief le bon comportement de *Simpo PDF converter*[1], en particulier grâce à la restitution scrupuleuse de la forme originale des fichiers PDF.

Caractéristiques / Outils	**Adobe Reader**	**Foxit Reader**	**Zilla PDF**	**Simpo PDF to text**
Libre	Oui	Oui	Oui	Oui
Conversion par lots	Non	Non	Oui	Oui
Conservation de la mise en page	Non	Non	Non	Oui

Tableau 1 : Comparaison des performances des outils de conversion sur la base de trois critères cruciaux pour le projet.

L'étape suivante consiste à normaliser les fichiers textuels afin qu'ils soient adaptés à la phase d'étiquetage morphosyntaxique. Ce traitement consiste, entre autres, à insérer un espace entre tous les mots lorsqu'ils sont suivis d'un signe de ponctuation. Ceci permet de distinguer les étiquettes propres aux mots par rapport aux étiquettes des ponctuations.

2.2 Etiquetage grammatical (étape 3)

L'étiquetage est le processus qui consiste à associer aux mots d'un texte une fonction grammaticale (nom, verbe, etc.) en s'appuyant sur des informations lexicales et contextuelles. Pour une telle tâche, nous appliquons l'étiqueteur de Brill (Brill 1994). Un exemple d'étiquetage à partir de données réelles issues de la thématique de la biodiversité est donné ci-dessous :

- Exemple *avant étiquetage* :

```
The processes of domestication of plant and animal ...
```

[1] http://www.simpopdf.com/

- Exemple *après étiquetage* avec les étiquettes DT (article), NN et NNS (noms au singulier et pluriel), IN (préposition), CC (conjonction de coordination):
The/DT processes/NNS of/IN domestication/NN of/IN plant/NN and/CC animal/NN ...

Outre un lexique constitué à partir du *Wall Street Journal*, l'étiqueteur de Brill utilise deux types de règles :

- *Règles lexicales* : ces règles permettent de définir l'étiquette du mot en s'appuyant sur ses propriétés lexicales (par exemple, les informations liées aux suffixes et/ou préfixes).

- *Règles contextuelles* : ce type de règles permet d'affiner l'étiquetage, c'est-à-dire de revenir sur les étiquettes précédemment affectées et de les corriger en examinant le contexte local.

L'étiquetage s'effectue en deux étapes. Durant la première étape, chaque mot du texte reçoit l'étiquette la plus probable dans le contexte considéré, soit par consultation du lexique si le mot est connu, soit par application des règles lexicales si le mot est inconnu au lexique. Pendant la seconde étape, le système revient sur ces premières affectations, examine le contexte local et corrige éventuellement les étiquettes précédemment affectées à l'aide des règles contextuelles (par exemple, « les mots suivis du modal *can* sont étiquetés comme des verbes »).

Une fois l'étiquetage grammatical effectué, l'étape suivante consiste à extraire la terminologie. Dans ce contexte, nous nous sommes concentrés sur l'extraction de la terminologie nominale, c'est-à-dire les groupes de mots pertinents au regard de la thématique du projet (Bourigault et Jacquemin, 1999).

2.3 Extraction de la terminologie (étape 4)

Dans nos travaux, nous utilisons le système EXIT (cf. Figure 2) qui est conçu pour extraire la terminologie à partir d'un corpus étiqueté (Roche *et al.* 2004). Ce logiciel est destiné à des utilisateurs experts d'un domaine. Quelques connaissances en linguistique de base et dans la création d'expressions régulières peuvent néanmoins être nécessaires pour modifier les fichiers d'options (notamment pour concevoir et/ou modifier des patrons d'extraction).

Lors de l'extraction des termes, EXIT propose des couples ou triplets d'unités prédéfinis (par exemple, des termes de type Nom-Nom, Adjectif-Nom, Nom-Préposition-Nom, etc.). Notons qu'outre ces paramètres, l'utilisateur peut appliquer des filtres statistiques (Information Mutuelle, Information Mutuelle au Cube, Rapport de Vraisemblance) qui permettent de présenter les termes les plus pertinents en tête de liste. Par ailleurs ce système peut appliquer un processus itératif afin de construire des termes plus complexes (composés de plusieurs mots).

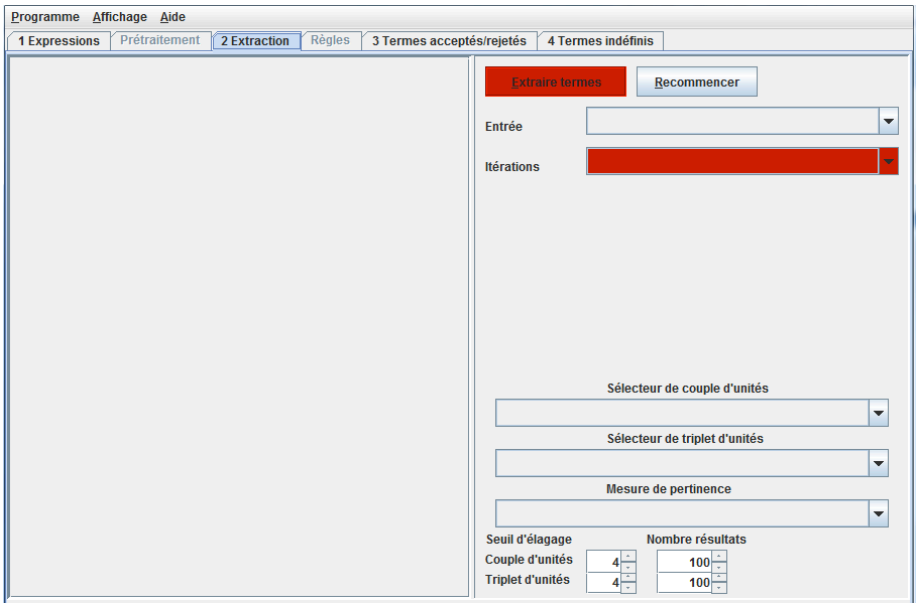

Figure 2 : Capture d'écran du logiciel EXIT

Le tableau 2 donne un exemple de termes pertinents, non pertinents, et non évalués par manque de connaissance experte (JNSP : « je ne sais pas »). L'évaluation a été effectuée par *consensus* des trois premiers auteurs de cet article.

Par ailleurs, le tableau 2 présente la répartition des évaluations de près de 200 candidats termes obtenus à partir d'un article scientifique lié à la biodiversité Méditerranéenne[2]. Ceci montre que 57% des termes ont été déclarés comme pertinents par rapport à cette thématique alors que 29% ne

[2] Jacques Blondel , The 'Design' of Mediterranean Landscapes: A Millennial Story of Humans and Ecological Systems during the Historic Period, Hum Ecol (2006) 34:713–729

le sont pas. Notons que de nombreuses erreurs sont dues à des problèmes de nettoyage lors de la normalisation du texte original. Néanmoins la quantité de termes pertinents se révèle tout à fait satisfaisante et constitue un élément de base pour la construction d'un dictionnaire spécialisé en biodiversité.

Termes pertinents (57%)	Termes non pertinents (29%)	JNSP (14%)
cumulative degradation	Lost Eden	adaptive variation
human action	second school	complex coevolution
human systems	Eden theory	refugia formation
Mediterranean Basin	David Attenborough	turbance event
Mediterranean landscapes	viewpoint stresses	scientists does
animal species	certainly lies	action can
feedback cycles	not acknowledge	resilience indicates

Tableau 2 : Extrait de termes obtenus après application du processus complet de fouille de texte.

3. Perspectives : transfert lexical des termes

Dans nos futurs travaux, nous proposerons une méthode de fouille du Web qui permettra de déterminer des termes en français représentant une traduction des termes extraits dans les textes en anglais (par exemple, le terme *Mediterranean landscapes* donné en section 2.3). Le transfert lexical d'une langue à une autre est un problème difficile particulièrement dans les domaines de spécialité (Claveau, 2008).

Notre approche s'appuie, entre autres, sur une méthode de fouille du Web et sur l'algorithme PMI-IR (Pointwise Mutual Information and Information Retrieval) de (Turney, 2001). PMI-IR consiste à interroger le Web via un moteur de recherche pour déterminer des synonymes appropriés. A partir d'un terme donné noté *mot*, l'objectif de PMI-IR est de choisir un synonyme parmi une liste donnée. Ainsi, le but est de calculer, pour chaque mot, le synonyme $choix_i$ qui donne le meilleur score. Pour ce faire, l'algorithme PMI-IR utilise, de la même manière que nos travaux, différentes mesures statistiques fondées sur la proportion de documents dans lesquels les deux termes sont présents (par exemple, l'Information Mutuelle ou la mesure de Dice). Plus la proportion de documents contenant ces deux mots est importante (par exemple, dans une fenêtre donnée) et plus *mot* et $choix_i$ sont considérés comme synonymes.

