WeBiText: Building Large Heterogeneous Translation Memories from Parallel Web Content

Alain Désilets
National Research Council of Canada
alain.desilets@nrc-cnrc.gc.ca

Benoit Farley
National Research Council of Canada
benoit.farley@nrc-cnrc.gc.ca

Marta Stojanovic
National Research Council of Canada
marta.stojanovic@nrc-cnrc.gc.ca

Geneviève Patenaude
National Research Council of Canada
genevieve.patenaude@nrc-cnrc.gc.ca

ABSTRACT

This paper investigates the extent to which a useful general purpose Translation Memory (TM) can be built based on very large amounts of heterogeneous parallel texts mined from the Web. In particular, we evaluate whether such a TM could add value over TMs built from other large, publicly available parallel corpora, such as the Canadian Hansard. In the case of Canadian translators working with English and French, we show that the answer to both questions is a resounding yes. Using field data collected through contextualized observation and interviews with translators at their workplace, we show how this concept is well grounded in existing workpractices of translators, especially Canadian ones. We also show that a TM based on 10 million pairs of pages from Government of Canada Web sites is able to cover 90% of the translation problems observed in our interview subjects. This turns out to be significantly better than coverage of a general purpose TM built from a smaller corpus, namely, the Canadian Hansard. The difference is most notable for the harder problems, such as specialized terminology. We also evaluate the approach on Web parallel corpora for other languages (European Commission Web sites, and 5000 Inuktitut-English pages harvested from the Nunavut domain), and find the approach to not be as advantageous there. We conclude that, while the concept of building TMs from Web corpora holds great promise, more research may be needed to make it work for language pairs other than English-French.

1. INTRODUCTION

This paper investigates the extent to which a useful general purpose Translation Memory (TM) can be built based on very large amounts of heterogeneous parallel texts mined from the Web. In particular, we evaluate whether such a TM could add value over TMs built from other large, publicly available parallel corpora, such as the Canadian Hansard.

We call this concept *WeBiText*, and one can think of it as a sort of *"Google of parallel text"*. We believe that it may be an attractive concept, especially for translators working as freelancers, or within small to medium organizations. Indeed, although conventional Translation Memories (TM) and Bitexts are routinely used in most large translation organizations, these tools have not been as readily adopted by smaller outfits. An important obstacle to deployment in this sort of environment is the lack of availability of parallel corpora sufficiently large and rich to cover most translation problems encountered by translators in their work. A WeBiText might address this issue by providing translators with TMs that are pre-seeded with large and varied amounts of parallel content harvested from the Web.

The paper is organized as follows. Section 2 describes prior work which is relevant to the concept of a WeBiText. Section 3 justifies the concept using qualitative and quantitative data gathered through contextualized observation and interview of professional translators at work in their normal environment. Section 4 describes the WeBiText framework, which we designed in order to deal with the unique technical challenges involved in building a Translation Memory from such heterogeneous and often unpredictable sources. Section 5 describes three instances of a WeBiText which we built based on different Internet domains, to cover different language pairs. Section 6 evaluates the degree to which these three WeBiText instances are able to cover the different types of translation difficulties encountered by translators in their work. Finally, Section 7 offers conclusions and directions for future research.

2. RELATED WORK

The idea of building a general purpose TM based on *publicly available* parallel texts has already been investigated in the context of the TransSearch project (Macklovitch et al, 2000). This is a bilingual concordancer (English-French) based on the Canadian Hansard (transcripts of the debates at the Canadian House of Commons), one of the most widely studied aligned corpora in translation technology research. The system has been deployed as a commercial product, and has been in daily use by translators since 2003. It is mostly used to find translations of short units (2-3 words), which are either generic phraseology expressions (ex: "as a result of", "at this point") or general language, highly polysemic words (ex: "meaningful", "hopefully") (Simard et al, 2005, Macklovitch et al, 2008). Although the system has a Web interface, the actual corpus was not harvested from the Web, and its size and diversity are much

smaller than what we aim to build in the context of the WeBiText project. This is important because, as pointed out by Macklovitch et al (2008):

> *"[...] the collections currently offered in TransSearch only cover a few technical domains, most notably court rulings and labour relations; otherwise, the majority of the text found in TransSearch are parliamentary debates. Hence, for translation problems involving technical terminology, Canadian translators would be well-advised to consult one of these large term banks [i.e. TERMIUM, Gdt], rather than TransSearch."*

Also, in a survey asking users about ways to improve TransSearch (Macklovitch, 2000), the same authors found that 58% of the respondents suggested including more data from non-Hansard domains. Also, the domains suggested by users tended to cover a wide range of fields. Consequently, a particularly interesting research question in the context of WeBiText, is whether increasing the size and diversity of the corpus through Web mining might address this particular craving.

Other projects have also aimed at building large, general purpose TMs based on *donated corpora*. This includes initiatives such as TAUS Data Association (TAD)[1], Very Large Translation Memory (VLTM)[2], and My Memory[3]. These systems have never been the object of scientific evaluation and publications. One possible drawback of this kind of approach is that, since it it relies on corpus donations, it may prove difficult for the TM to rapidly achieve sufficient critical mass to be useful. In contrast, by harvesting the large amounts of parallel texts that already exist on the Web, a WeBiText may be able to reach critical mass rapidly without relying on the goodwill of hundreds, if not thousands of translation organizations. The downside however, is that Web parallel corpora may be much noisier than donated corpora, since the data has not been specifically prepared for inclusion into a TM.

