

TransSearch : un moteur de recherche de traductions

Julien Bourdaillet Fabrizio Gotti Stéphane Huet
Philippe Langlais Guy Lapalme

RALI - DIRO - Université de Montréal
C.P. 6128, succursale centre-ville
H3C 3J7, Montréal, Québec, Canada

{bourdaij, gottif, huetstep, felipe, lapalme}@iro.umontreal.ca

Résumé. Malgré les nombreuses études visant à améliorer la traduction automatique, la traduction assistée par ordinateur reste la solution préférée des traducteurs lorsqu’une sortie de qualité est recherchée. Cette démonstration vise à présenter le moteur de recherche de traductions TransSearch. Cette application commerciale, accessible sur le Web, repose d’une part sur l’exploitation d’un bitexte aligné au niveau des phrases, et d’autre part sur des modèles statistiques d’alignement de mots.

Abstract. Despite the impressive amount of studies devoted to improving the state of the art of machine translation, computer assisted translation tools remain the preferred solution of human translators when publication quality is of concern. This demonstration presents the translation search engine TransSearch. This web-based commercial application relies on a sentence-aligned bitext and a statistical word alignment techniques.

Mots-clés : Traduction automatique statistique, repérage de traductions, alignement de mots, requêtes linguistiques.

Keywords: Statistical machine translation, translation spotting, word alignment, linguistic queries.

1 Introduction

L’application TransSearch développée au RALI et commercialisée par la société Terminotix¹ offre à ses utilisateurs une *mémoire de traduction* interrogeable à travers le Web. De tels outils sont très populaires dans la communauté des traducteurs professionnels et dans le domaine de la traduction assistée par ordinateur (Macklovitch *et al.*, 2008). Dans l’application actuellement commercialisée, lorsqu’un utilisateur soumet une *requête* l’application répond en affichant l’ensemble des paires de phrases source-cible qui contiennent la requête. Il appartient alors à l’utilisateur de fouiller dans ce matériel afin d’identifier la meilleure réponse correspondant à son besoin traductionnel. Nous avons développé au RALI une nouvelle version de cette application qui tire profit de techniques d’alignement statistique de mots à l’état de l’art (Huet *et al.*, 2009; Bourdaillet *et al.*, 2010). Cette version transforme TransSearch en un *moteur de re-*

