How2Say: Une exploration interactive terminologique des corpus multilingues

Jacques Guyot[1], Gilles Falquet[2] , Nizar Goulha[1]

[1] Fondation Olanto - Genève.[2] University of Geneva - Centre Universitaire d'Informatique
jacques@olanto.org, gilles.falquet@unige.ch, nizar@olanto.org

Résumé

La terminologie est largement dépendante du contexte d'utilisation et varie donc d'une organisation à une autre. Nous présentons un outil qui facilite l'exploration terminologique. A partir d'un terme, il permet de chercher interactivement des traductions, les termes composés, des synonymes et éventuellement d'autres termes qui lui sont associés. Cette recherche se fait directement sur le corpus des documents produits par l'organisation avec une interface simplifiée pour le terminologue.

1 Introduction

La fondation Olanto met à disposition des logiciels en licence libre. Ces logiciels sont élaborés avec l'expertise des traducteurs et celle des chercheurs universitaires. Olanto élabore une suite d'outils pour couvrir le domaine de la TAO. Actuellement, elle dispose d'un concordancier (myCAT) incluant un outil de référencement, d'un module de traduction automatique (myMT[1]), d'un aligneur pour préparer des TMX à partir de corpus multilingues et finalement d'un gestionnaire de terminologie (myTERM). Ces logiciels sont développés pour répondre à des organisations dont les corpus se chiffrent en centaines de milliers de documents et dont les langues peuvent multiples. Les ressources générées par l'alignement des documents produisent des millions de phrases alignées. Dans un article (Guyot, 2014), nous avons montré comment ces ressources pouvaient être utilisées pour éliminer les incohérences produites par l'application de la transitivité appliquée sur deux lexiques bilingues pour produire un nouveau lexique bilingue (par exemple : FR-EN et EN-DE → FR-DE).

Entrée en anglais	Entrée en français	Expression suggérée par How2Say	Expression traduite automatiquement par myMT
chain of causation	chaîne causale	lien de causalité	lien de causalité
ex gratia payment	paiement à titre gracieux	versement à titre gracieux	versement à titre gracieux
act of God	cas de force majeure; cas fortuit; [on dit aussi parfois] acte de Dieu	acte de Dieu	acte de Dieu

Table 1: comparaison entre la recommandation du dictionnaire et l'usage dans le corpus.

Lors de la mise au point de la mesure de cooccurrence des termes, nous avons utilisé un dictionnaire juridique Anglais-Français (ONUG, 200) mis à disposition par l'ONU avec le corpus MULTI-UN (Eisele, 2010). Il est rapidement apparu que certaines traductions du dictionnaire n'étaient pas celles utilisées dans le corpus (voir table 1).

Bien qu'une majorité des termes soient bien ceux proposés par le dictionnaire, il semble important que le traducteur ou le terminologue puisse vérifier l'usage réel des traductions dans le corpus. De plus, les processus utilisés permettent aussi de mettre en évidence les termes composés à partir du terme recherché et parfois de trouver des synonymes ou des termes associés. Il est aussi important que l'utilisateur puisse avoir accès à ces informations de la façon la plus simple et interactivement.

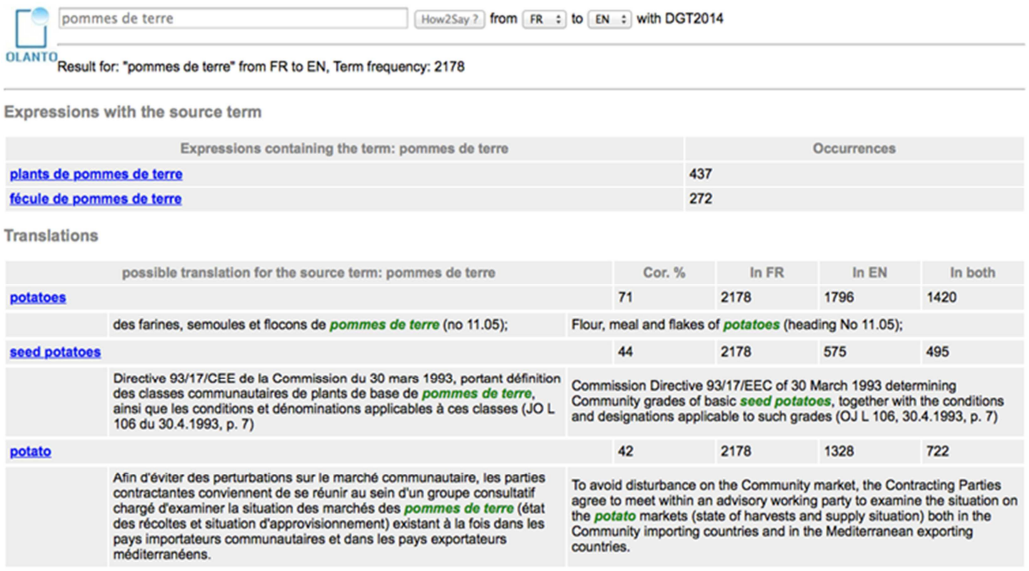

Figure 1: Interface Utilisateur de How2Say, http://olanto.org/fr/logiciels/how2say/demo