The idea of harvesting parallel corpora from the Web has also been investigated (Nie et al, 1999, McEwan et al, 2002, Resnick and Noah, 2003, Fattah et al, 2004). However, this work focused on how such corpora can help with tasks like Cross Lingual Information Retrieval, automatic building of bilingual dictionaries, and training of Statistical Machine Translation systems. In contrast, the present paper evaluates the extent to which TMs based on Web corpora can help translators resolve typical translation problems that they encounter in their work.

3. JUSTIFICATION OF CONCEPT

The concept of a WeBiText was inspired and motivated by a *Contextual Inquiry* study carried out in order to better understand translators' workpractices, their use of technology, and how new technologies might assist them better in their work (Désilets et al, 2008). We now describe this study, and how it helped us in the context of the WeBiText project.

Contextual Inquiry (CI)[4] is a well known technique in Human Computer Interaction, where researchers observe and interview potential end-users while they carry out their normal day to day work. This allows the observer to ask probing questions which are relevant and well-grounded in the actual reality of the subject's work. In our case, we conducted interviews with 11 translators coming from a broad range of work environments: home based freelancers (2 subjects), medium sized agencies (1 subject), large government translation departments (4 subjects), academia (2 subjects), and even amateur communities of volunteer translators (2 subjects). Seven of the subjects were based in Canada, and two were based in Japan. Four of the subjects were employed by the Canadian Government, and an additional two translated content for the Canadian Government (but were not employed by it). Eight of the subjects translated from English to French, two from Japanese to English, and one from English to Inuktitut (the language of the Inuit people).

Each subject was interviewed in the context of carrying out two translation tasks: a *natural* and a *controlled* task (50 minutes each). In the *natural* task, we asked the subject to work on whatever document was in her in-tray at the time. The purpose of this task was to maximize the ecological validity of the data, by ensuring that we observed the subject working on a document which is representative of what she usually translates. Topics of the source document varied widely from Aboriginal rights, administration, customs and immigration, education, financial, health, human rights, legal, and municipal affairs. In the *controlled* task, we asked all subjects to translate a same short document (a nontechnical newspaper article). The purpose of this task was to provide a common point of reference across all subjects.

The transcripts of these interviews provided us with very rich qualitative data about how translators work. But it was also a good source of quantitative data. In particular, it allowed us to collect a list of *translation problems* experienced by translators in

Type of problem	**Examples**	**Number Observed**
Terminology	subsidiary, fuel-oil	16
Phraseology	on short notice, for more than a decade	21
General language, polysemic words	grave, fiery, step	25
Cultural or Country-Specific Realities	Go Huskies!, liberal Indian Affairs critic	10
Named entities	Sun (name of a computer company), Xinjian Uighur autonomous region	8

Table 3.1: Types of translation problems observed.

[1] TAUS Data Association (TAD):

http://www.translationautomation.com/tda/mission-a-activities.html

[2] Very Large Translation Memory (VLTM) project:

http://www.wordfast.net/index.php?whichpage=jobs&lang=frfr

[3] My Memory: http://mymemory.translated.net/doc/

their actual day to day work. By translation problem, we mean a word or expression for which the subject had to consult various resources in order to find a proper equivalent in the target language. We also refer to an appropriate equivalent in target language as a *solution* to the translation problem. Table 3.1 provides examples of the different types of problems we observed in the course of our Contextual Inquiries with translators. In total, we observed 80 problems in the course of the natural tasks. Readers who are familiar with the world of translators may be surprised by the relatively small number of problems in that table, given that we observed 11 subjects for 50 minutes each. But one must understand that in the course of the interviews, researcher frequently interrupted the subject to ask probing questions, and many of them prompted long and detailed answers and tangents. As a consequence, the amount of actual work activity observed during the interviews represent a much shorter time than the 50 minutes of elapsed time. To put the number in perspective, those 80 problems amounted to approximately two problems for every 100 words translated by our subjects.

Transcripts of the Contextual Inquiries reveal many interesting things about how translators search in parallel corpora. First and foremost in the context of WeBiText, is the fact that many of our subjects already use Google to manually search the Web for parallel texts. Here is an example illustrating how this was done typically. Suppose a subject was looking for a French equivalent for term *"adjunct professor"*. In that situation, he might compose a Google query as follows:

site:gc.ca "adjunct professor"

in order to retrieve all Government of Canada (GOC) pages containing that exact term (Figure 3.1). The reason for focusing the search on gc.ca is that, by law, most (but not all) GOC pages must be published in both French and English. Once the translator had a list of Google hits, he would scan it for a URL that contains a pattern indicating the availability of a French version. For example, the *"lang=En"* part of the following URL:

http://www.ec.gc.ca/scitech/default.asp?lang=En&n=901AA9A7-1

The translator would then click on that link, and have to search in the page for an occurrence of the query term *"adjunct professor"*. He would then locate a clear and language neutral "landmark" that lied close to the term (Figure 3.2). For example, he might notice that the term occurs in the caption of the first picture in the right margin. Then, the translator would go back to the top of the page and click the *Français* link, which points to the French page (Figure 3.3). He would then scroll through the French page, looking for the language neutral landmark (Figure 3.4), and scan around it for a French translation of the desired term ("professeur adjoint" in this case).

As the reader can imagine, such manual searches were very time consuming. Typically, retrieving a single pair of sentences in that fashion required a minimum of 30 seconds and in one extreme case, the translator took 5 minutes. Yet, as shown in Table 3.1, the practice was fairly common in our subjects. Indeed, 7 of our 11 subjects used this technique at least once during the 50 minute natural task interview. In comparison, only 3 of our subjects used a custom TM. By *custom TM*, we mean a Translation Memory which the translator, his employer, or his client, built themselves using TM systems like Logitrans, Multitrans and Trados (each of which was used by at least one of our subjects).