¹www.terminotix.com

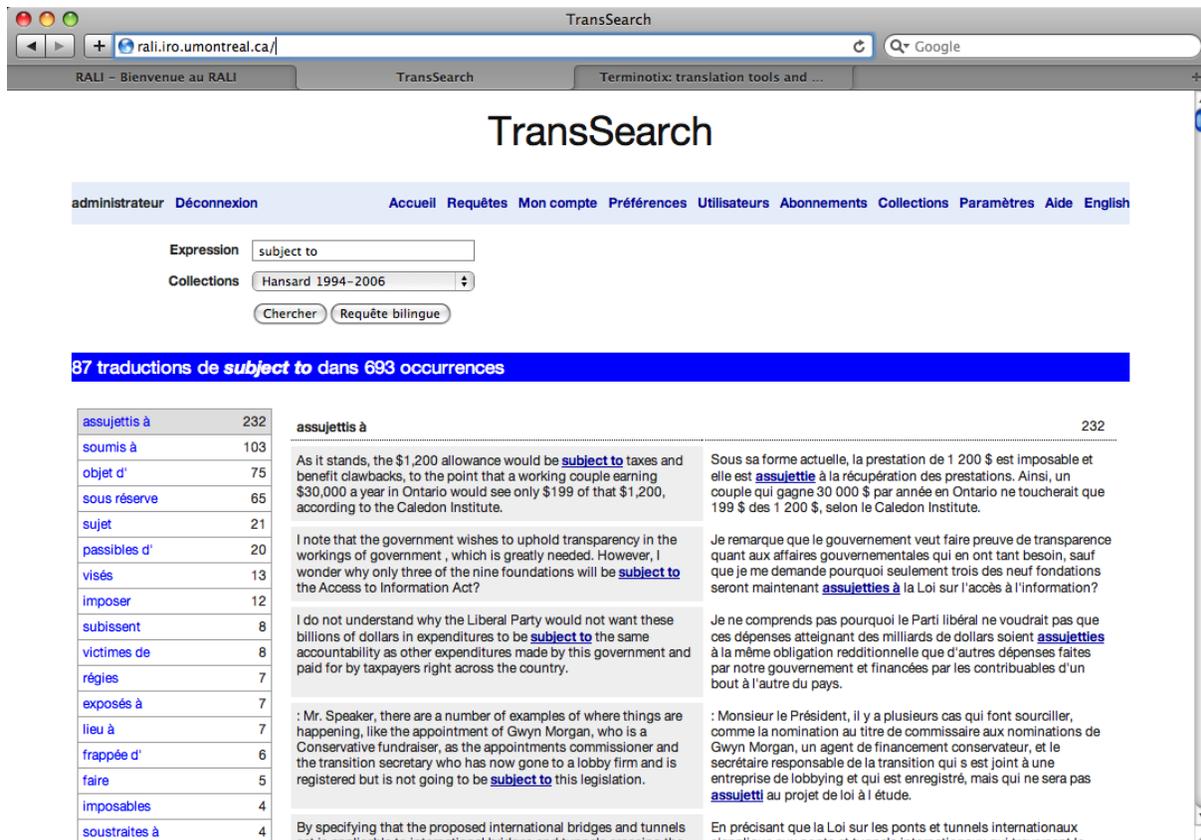


FIG. 1 – Interface utilisateur du moteur de recherche de traductions TransSearch suite à la soumission de la requête `subject to`. Les 87 traductions repérées s’affichent dans la colonne de gauche.

cherche de traductions, offrant ainsi à ses usagers de nouvelles perspectives d’utilisation. Nous présentons ici cette application novatrice qui sera très prochainement commercialisée.

2 Des mots aux requêtes

La figure 1 présente une capture d’écran du nouveau TransSearch. Dans cet exemple, un utilisateur a soumis la requête `subject to` au système. Celui-ci répond alors par une distribution des traductions de la requête (à gauche) automatiquement identifiées par le système. Dans cet exemple, 87 traductions différentes sont repérées dans les textes des débats parlementaires canadiens (Hansards). Chaque traduction est accompagnée de sa fréquence d’association à la requête.

À l’instar du premier système, le nouveau TransSearch fait également office de concordancier bilingue. Pour chacune des traductions d’une requête, il est en effet possible d’accéder à l’ensemble de ses occurrences en corpus. Ainsi dans l’exemple de la figure 1, l’application montre à l’utilisateur les 232 contextes dans lesquels la traduction identifiée `assujettis à` apparaît. L’utilisateur peut bien sûr sélectionner une autre traduction afin d’en examiner les différents contextes d’occurrence.

Il convient de souligner que la distribution des traductions retournée par le nouveau système modifie