L'approche que nous souhaitons adopter dans ce projet se décline de la manière suivante :

- ***Extraction des candidats à la traduction*** **:** Dans un premier temps, nous interrogeons un moteur de recherche (par exemple, *Google*) pour rassembler les pages Web contenant un terme à traduire (par exemple, *Mediterranean landscapes*). Nous cherchons alors les pages contenant le terme à traduire et certains marqueurs paralinguistiques comme les parenthèses afin d'en extraire le contenu. Ce dernier peut se révéler un excellent candidat à la traduction après avoir vérifié de manière automatique qu'il représente une expression écrite en français. L'étape suivante consiste à nettoyer les données textuelles extraites en supprimant des données représentant du bruit (signes de ponctuation, marqueurs linguistiques, etc.). Ceci permet d'obtenir un certain nombre de candidats normalisés. Par exemple avec l'expression *Mediterranean landscapes* à traduire, nous pouvons obtenir le candidat *Paysages Méditerranéens*. Notons que ce principe a déjà été adopté avec succès dans un processus de traduction de termes liés au domaine informatique.

- ***Classement des candidats à la traduction :*** Dans un second temps, nous allons appliquer des méthodes de fouille du Web sur la base de l'approche de Peter Turney décrite précédemment afin de classer les termes candidats à la traduction. Les mesures que nous souhaitons utiliser calculent une forme de dépendance entre un terme à traduire (*exp*) et les candidats proposés à l'étape précédente (*cand*). Cette dépendance peut par exemple s'appuyer sur la mesure de Dice où *nb*(*exp*) représente le nombre de pages retournées par un moteur de recherche avec la requête *exp*:

  $$Dice(exp,cand) = 2 \times nb(exp, cand) / (nb(exp) + nb(cand))$$

  Notons que *nb*(*exp*, *cand*) peut représenter le nombre de pages où les termes *exp* et *cand* sont présents dans la même page (dépendance *souple*) ou strictement voisins (dépendance *stricte*).

  Des expérimentations sur des données réelles liées au domaine informatique avec la dépendance stricte ont montré que le candidat proposé est pertinent dans 83% des premiers cas proposés par notre système (sur la base de 358 couples termes/candidats). Nous souhaitons maintenant expérimenter notre approche avec les termes spécifiquement liés à la biodiversité que nous avons obtenus dans la première partie du projet.

4. Conclusion et Perspectives

L'objectif de notre projet porte sur la construction d'un glossaire trilingue en Méditerranée. Il consiste, dans un premier temps, à extraire d'un fichier PDF des termes standards liés à la biodiversité. Pour parvenir à ce résultat, nous avons mis en place plusieurs étapes successives de traitements. La première étape consiste à mener une analyse rigoureuse et méthodique des différents outils de conversion d'un fichier PDF en un fichier textuel de bonne facture. Les deuxième et troisième étapes concernent respectivement la normalisation et l'étiquetage grammatical des données textuelles. Enfin, la dernière étape a trait à l'extraction de la terminologie.

Dans la suite du travail, nous souhaitons focaliser notre recherche sur les méthodes de traduction qui pourront s'appuyer sur des approches pointues de fouille du Web (Roche et Prince, 2010) mais aussi sur des méthodes d'alignement de corpus (Och et Ney, 2004). A plus long terme, il nous faudra prendre en compte la traduction en arabe qui se révèlera plus complexe compte tenu des spécificités de cette langue.

5. Bibliographie

Bourigault D., Jacquemin C. Term Extraction + Term Clustering: An Integrated Platform for Computer-Aided Terminology. Proceedings of the European Chapter of the Association for Computational Linguistics (EACL'99), p.15-22, 1999

Brill E.. Some advances in transformation-based part of speech tagging. In Proc. of AAAI, volume 1, pages 722–727, 1994.

Claveau V. Automatic Translation of Biomedical Terms by Supervised Machine Learning. Proceedings of the Language Resources Conference (LREC), 2008

Och, F.J., Ney H. The Alignment Template Approach to Statistical Machine Translation. Computational Linguistics. Vol 30 n°4, 417-449, 2004

Roche M., Heitz T., Matte-Tailliez O., Kodratoff Y. EXIT: Un système itératif pour l'extraction de la terminologie du domaine à partir de corpus spécialisés. Dans les Actes des JADT, Volume 2, p946-956, 2004

Roche M, Prince V. A Web-Mining Approach to Disambiguate Biomedical Acronym Expansions. Informatica, Vol 34, No2, p243-253, 2010

Turney P. Mining the Web for synonyms: PMI–IR versus LSA on TOEFL. In Proc. of ECML, pages 491–502, 2001.Kolski C., *Interfaces homme-machine*, Paris, Hermès, 1997