

AMTA 2010 Technology Showcase

Company or organization	National Research Council of Canada
Products or software	WeBiText
Version	N/A
Available for licensing Y/N	Free web-based service
Kind of product (e.g., MT, OCR, Dictionary, etc.)	Free concordancer which allows searching in large multilingual web sites.
Description	WeBiText is a free, web-based concordancer, that allows translators to search in large, high quality, multilingual web sites like those of the Government of Canada, European Union organizations and other international organizations.
Languages or language pairs supported	29 languages, mostly European, but including some less common ones like Inuktitut (language of the Inuit people of Canada) and Haitian Creole.
Unique about this system	Contrarily to conventional Translation Memories or concordancers, WeBiText comes pre-loaded with large amounts of high quality multilingual content, covering a wide range of topics.
New this year	No. It has been available since 2008, but never advertised widely until December 2009.
Part of a larger system? Describe	No
Made up of other systems? Describe	No
Standards supported	None.
Used by	Several hundreds of translators. Overall, 64% of queries originate from freelance, but with 27% originating from translators working for the Canadian federal government.
Presentations at AMTA—what/when	We are submitting a full paper on this system at the workshop on parallel corpora.
Contact for more information	
Name	Alain Désilets
Telephone	819-712-2813
Email	Alain.desilets@nrc-cnrc.gc.ca

Please use the rest of this page and up to one additional page to provide other information.

Here is an extended abstract of the talk I will be submitting for the workshop on parallel corpora. It should give you a better idea of the project.

In this talk, we present WeBiText (www.webitext.ca), a free concordancer developed at the National Research Council of Canada, which allows translators to search in large, public, high quality multilingual web sites like those of the Government of Canada or European Union institutions (Désilets et al, 2008).

We start by validating the product concept using field data collected in a Contextual Inquiry study where we observed professional translators while they carried their day to day work. In that study, we noticed that translators often used web search engines like Google to manually search in large multilingual web sites. It was also apparent that this was a time consuming process (approximately 2 minutes for a single pair of sentences) which was amenable to automation.

We then provide an overview of the system's functionality and show how it automates the time-consuming manual search process, and allows translators to find several pairs of sentences in a few seconds, with a single click of the mouse. In particular, we highlight the following interesting points:

- Search in 29 languages (including some small languages like Inuktitut and Haitian Creole)

- Search in a growing list of 63 web sites from reputed organizations (ex: government of Canada, European Union institutions, international organizations) which cover a wide range of domains (ex: legal, technology, health, policy, public administration)
- View parallel sentences in context (i.e. see sentences that preceded or followed)
- View a pair of web pages side by side

We then proceed with the results of a log analysis which shows how WeBiText is currently used by translators. We show that it is currently experiencing rapid growth in the number of daily searches. We also present a back of the envelope evaluation of the economic value of this traffic, in terms of hours of work saved to translators, which we estimate at \$1 million annually as of June 2010. We then look at the provenance of those queries and show that home based freelancers account for 64% of them, with 28% for Canadian Government users (both federal and provincial), 4% from private LSPs, 0.5% universities and 3.5% miscellaneous. We discuss the advantages that WeBiText offers to these various constituencies.

Based on a sample of queries from our logs, we show that WeBiText produces at least one pair of correctly aligned sentences for 84% of the cases. We also look at the nature of queries being submitted by users, and show that they correspond to a somewhat even mix of general and specialised language translation problems. This is in contrast to the predominance of general language queries that was found in the logs of TransSearch, a similar tool based on the Canadian Hansard (Macklovitch et al, 2008, Simard and Macklovitch, 2005). Finally, we also discuss the most common types of feedback that we receive from our end users, through the Contact Us and Feedback links.

We conclude the talk with a list of future developments and planned improvements for the system.

References

Désilets, A., Farley, B., Patenaude, G., Stojanovic, M. "WeBiText: Building Large Heterogeneous Translation Memories from Parallel Web Content.", Proc. of ASLIB Translating and the Computer (30), London, UK, Nov 27-28, 2008.

Macklovitch, A., Lapalme, G., Gotti, F., (2008), "TransSearch: What are translators looking for?", in Proc AMTA'08, Waikiki, Hawaii, October 21-25, 2008.

Simard, M., Macklovitch, E., (2005). "Studying the Human Translation Process Through the TransSearch Log-Files.", In Proc. AAAI Symposium on Knowledge Collection from Volunteer Contributors, Stanford, USA, March 2005.