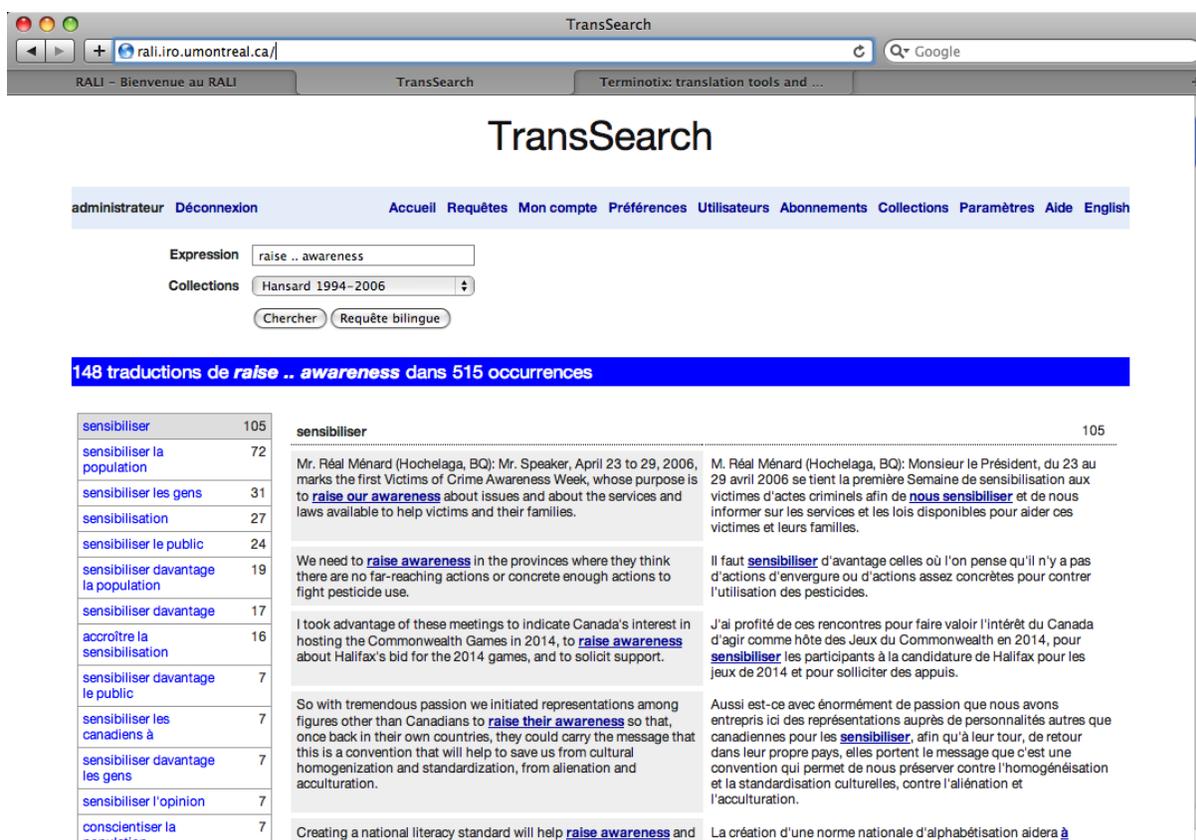


FIG. 2 – Réponse de TransSearch à la requête raise .. awereness. L’ellipse indiquée par .. permet de repérer des occurrences telles que raise our awareness ou raise the public’s awareness.

profondément la nature même de l’application. L’utilisateur peut de manière très synthétique observer les fréquences d’usage de toutes les traductions identifiées, ce qui nécessitait un effort manuel (très) coûteux dans la version actuellement commercialisée.

L’exemple subject to montre que le système est capable de répondre à des requêtes comptant plusieurs mots. L’application est donc plus puissante qu’un simple dictionnaire bilingue où les entrées sont constituées uniquement de mots simples.² Il est également possible de soumettre au système des requêtes plus évoluées, qui autorisent la recherche d’expressions non nécessairement contiguës. Par exemple, la figure 2 présente la réponse du système après la soumission de la requête raise .. awareness. En séparant les mots à rechercher par le symbole .., l’utilisateur spécifie au système que, lors de la recherche, les occurrences des 2 mots de la requête peuvent être distantes d’au plus 5 mots au sein d’une même phrase. Un autre opérateur permettant des recherches sur une distance non limitée est également disponible.

D’un point de vue linguistique, ces opérateurs sont très intéressants puisqu’ils permettent à l’utilisateur de repérer des expressions dont les mots pourraient être séparés par des mots écrans, par exemple their dans l’occurrence raise their awareness. Enfin, un autre type de requête évoluée est accessible à l’utilisateur de TransSearch : les requêtes avec expansion morphologique. La figure 3 présente la réponse du

²Dans certains dictionnaires bilingues, on trouve des expressions de plusieurs mots, mais celles-ci ne sont généralement pas des entrées principales, mais uniquement des exemples d’usage.

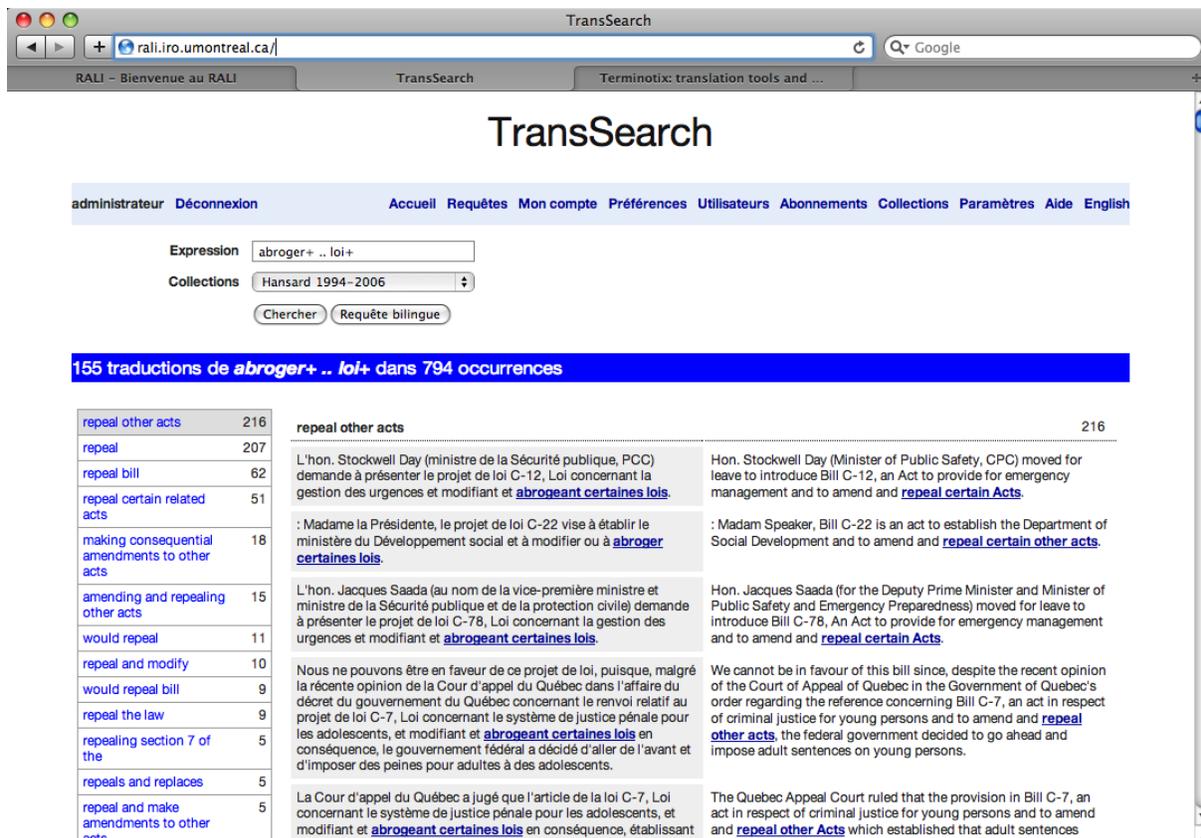


FIG. 3 – Réponse de TransSearch à la requête `abroger+ .. loi+`. L'opérateur `+` permet de repérer les dérivations morphologiques des mots qu'il qualifie, étendant ainsi la portée des recherches effectuées.

système à la requête `abroger+ .. loi+`. Pour les termes auxquels est ajouté le symbole `+`, le système va considérer l'ensemble de leurs dérivations morphologiques existantes dans la langue du terme. Ainsi, à cette requête utilisateur correspondent toutes les expressions contenant une forme fléchie de `abroger`, suivie d'une forme fléchie de `loi`. Cet opérateur spécifiquement linguistique est particulièrement utile pour des langues riches morphologiquement comme le français ou l'espagnol, et permet à l'utilisateur de repérer des expressions sans égard à leur possibles flexions.

Références

- BOURDAILLET J., HUET S., LANGLAIS P. & LAPALME G. (2010). TransSearch : from a bilingual concordancer to a translation finder. *À paraître dans Machine Translation*.
- HUET S., BOURDAILLET J. & LANGLAIS P. (2009). Intégration de l'alignement de mots dans le concordancier bilingue TransSearch. In *16e Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Senlis, Picardie, France.
- MACKLOVITCH E., LAPALME G. & GOTTI F. (2008). TransSearch : What are translators looking for? In *18th Conference of the Association for Machine Translation in the Americas (AMTA)*, p. 412–419, Waikiki, Hawai'i, États-Unis d'Amérique.