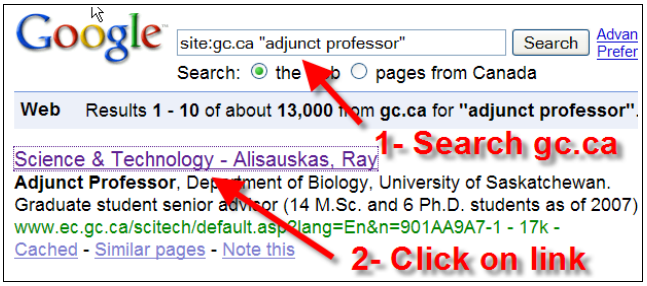

Figure 3.1: Googling for the problem.

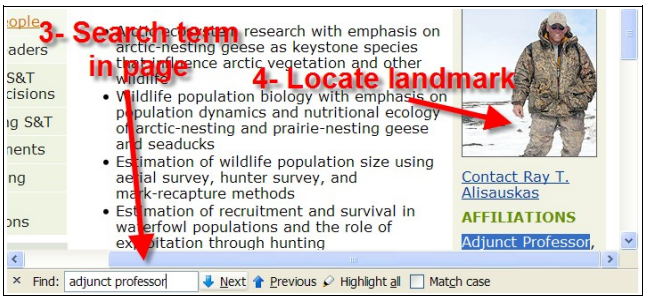

Figure 3.2: Finding occurrence of the problem in a source language page.

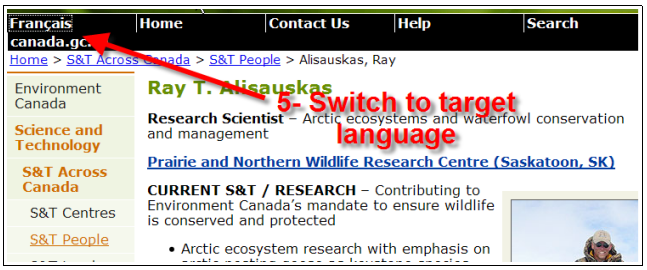

Figure 3.3: Switching to target language page.

Figure 3.4: Locating solution in target language page.

The number of times that a type of tool was used (as opposed to the number of subject that used it) is also much higher for Google than custom TMs (13 versus 4). One might of course wonder if translators who used Google to search parallel texts simply did not know about, or could not afford, a proper TM system. However, looking at just the 6 subjects who did have access to a proper TM, we again see that the number who used Google is 5 versus 3 for TMs, and the number of observed uses is 8 for Google versus 4 for TMs.

The table also shows that usage of TransSearch is about as frequent as that of custom TMs, but lower than that of Google. Again here, one might suspect that subjects who used Google instead of TransSearch did not know about that tool, or could not afford to pay its annual subscription. But looking only at the 7 users who had access to TransSearch, we see that again, usage of Google is slightly larger than usage of TransSearch, both in terms of number of subjects that used them, and the number of times they were used.

When asked why they used Google to search for parallel texts, our subjects evoked a variety of reasons. For freelance and volunteer amateur translators, this was a way to tap into parallel corpora which are much larger than their own private translation archives. For translators working in organizations that had large TMs, this was seen as a way to tap into additional high quality corpora (the assumption being that if something is published on a government Web site, its level of quality must be high).

Of course, given the very small sample size of our study, one should be careful not to interpret these numbers at too fine a level. Also, keep in mind that our subjects were very predominantly Canadians translating between English and French, and there is no guarantee that what we observe is representative of what translators do in other countries, or for other language pairs. However, at a higher level, our data does provide a strong indication that manually searching for parallel texts on the Web is a common practice, at least among Canadian translators, and one that is not well supported by existing tools. This by itself, was sufficient to convince us that automating it would provide value. However, other observations helped reinforce that assessment.

In particular, we noticed that our subjects mostly consulted parallel corpora (whether it be general purpose TM or a more specialized custom TM), to solve short, 1-3 word translation problems. This is consistent with what Simard et al (2005) found in their log analysis of the TransSearch general purpose TM, but our observations extend that result and indicate that this is also the case for searches done in specialized, custom TMs. This information is good news for a WeBiText, because it seems unlikely that one would find the translation of long expressions (ex: a complete sentence) in a general purpose parallel corpus, unless that corpus happened to contain documents that are very similar to the document being translated (ex: an older version of the exact same document). In contrast, it seems more likely that one could find translations of short terms or expressions, in a sufficiently large general purpose corpus collected from the Web.

Whereas Macklovitch et al (2008) found that the TransSearch general purpose TM was mostly used to solve problems related to phraseology and general vocabulary polysemic words, we observed that subjects used Google to search for solutions to the whole range of problem types (with the exception of general vocabulary polysemic words). Frequencies are listed in Table 3.3.

		TS	Google	Custom TM
All Users (11 total)	**Used by**	3	7	3
	Times used	5	13	4
Users with TM access (6 total)	**Used by**	3	5	3
	Times used	5	8	4
Users with TS access (7 total)	**Used by**	3	5	3
	Times used	5	8	4

Table 3.2: Use of different types of corpus based tools (TransSearch, Google, and Custom TM) by different types of users (with or without access to a custom TM and TransSearch) during the natural task.

Terminology	2 cases
Phraseology	1 case
General language, polysemic terms	0 case
Cultural or country-specific realities	3 cases
Named entities	5 cases
Complete quote	2 cases

Table 3.3: Frequency of Google use for different problem types.

While the numbers are small, the fact that Google was used at least once for all but one of the types of problem is an indication that a Web parallel corpus may be more versatile than a more limited corpus like the Hansard.

Our Contextual Inquiries also indicate that, in the context of a WeBiText, trustworthiness of the sources might be an issue. Indeed, our interviews clearly indicate that translators worry about the provenance of solutions to translation problems, and that they tend to prefer sources that they know and trust. But with a WeBiText, parallel sentences may possibly come from any source on the Web, including some that may be less than trustworthy. At the same time, we observed that when they are not able to find solutions in trusted sources, translators have no qualms about searching in less controlled ones, and on the Web in particular. We noticed that translators are very good at dealing with noisy results, where many bad or irrelevant solutions to a translation problem are mixed with a few good ones. It seems that the presence of bad solutions (i.e. lack of precision) is less detrimental than the absence of a solution altogether (i.e. lack of coverage). Indeed, translators seem very adept at quickly scanning a list of potential solutions and choosing the one that is most appropriate for their current needs. Moreover, the basis for choosing a particular solution from a list of hits seems more rooted in the translator's own experience and knowledge about the

topic domain, than in a careful assessment of the sources from which the different suggestions came from.

Consequently, while trustworthiness of sources may be an issue, we do not think it is an absolute obstacle to adoption of the WeBiText concept by translators. However, any features that would make it easier for translators to assess the trustworthiness of the source for a particular proposed solution might help with system adoption (and we shall have more to say about this in Section 7).

To summarize this section, our observation data on translator workpractices provides a very strong case for a tool that would automate searching of parallel Web sites. This is exactly what WeBiText attempts to do.

4. THE WeBiText FRAMEWORK

In this section, we describe a framework which we designed for building TMs based on parallel corpora harvested from the Web. In particular, we describe some of the unique technical challenges we had to solve in dealing with corpora coming from such heterogeneous and often unpredictable sources.

4.1 On-demand versus pre-processing approaches

Because the main idea behind WeBiText is to act as a sort of "Google of parallel text", it might need to crawl and index a great number of Web sites containing very large amounts of parallel multilingual texts. This is no small feat, considering the amount of computer power, bandwidth, and human resources that companies like Google and Yahoo! have invested into it.

As a way to experiment with the WeBiText concept without the hassles of large scale crawling and indexing, our first version used an approach which we call *on-demand*. Given a problem *P* entered by a user, a target language *T*, and an Internet domain *D* where to search for a solution, this on-demand version carries out the following steps:

- *Step 1*: Send query *P* to a Web search engine (Yahoo! in this case), restricting the search to domain *D*. Retrieve a list of 100 hits.
- *Step 2*: Using some simple heuristics, find 10 of those hits which look like "good" pages with an equivalent page in target language *T*. The exact heuristics used to identify "good" pages will be explained in section 4.2.
- *Step 3*: Retrieve the content of the 10 selected hits, and their corresponding translations in target language *T*.
- *Step 4*: Align the 10 pairs of pages, find sentence pairs that contain an occurrence of problem *P* on the source language side, and present those pairs to the user.

Alignment in Step 4 is carried out using the aligner from th PORTAGE Machine Translation system (Ueffing et al, 2007). This implement a single pass alignment algorithm based on the length of sentences (in number of characters), and it is roughly equivalent to the approach described in Gale and Church (1993).

The advantage of the on-demand approach is, of course, that we do not have to do any crawling of Web sites, and that in principle, users can search on any Internet domain that has parallel content. The disadvantage is that it is really slow, due to steps 3 and 4. This is the reason why heuristics must be used to narrow down pairs of pages in Step 2, but even with this, a typical query takes about 30 seconds to process (assuming no other query is being processed concurrently). Another disadvantage of the on-demand approach is that the system can only return parallel sentences contained in at most 10 pairs of pages, even if other (and possibly better) solutions for *P* might be found in page pairs that were further down the list of 100 hits returned by Yahoo!. Finally, a third disadvantage is that search engines like Yahoo! limit the number of queries that can be submitted per day from a same IP address or account. In the case of the current implementation of WeBiText, this concretely results in a maximum of 1000 queries per day, all users included.

In order to deal with these limitations, we built another version of WeBiText, based on a *pre-processing* approach. Here, the system crawls a limited list of Internet domains which are known to have good parallel content, and pre-aligns all the parallel pages it finds. The advantages of the pre-processing approach is that it is fast, and can deal with unlimited number of queries per day. The disadvantage is that users can only search on Internet domains which have already been pre-processed by the system.

We are currently working on a third version of the WeBiText framework, which will combine those two approaches into an *hybrid approach*. When a user submits a query for an Internet domain that has already been indexed and aligned by WeBiText, the system will quickly return pre-crawled and aligned results. When the Internet domain is not already in the system's database, it will use the slower on-demand approach. In addition, the system will then start an automatic background process to evaluate if this new Internet domain is worth pre-processing, and if so, initiate the process of crawling, indexing and aligning its content. This will in turn ensure that future queries on that same domain can be answered more rapidly.

4.2 Dealing with noisy sources

A unique challenge when building TMs from Web data is that it can be much noisier than data coming from more controlled sources. In this section, we describe the various causes of such noise, and solutions we took to deal with them.

A first source of noise is caused by the fact that equivalent pages in different languages are not explicitly paired to each other. Yet, given a page in source language, the system must somehow find its translation in the target language, if it exists. A simple way to do this might be to parse the content of the source page and look for an inter-language hyperlink whose anchor text is the name of the target language. While this might work in many cases, it assumes that we have already downloaded the content of the source page, which might lead to unnecessary and costly downloads in Step 2, if it turns out that the page does not have such a link after all. Also, there are many cases where linkage between linguistic versions of pages is done through HTML code that is harder to parse and interpret. For example, some sites use a javascript enabled language picklist, in which case, it might be hard to automatically deduce which page corresponds to a particular language choice in that list. Other sites use country flag icons for inter-language links, and do not bother to specify the language name using the ALT text of the image. This makes it

virtually impossible for a machine to figure out which image corresponds to which language. Finally, this approach does not work for PDF documents which typically do not contain hyperlinks, and in our experience, there are many parallel PDF documents on the Web.