2 Principes de fonctionnement

Le corpus est composé des phases alignées dans toutes les langues disponibles. Ce corpus est d'abord indexés (les « stopword » sont retirés). Les associations entre les phrases ainsi que les langues sont conservées. Pour une recherche portant sur le terme *w*, dans la langue source *so* et pour la langue cible **ta**, nous avons le processus suivant (voir figure2):

1. Avec l'index, nous cherchons les phrases contenant *w* dans la langue *so* que l'on note *result(w,so)***.**
2. Avec les liens de traduction que l'on a conservés, on peut trouver *pour result(w,so)*, les phrases qui sont lui associées pour la langue *ta* que l'on notera *trad(w,so,ta)*. On a donc un concordancier entre *so* et *ta* pour les phrases contenant *w*.
3. La recherche des termes composés contenant *w* est faite en calculant les n-gram de *result(w,so)*. En les ordonnant par fréquence et en conservant que ceux qui contiennent *w*, on dresse facilement une liste des candidats des termes composés.
4. La recherche de la traduction du terme *w* pour la langue *ta* est similaire dans sa première étape. On cherche les n-gram de *trad(w,so,ta)* en les ordonnant par fréquence. On ne peut à priori en éliminer aucun. La deuxième étape est de calculer pour chaque n-gram sa corrélation avec *w* (Guyot, 2014). Pour chaque terme *v* de la liste, on cherche *result(v,ta)* avec l'index. la corrélation est égale à $r_{wv} = \frac{nn_{wv} - n_w n_v}{\sqrt{nn_w - {n_w}^2}\sqrt{nn_v - {n_v}^2}}$ où *n* est la taille du corpus, n_w est la taille de *result(w,so)*, n_v

est la taille de *result(v,ta)* et n_{wv} est taille de l'intersection entre *result(w,so)* et *result(v,ta).* Les résultats sont triés en fonction de la corrélation et présenté à l'utilisateur.

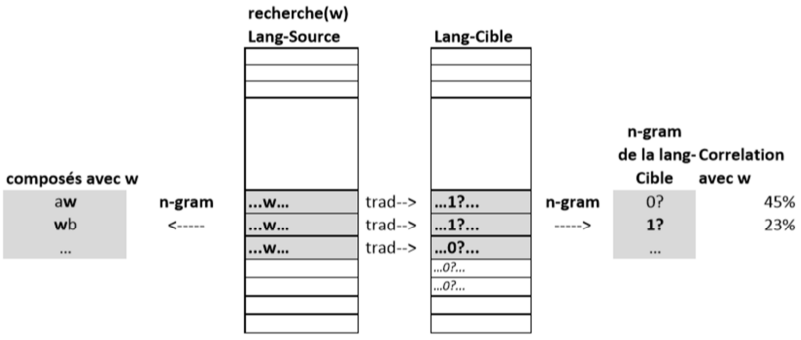

Figure 2: processus utilisés par l'explorateur terminologique

How2Say n'est pas un extracteur terminologique. En effet, il n'analyse pas un document pour en extraire les termes, mais il travaille à partir d'un terme sur l'ensemble du corpus. Il est aussi intéressant de noter dans la table 1, que le système de traduction automatique est en accord avec les propositions faites par l'explorateur de terminologie. L'avantage de l'explorateur est sa simplicité de mise en œuvre par rapport à la SMT. Il ne nécessite pas d'entrainement par pair de langues et par direction. Par contre, il est limité à la traduction de termes. Ceci permet de construire facilement, pour toutes les pairs de langues, un traducteur terminologique (si le corpus le permet). Dans le cas du DGT-2014[ii], Nous avons virtuellement pour les 24 langues, 552 traducteurs qui traduisent, par exemple, le terme polonais **ziemniaków** en grec **γεωμήλων** (ou **πατάτες**).

Une autre difficulté rencontrée est le maintien d'un temps de réponse raisonnable (<20sec pour les cas les plus complexes) avec des corpus de plusieurs dizaines de millions de phrases sans gaspiller les ressources. Malgré la parallélisation des processus, nous avons dû limiter les tailles des résultats retournés par l'index.

Le calcul de la corrélation peut aussi s'appliquer pour la recherche de termes utilisés dans le même contexte (la langue source étant égale à la cible). Le système trouve que *munition* et corrélé avec *arme.* Cette possibilité est encore limitée car le paramétrage est focalisé sur la traduction. Mais nous allons élaborer une nouvelle version orientée vers cette tâche ainsi qu'un Web Service pour que l'on puisse exploiter plus systématique l'outil. Ceci permettra de cartographier un corpus avec un réseau terminologique.

Références

Guyot J., Ghoula N., Falquet G. (2014) Terminology management revisited. Proc of ASLIB 2014, November 2014

Eisele A., Chen Y. (2010) MultiUN: A Multilingual corpus from United Nation Documents LREC 2010

ONUG (2000) - Vocabulaire juridique bilingue , droit (et CDI) (E-F), 2000 TERM/49

[i] myMT utilise le moteur de traduction statistique Moses voir : http://www.statmt.org/moses/

[ii] https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory