

The Bilingual Concordancer TRANSSEARCH

Guy Lapalme, Philippe Langlais, Fabrizio Gotti

RALI - DIRO - Université de Montréal

CP 6128 Succ. Centre-Ville

Montréal, Québec, Canada, H3C 3J7

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

Abstract

TRANSSEARCH is a web-based translation search engine. When a user submits a translation query, the system replies with a set of sentence pairs whose source sentence contains the query. The source expression is highlighted and, with the help of statistical word alignment techniques, the corresponding target expression is also identified. When many sentences share the same translations, the translations are grouped and sorted in decreasing order of frequency to give the translators a variety of different translations whose contexts can be further explored.

1 Introduction

Despite the impressive amount of studies devoted to improving the state of the art in machine translation, translation memories remain the preferred solution for human translators when publication quality is of concern. This demonstration presents the translation search engine TRANSSEARCH. This web-based commercial application, aimed at language professionals, relies on a sentence-aligned bitext and statistical word alignment techniques. It is currently used by more than 3000 professional translators.

TRANSSEARCH is a web-based translation memory, developed by RALI¹ and commercialized by Terminotix², a Montréal-based company specializing in computer-aided translation tools. TRANSSEARCH is a translation search engine very popular among professional translators

(Macklovitch et al., 2000; Macklovitch et al., 2008). In the production version of TRANSSEARCH, when a user submits a translation query, the system replies with the set of sentence pairs whose source sentence contains the query. The translator then has to mine this material to discover the corresponding translation in each target sentence. This demonstration presents a new version of this application that exploits statistical word alignment techniques to turn TRANSSEARCH into a *translation search engine*.

Colloquial expressions, especially adverbial phrases such as *in keeping with*, are difficult to translate correctly because they depend on the context of use and are hard to find in bilingual dictionaries. Translators want to see many contexts of use to determine the most appropriate translation for their case. Bitexts of aligned sentences of previously translated texts are a useful source of inspiration to find solutions to difficult translations problems.

Figure 1 is a screenshot of TRANSSEARCH after the user has submitted the query *in keeping with*. TRANSSEARCH found 668 occurrences of the query for which 169 different translations were discovered by the system in the proceedings of the Canadian parliament (Hansards). The system presents the distribution of the translations of the query on the left. Each translation is presented with its frequency of occurrence. For each translation, the user can consult its contexts of occurrence on the right. For instance, the translation *conforme à* occurred 146 times in the corpus. To see the occurrences of another translation, e.g. *fidèle à* for

¹<http://rali.iro.umontreal.ca>

²<http://www.terminotix.com>

TRANSSEARCH³ BETA TERMINO TIX rali

UTILISATEUR : lapalme REQUÊTES | MON COMPTE | PRÉFÉRENCES | AIDE | QUITTER

Signet / Favori personnalisé : **TransSearch** (qu'est-ce que c'est ?) Requête bilingue

Collection de documents : Les Hansards canadiens

Expression : Chercher

169 traductions de *in keeping with* dans 668 occurrences

conforme à	146	conforme à	146
conformément à	97	This would be in keeping with the administrative process, which has been put in place for the Ukrainian-Canadian community.	Cela serait conforme au processus administratif qui a été mis en place pour la communauté ukraino-canadienne.
respecte	52	Is it because of their ideology that they are all right with left-wing thuggery in Venezuela and they are opposed to some sort of ideological perspective that is more in keeping with market-based economies?	Est-ce en raison de leur idéologie qu'ils s'accommodent de la brutalité gauchiste du Venezuela et qu'ils s'opposent à une perspective idéologique qui est plus conforme aux économies de marché?
correspond à	36	Clearly that is not in keeping with the Confederation deal to which our ancestors all agreed.	Il est clair que cela n'est pas conforme à l'accord sur la Confédération auquel ont adhéré nos ancêtres.
dans le sens	15	This is in keeping with what Bourinot refers to as the	Leur intention était conforme à ce que Bourinot désigne
en conformité avec	14		
fidèle à	11		
compatible avec	10		
inscrit dans	10		
dans le cadre de	9		
dans le respect de	9		
compte tenu	9		
à fait dans	9		

conforme à	146	fidèle à	11
conformément à	97	However, given that the measures in Bill C-51 are good for Quebec, the Bloc Québécois, in keeping with its responsible approach, will support this bill.	Cependant, comme les mesures contenues dans le projet de loi C-51 sont acceptables pour le Québec, le Bloc québécois, fidèle à son attitude responsable, appuiera ce projet de loi.
respecte	52	It is a good piece of business and in keeping with the spirit with which his father gave great leadership to the country many years ago.	C'est une bonne motion, qui est fidèle à l'esprit de leadership dont faisait preuve son père il y a de cela de nombreuses années.
correspond à	36	The NDP has a vision that is much more in keeping with the values of Canadians, such as improving our health care system, actually dealing with the housing crisis and the homelessness crisis, and reinvesting in Canadian cities.	La vision du NPD est plus fidèle aux valeurs des Canadiens. Elle préconise, entre autres, l'amélioration du système des soins de santé, la prise de mesures concrètes contre la crise du logement et la crise de l'itinérance, et le réinvestissement dans les villes canadiennes.
dans le sens	15		
en conformité avec	14		
fidèle à	11		
compatible avec	10		
inscrit dans	10		
dans le cadre de	9		
dans le respect de	9		
compte tenu	9		
à fait dans	9		

Figure 1: Searching for two French translations of *in keeping with* with TRANSSEARCH. The top part shows an excerpt of the results for *conforme à* and the bottom part for *fidèle à*. In each screen, the list on the left indicates identified translations with their frequency of occurrence and the panel on the right shows the corresponding contexts in which this translation was used. Clicking on a different translation updates the panel on the right with the new contexts. Sentences that appear over a darker background are sentences in their original language while the sentence over a white background are their translations.

which there were 11 occurrences, one only clicks on it and the panel on the right is updated immediately with the occurrences of this new translation.

2 Underlying Technology

TRANSSEARCH must first identify the set of sentence pairs containing the query (Step ② in Figure 2), this is a straightforward application of search engine methods provided each pair is considered a document to be indexed. Given the fact that the system is aimed at language professionals, TransSearch carefully handles morphological variants of French and English query words. For example, the query *go+* will search for *go*, but also for *goes*, *went*, *going* and *gone*. This is the output of the current version of TRANSSEARCH.

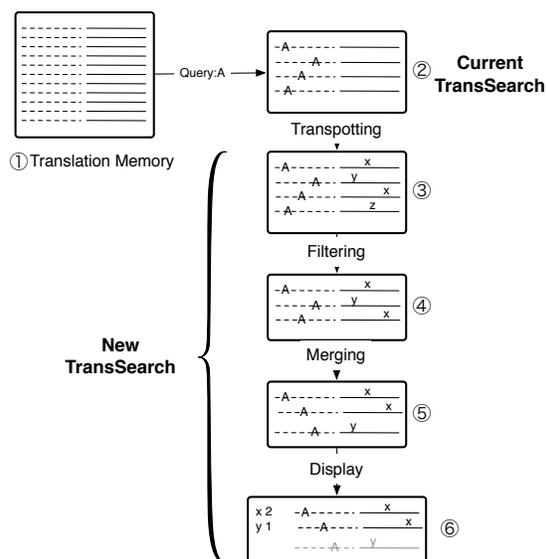


Figure 2: Processing steps of TRANSSEARCH starting from the translation memory of bilingual pairs (Step ①) to the display of groups of translations (Step ⑥).

Once the sentence pairs are retrieved, the translation must be identified in the sentence written in the target language. This will be the French sentence if the query is in English³. This process is called *translation spotting*, relabeled here as *transpotting* (Step ③ in Figure 2). We call *transpot* the target word-tokens automatically associated with a query in a

³The user does not have to specify the language of the query, TRANSSEARCH assumes that it is the language in which the query was found most often.

given pair of sentences. For instance, in Figure 1, *conforme à* and *respecte* are 2 out of 169 distinct transpots displayed to the user for the query *in keeping with*. Transpotting is performed using word alignment techniques that combine HMM and IBM translation models. TransSearch also assumes that, since the query is a sequence of contiguous words, so should be the transpot. This hypothesis simplifies transpotting, which must be performed on the fly when the user queries the system. Because between 20% and 40% of the identified transpots are not acceptable translations, they are filtered (Step ④ in Figure 2) in order to keep only *correct* and *diversified* translations. Filtering is performed using a supervised classification process. (Bourdaillet et al., 2010) describes in detail the many alternatives that have been tested and evaluated for the transpotting and filtering processes.

Once translations have been identified using these methods, it becomes possible to merge the translations (Step ⑤ in Figure 2) that are identical or at least very close, for example if they differ only by inflection, conjugation or by a few words. Figure 1 shows some flexional variations of *conforme à*. Merging is very important for the user because it groups *uninteresting* variations of the same translations and, in some cases, hides deficiencies of the word alignment process. This can be seen in the third context of the top at Figure 1 in which *conforme à* should have been underlined instead of only *conforme*.

TRANSSEARCH displays each translation only once (Step ⑥ in Figure 2), but allows the user to view all its contexts of occurrence. The translations are sorted in decreasing order of their frequency, making the assumption that the most frequent translations are the most likely ones. They are displayed in a web interface, loosely inspired from the one pioneered by Linear B (Callison-Burch et al., 2005). Similar interfaces have been developed for other web based concordancers such as Linguee⁴ and Tradooit⁵.

CSS and Javascript are used to selectively display translation pairs. The list of transpots is presented to the user who can then browse the associated pairs of sentences by clicking on each translation. Both the query and its translation are highlighted in all

⁴<http://www.linguee.com/>

⁵<http://www.tradooit.com/>

pairs of sentences. TRANSEARCH is also available as a web service easily invoked from other applications. Terminotix has developed a TRANSEARCH toolbar to be used from within Microsoft Word. See (Bourdaillet et al., 2010) for details of the merging processes and the evaluations by both machine and man that have been carried out in the course of developing TRANSEARCH.

TRANSEARCH has also been used in studies of linguistic phenomena such as discourse connectives (Meyer et al., 2011) and idiomatic expression (Huet et al., 2010).

Corpus	docs	M pairs	M words
Hansard	2 854	10.5	339.6
Senate	1 025	1.2	47.6
Canadian courts	13 068	3.2	142.6
Total	16 947	14.9	528.2

Table 1: Size of the Transbases used by TRANSEARCH

TRANSEARCH transbases (translation data bases of French and English sentence pairs) contain more than half a billion words (Table 1). They contain the Canadian Hansards (Canadian parliamentary debates) since 1986, but also from the debates of the Senate of Canada since 1995. As TRANSEARCH is also used by the translators of the Canadian Translation Bureau that produces the Hansards, these databases are updated daily when new texts are produced and translated. TRANSEARCH also contains more than 13K decisions from Canadian courts since 1968.

3 Conclusion

This paper describes the bilingual concordancer TRANSEARCH which features word alignment. Interestingly, this transforms the nature of the application: it now behaves like a translation finder with a concordancer feature. The application goes beyond a bilingual dictionary thanks to its ability to find phrase translations while providing their contexts of occurrence. TRANSEARCH can be accessed on a 5-day trial basis at:

<http://transsearch3.com>

Acknowledgments

We thank Julien Bourdaillet and Stéphane Huet for their research contributions as postdoctoral fellows at RALI. We also thank Gilles Gamas, Micheline Cloutier and Jean-François Richard from Terminotix who supported this research via their contribution to a Collaborative Research and Development (CRD) grant from the Natural Science and Engineering Research Council (NSERC) of Canada. We thank also Jacques Steinlin and Elliott Macklovitch for their collaboration to this work.

References

- Callison-Burch, C., C. Bannard, and J. Schroeder: 2005, 'A Compact Data Structure for Searchable Translation Memories'. In: *10th European Conference of the Association for Machine Translation (EAMT)*. Budapest, Hungary, pp. 59–65.
- Bourdaillet, J., S. Huet, P. Langlais and G. Lapalme: 2010, 'TransSearch: from a Bilingual Concordancer to a Translation Finder' *Machine Translation*, 24:3-4, pp. 241–271.
- Huet, S. and P. Langlais, 'Identifying Translations of Idiomatic Expressions using TransSearch', In *8th International Workshop on Natural Language Processing and Cognitive Science*, Copenhagen, Denmark, Aug. 2011.
- Macklovitch, E., M. Simard, and P. Langlais: 2000, 'TransSearch: A Free Translation Memory on the World Wide Web'. In: *2nd International Conference on Language Resources and Evaluation (LREC)*. Athens, Greece, pp. 1201–1208.
- Macklovitch, E., G. Lapalme, and F. Gotti: 2008, 'TransSearch: What are translators looking for?'. In: *8th Conference of the Association for Machine Translation in the Americas (AMTA)*. Waikiki, Hawaii, USA, pp. 412–419.
- Meyer T., Roze C., Cartoni B., Danlos L. and Popescu-Belis A. 'Disambiguating discourse connectives using parallel corpora: senses vs. translations' In *Proceedings of Corpus Linguistics 2011*, Birmingham, United Kingdom.