Instead of relying on inter-language links, WeBiText uses the URL of a page in the source language to deduce the URL of the corresponding page in the target language. It does this by using several substitution heuristics. Below are some examples in the case of an English to French conversion:

- english -> francais
- _en -> _fr
- _e -> _f
- _eng -> _fr

The system executes all substitutions that apply, because the name of the language is often specified in several places in the path of the URL. If no substitution patterns apply, then the target language page is not added to the list of 10 "good" pages.

This simple URL pattern substitution approach is similar to that used by Nie et al (1999), who estimated the recall of this kind of heuristic to be at least 50%, and precision to be over 95%. Likewise, we found these simple heuristics to work very well on a variety of Web sites that contain parallel content. Note however that they do not deal with situations where more than one linguistic versions are included side by side in a same document, which is often the case for PDF documents. Also, there are cases where the source page URL gives no indication as to the URL of the target language page, in which case, it can only be deduced through analysis of the inter-language links. Consequently, we are currently working on adding this kind of link analysis for situations where the above heuristics fail.

A second source of noise is the actual content of the parallel pages. Indeed, we have often found that, while a page in say, English, might have a corresponding page in say, French, the French page turns out to not be a translation of the English page at all. For example, it might simply consist of a short note saying that this content is not currently available in French. In other cases, all the French links on all the English pages point to one and a same page, namely, the home page of the French site (ex: all IBM USA pages pointing to the home page of IBM France). Other times, the "French" page might be in fact an untranslated, exact duplicate of the English page. And although we have not encountered this in practice, there is also the possibility that some of the translations might have been produced automatically using Machine Translation, and be therefore of poor quality. We deal with the first two issues by doing a basic sanity check on the length of the source and target page during Step 2. If the lengths (in number of characters) differ by more than 20%, this pair is not considered for inclusion in the set of 10 "good" pairs. We deal with the third issue (identical source and target pages) during Step 4, by excluding from the results presented to the user, any pair of sentences where the source sentence is exactly identical to the target sentence. We currently do not have a way to deal with the fourth issue (MT generated target pages).

5. Three case studies

We now present three case studies which illustrate the potential of the WeBiText concept.

The first case study, called *WeBiText-GC*, is a TM based on all Government of Canada (GOC) Web sites (gc.ca domain), using an *on-demand* approach. The reason why we focused on GOC sites is that, in our Contextual Inquiries, most of the time when we witnessed a translator conducting a manual search for parallel Web sentences, it was done on gc.ca. According to Google, this domain contains 15 million English pages, and 9 million French pages. Most of these French pages have a corresponding translation into English, because of a GOC policy requiring Web content to be posted in both languages. Because of the large size of this domain, we opted initially for an on-demand approach, in order to avoid having to crawl and index it upfront.

WeBiText-GC has been in daily use by a small number of pilot users for 10 months now. It averaged 45 queries per day in August and September of 2008. Overall feedback is encouraging. At this point, the most common complaint is its lack of speed.

The second case study, called *WeBiText-EU*, also uses an on-demand approach and is based on approximately 30 million pages from the European Union (europa.eu domain), and in principle, supports 23 European languages. Note however that coverage of the 23 languages is uneven. For example, while Google finds 19 million pages for English, it only finds 3.5 million pages for French, 1.7 million for Spanish, and 0.9 million for Hungarian. In other words, at most one out of every five English pages can have been translated to French, and this ratio is ten for Spanish, and twenty for Hungarian.

An interesting point about this case study is that, starting from code which was originally designed specifically to deal with English-French parallel pages found on the gc.ca domain (Government of Canada), we were able to create this WeBiText-EU TM in a single morning. All that was required was to point WeBiText to the europa.eu domain instead of gc.ca. This is a great testimonial to the versatility of the on-demand approach. At this point however, WeBiText-EU is still very much an early prototype, and has not been used by pilot users.

The third case study, called *WeBInuk*, is a TM which uses a pre-processing approach, and is based on a collection of Inuktitut-English pages found in the Nunavut domain. The reason for choosing that particular corpus is that Inuktitut, the language of the Inuit people, is one of three languages with an official status in Canada. But contrarily to French and English translators who are very well endowed with tools like TERMIUM (the publicly available terminology database of the Government of Canada), Inuktitut translators are very resource deprived. Indeed, they do not even have access to a comprehensive bilingual dictionary. Contrarily to WeBiText-GC and WeBiText-EU, we had to use a pre-processing approach for WeBInuk because there was no single large Internet domain that contained a lot of parallel English-Inuktitut content. Also, Web search engines like Google and Yahoo! do not deal very well with Inuktitut documents which are often encoded using legacy fonts as opposed to proper Unicode characters. Indeed, we found we had to develop our own custom tools to reliably convert Inuktitut pages into a common, indexable Unicode format. The WeBInuk database contains approximately 5000 pairs of pages, crawled from 61 Internet

domains. Those domains were identified by querying Google for common Inuktitut words, using several different encodings (Unicode, and several Inuktitut legacy fonts). WeBInuk is still at the early prototype stage and has not yet been tried by pilot users.

Note that WeBInuk is a follow up to a prior system called NunHanSearch, which was co-built by one of the authors (Benoit Farley, in collaboration with Joel Martin) based on eight years of Hansard of the Nunavut parliament (1999 to 2007). That system has been in use by a small community of Inuktitut translators since April 2007. In this paper, we will use NunHanSearch as a baseline against which to compare the performance of WeBInuk.

All three of the above WeBiText instances are freely available for testing by pilot users are www.WeBiText.ca.

6. Evaluation

We now describe a series of experiments which we carried out to measure the potential usefulness to translators of the WeBiText instances described in the previous section.

6.1 Evaluation procedure

We evaluated the extent to which different WeBiText instances are able to find solutions for typical translation problems that translators encounter in their work. As a basis for comparison, we also evaluated TransSearch and NunHanSearch, in order to assess the extent to which parallel Web corpora can improve coverage over smaller, publicly available parallel corpora like the Canadian or Nunavut Hansards. It is worth noting that all of the evaluated systems use very similar parallel text alignment technologies, and that any difference in performance is mostly due to differences in the corpora.

If we are to evaluate the usefulness of WeBiText to translators, we must first understand how translators themselves evaluate lists of suggestions offered by Computer Assisted Translation tools. Our Contextual Interviews indicate that when translators use a tool to search for a solution to a particular problem, their main goal is to find *one acceptable solution* in the first 10 to 20 suggestions (Désilets et al, 2008). Also, translators seem to have a relatively high tolerance for noise in the list of suggestions. In other words, they do not mind seeing a list of mostly poor suggestions, as long as it contains a few good ones in the top 10. Translators seem to be highly skilled at quickly scanning lists of potential solutions to a translation problem, and identifying which ones (if any) are most appropriate for their current situation.

Given those observations, we decided to evaluate the performance of WeBiText using a measure which we call *problem coverage in the top N* (or *problem coverage* for short). We define it as follows:

$$PCov_N(P) = Avg_{p \text{ in } P}(success(p, N)) \quad (1)$$

where:

P = a set of translation problems

success(p, N) =

1 if the tool proposes a relevant solution for p in its top N hits

0 otherwise

In other words, given a set of translation problems, coverage is proportion of times (or, if if you prefer, the probability) that the tool includes at least one relevant solution in its top N suggestions.

Although translators seem mostly interested in rapidly finding one solution to a problem, we found that they also like to be given a choice of different relevant solutions, so they can select the one that seems most appropriate for their current situation. Multiple suggestions are also often used as a source of inspiration for creating new solutions that are ideally suited to the situation at hand. Therefore, another reasonable evaluation measure would be *solutions recall in the top N (*or *solutions recall* for short*)*, which we define as:

$$Rec_N(P) = Avg_{p \text{ in } P}(num\_found(p, N)/total\_sols(p)) \quad (2)$$

where:

num_found(p, N) = number of relevant solutions that the system proposed for problem p in its top N hits.

total_sols(p) = total number of relevant solutions that exist for p

In other words, solutions recall is the proportion of all relevant solutions that a tool is able to suggest in its top N list. Although this measure is well supported by our Contextual Inquiries transcripts, it is not evaluated in the present study, and is left for future research.

A third possible measure is *solutions precision in the top N*, which we define as follows:

$$Prec_N(P) = Avg_{p \text{ in } P}(relevant(d)/total\_proposed(d)) \quad (3)$$

where:

relevant(p) = number of suggestions proposed for p which are in fact relevant.

total_proposed(p) = total number of unique solutions proposed for p by the system.

In other words, solutions precision is the proportion of top N solutions proposed by a tool, which are in fact relevant for the problem at hand. Given what we observed previously about translator's tolerance for noisy lists of suggestions, we consider that this last measure is not particularly relevant in the WeBiText context, and we therefore did not measure it in the present study.

We evaluated problem coverage in the top 10 (i.e., N = 10) for the three WeBiText instances, on various sets of translation problems, which we now describe. English was the source language for all problems in those sets.

The first set, which we call *ContInq*, corresponds to the set of all problems for which subjects searched in some resource (whether it be a dictionary, terminology databases, TM, or the Web) during the natural task of our Contextual Inquiries. There are 80 such cases in that set. The purpose of this set is to evaluate the extent to which a WeBiText instance is able to address the full range of difficulties that translators encounter in their typical work.

A second set, *CINonGOC* corresponds to a subset of 29 problems from *ContInq*, which were encountered by subjects who were not translating content for the Government of Canada. The purpose of this set is to assess the degree to which WeBiText-GC (which is

based on GOC sites) can still be useful for translators working on non-GOC content.

A third set, *TSNoHits*, consists of a sample of 75 randomly picked queries taken from the TransSearch logs, for which that system returned no hits at all in 2000 (we obtained this set from the TransSearch research team at RALI). The purpose of this set was to evaluate the extent to which a TM based on all of the GOC sites could address these problems better than a TM based on a corpus like the Canadian or Nunavut Hansard only.

A fourth set, *GdtTerms*, consists of 195 terms selected randomly from the Grand Dictionnaire Terminologique (Gdt)[5], a very popular terminological database available to the public. The purpose of this set is to assess the degree to which a tool like WeBiText could help solve specialized terminological problems better than a general purpose TM based on a corpus like the Canadian or Nunavut Hansard.

Note that in order to evaluate coverage, we needed to determine the relevance of suggestions for each problem in the evaluation sets described above. We now describe how this was done.

In our evaluation of WeBiText-GC, we deemed that a solution was relevant if it provided a translation to the problem in the appropriate sense and topical domain. Below are typical situations in which a solution would have been deemed irrelevant:

- The target language sentence provided a translation of the problem, but it was for the wrong sense or topical domain.
- The target language sentence was in fact not properly aligned with the source language sentence containing the occurrence of the translation problem.
- The target language sentence turned out to be written in the source language (i.e., the whole sentence was never actually translated).
- The problematic part of the source language sentence had not been translated in the target language (this sometimes happens when translators "avoid" a difficulty by simply not rendering that particular part of the message in the target language).

Note that we did not try to evaluate the quality of the solutions proposed because this can be highly subjective, and translators often do not agree among themselves on what constitutes a quality solution to a given translation problem. Also, the first point about proper sense and domain could not be assessed for the *TSNoHits* set, because we have no information about the context in which those queries were performed. In contrast, for *ContInq* and *CINonGOC*, since the problems came from our Contextual Inquiries, we knew exactly what sense and domain the subject was looking for. In the case of *GdtTerms*, we randomly picked one of the senses and domains from all the ones listed for a particular term.

[5] GDT: http://www.granddictionnaire.com/btml/fra/r_motclef/index800_1.asp

6.2 Performance of different WeBiText instances

We now describe how various instances of WeBiText fared in terms of problem coverage for various sets of translation problems. As a point of comparison, we also evaluated coverage of TranSearch-Hansard and NunHanSearch on those same problem sets, to see the effect of using a larger and more varied corpus than the Canadian or Nunavut Hansard. The results of these experiments is summarized in Table 6.1.

It is worth noting that since the Canadian Hansard is posted on gc.ca, the corpus that WeBiText-GC can tap into is actually a super set of the Hansard. This however does not necessarily mean that coverage of WeBiText-GC will always be superior to that of TransSearch-Hansard, because of the real-time sampling being done in Step 2. A similar point holds for WeBInuk versus NunHanSearch, except that in this case, there is no real-time sampling being done, and coverage of WeBInuk should therefore always be at least equal to that of NunHanSearch.

Note again that all the WeBiText instances as well as TransSearch and NunHanSearch, use essentially the same kind of alignment algorithm, and that differences between them can therefore mostly be attributed to the difference in corpus.

The first point we notice is the very high coverage exhibited by both WeBiText-GC and TransSearch-Hansard on the *ContInq* problem set. Indeed, WeBiText-GC was able to find a relevant solution for 90% of the translation problems experienced by our subjects. Even TransSearch-Hansard, which is based on a much smaller corpus, was able to deal with 86.3% of the problems. The latter is somewhat surprising, considering that previous studies of TransSearch logs found that the system returned empty results for 39% of the queries (Macklovitch et al, 2000). We can offer two possible explanations for this. One possibility is that the translation problems contained in *ContInq* are, in average, easier to solve than those contained in the logs of TransSearch. In other words, it could be that translators turn to a tool like TransSearch when they cannot find an answer in other resources like dictionaries and terminology databases. The difference could also be caused by the fact that 7 of our 11 subjects were translating texts for the Government of Canada, and that the Canadian

	ContInq	CINonGOC	TSNoHits	GdtTerms
WeBiText-GC	0.901	0.965	0.480	0.349
TransSearch-Hansards	0.863	0.896	0.093	0.133
WeBiText-EU (French)	0.662	0.689	0.173	0.138
WeBiText-EU (Spanish)	0.550	0.621	0.160	0.118
WeBINuk	0.513	0.483	0.040	0.036
NunHanSearch	0.500	0.414	0.027	0.031

Table 6.1: Problem coverage in the top 10 of six TMs on various sets of translation problems.

Hansard and gc.ca would therefore have been particularly well suited to address their translation problems. However, looking at the *CINonGOC* column, we can see that the first explanation is probably the correct one. Indeed, for subjects who were *not* translating GOC documents, coverage of both WeBiText-GC and TransSearch-Hansard is the same as (and in fact slightly superior to) that observed for subjects translating GOC content.

Another interesting trend is that the coverage of WeBiText-GC is consistently better than that of TransSearch-Hansard. ANOVA shows that the difference between those two systems are statistically significantly ($p < 0.05$) for all problem sets except *CINonGOC* ($p = 0.16$). Note however that the latter was the smallest of the four problem sets (29 cases only), and this may be the main reason why statistical significance could not be reached there.

The difference between WeBiText-GC and TransSearch-Hansard is most strongly marked on the *TSNoHits* set. Here, we see that WeBiText-GC is able to resolve 48% of the problems for which TransSearch returned empty results in 2000. Note also that, as of this year (2008), TransSearch-Hansard is able to address 9% of those, owing to the fact that its database has been updated with more recent Hansard content. Given that the corpus used in WeBiText-GC is a superset of the Canadian Hansard, we can conclude that it would be able to at least return hits for the 61% of the queries for which TransSearch was reported to return results in 2000 (actually, this would only be true if WeBiText-GC was modified to use a pre-processing approach without realtime sampling in Step 2). In addition, our results indicate that WeBiText-GC would be able to return hits for 48% of the remaining 39%, for a total of 80%. In contrast, the most recent version of TransSearch (which is based on the Hansard only) would be able to return hits for only 65% of the queries (the previously productive 61%, plus an additional new 9% of the remaining 39%). This represents an absolute 15% difference, which is non-negligible.

The difference between the two corpora is also very strongly marked on the *GdtTerms* set, where WeBiText-GC was able to find solutions for 34.9% of the terms. This is almost three times the percentage solved by TransSearch-Hansard (13.3%). This difference is a clear indication that a Web corpus like gc.ca can offer value when it comes to searching for specialized terminology such as: “high wing monoplane”, “anosmia” and “cryostat” (all of which were solved by WeBiText-GC, but not by TransSearch-Hansard).

The fact that WeBiText-GC performed better than TransSearch-Hansard on *TSNoHits* and *GdtTerms* is a clear indication that the gc.ca Internet domain is a more varied and rich source of parallel data than the Canadian Hansard, for solving the harder translation problems, including specialized terminology. Again, let us point out that, since the two systems use approximately the same underlying technology, these differences in coverage are mostly attributable to the differences in the corpora used.

Looking now at WeBiText-EU, we can see that both the French and Spanish versions perform much worse than WeBiText-GC. All differences between WeBiText-GC and either versions of WeBiText-EU were found to be statistically significant ($p < 0.005$). Again, the higher coverage of WeBiText-GC cannot be explained by the fact that our subjects were mostly translating content from the Canadian Government. Indeed, the differences are also present for the *CINonGOC* problem set, and are statistically significant. Comparing coverage of both WeBiText-EU instances to that of TransSearch-Hansard, we find a somewhat mixed picture. While TransSearch-Hansard performed significantly better than either of the WeBiText-EU instances for *ContInq,* and *CINonGOC*, it performed significantly worse than WeBiText-EU Spanish on *TSNoHits*. All other pairwise comparisons were found to not be statistically significant.

	WeBiText-GC	**WeBiText-EU Fr**
a) Pr(not in En)	0.400	0.534
b) Pr(no Fr page|in En)	0.019	0.074
c) Pr(not in aligned|Fr page)	0.097	0.272
d) Pr(none relevant|in aligned)	0.048	0.164

Table 6.2: Conditional probability of failure at four different points in the processing: a) Finding problem on English side, b) Finding corresponding French pages, c) Aligning pages and keeping only sentence pairs that contain the problem, and d) Checking if any of the aligned sentences contain a relevant solution.

It therefore appears that overall, europa.eu is a clearly less useful Web corpus than gc.ca, for the purpose of building a TM, and that it offers no decisive advantage over a corpus like the Canadian Hansard, for the problem sets we investigated.

It is interesting to analyze the reasons for failures with the europa.eu corpus, versus gc.ca. Looking at Table 6.2, we see that both corpora have a comparable coverage of the translation problems on the English side. However, the conditional probabilities of failure in the remaining three steps are much higher for europa.eu than gc.ca. The fact that steps b) and c) fail more often on europa.eu is consistent with our cursory assessment of that domain. Indeed, it seems that content is not translated as comprehensively there as it is on gc.ca. As pointed out earlier, at most one out of five English page can have been translated to French on europa.eu. Often times, we have also found that when an English page has a corresponding French page, the latter is either a notice saying that this content is not currently available in that language, or the content of the two pages are seriously out of sync with each other. These findings are confirmed by the following quote from the EUROPA Frequently Asked Question page on languages[6]:

> *Frequent visitors to EUROPA tell us they are surprised at the many differences between sub-sites in terms of the languages available. It will come as no surprise to you that only part of EUROPA's pages are currently available in all 23 official languages[...]. While you might appreciate that fact many sub-sites are at least available in the 11 'old' languages, we receive many complaints about sub-sites being in only three, two, and sometimes only one language (usually English).*

[6] EUROPA languages FAQ:

http://europa.eu/abouteuropa/faq/q10b/index_en.htm

We currently do not have an explanation for why probability of failure in the last step (checking for relevant solution) is also much higher for europa.eu than gc.ca. It could be that, since europa.eu returns fewer aligned sentences to start with, the probability that at least one of them is relevant is accordingly smaller.

Next, we turn our attention to the evaluation of WeBInuk. Like in the case of WeBiText-GC, we compare its coverage to that of a benchmark system built on top of a smaller, non-Web corpus, namely NunHanSearch (which is based on 8 years of the Nunavut Hansard).

Looking at the last two rows of Table 6.1, we see that a TM based on the Hansard plus 5000 additional Web pages, only offers marginally higher coverage than one based on the Hansard alone. None of the differences between coverage of both TMs were found to be statistically significant ($p < 0.05$), even in the case of the 195 *GdtTerms* ($p = 0.08$). In other words, it appears that 5000 parallel Web pages is not sufficient to significantly add value to a corpus the size of the Inuktitut Hansard, and that additional effort would have to be invested in order to locate more sources of such parallel text.

7. Conclusion and future research

The purpose of our work was investigate whether a useful general purpose Translation Memory (TM) could be built based on very large amounts of heterogeneous parallel texts mined from the Web. In particular, we wanted to evaluate whether such a TM could add value over TMs built from other large, publicly available parallel corpora, such as the Canadian Hansard.

In the case of Canadian translators working with English and French, our evidence allow us to answer both questions with a resounding yes. Using user field data gathered through Contextual Inquiries with translators, we have showed how this concept is well grounded in a manual practice that Canadian translators already use on a routine basis. We also showed that for the French-English language pair, a TM built based on Government of Canada sites is able to address a large proportion of translation problems encountered by translators in their day to day work. We also showed how this TM offered greater problem coverage than a general purpose TM based on a smaller (but still sizable) corpora, namely the Canadian Hansard. This is especially true for harder problems such as specialized terminology.

In the case of other language pairs, however, we find the evidence to be more mitigated. When we built TMs based on Web parallel corpora for other languages (European Commission Web sites, and 5000 Inuktitut-English pages harvested from the Nunavut domain), we found coverage to be significantly worse than that of the above English-French TM. We also found that those Web TMs did not perform significantly better than TMs based on Hansard corpus for the same language.

In short, we conclude that, while the concept of building TMs from Web corpora holds great promise, more research may be needed to evaluate its actual potential for language pairs other than English-French.

A first step in that direction would be to assess the degree to which good parallel text can be found for other language pairs. Could it be that gc.ca is a unique, outstanding resource that has no other equivalent in any other language? The answer to that question is simply not known at the moment. Note however that, even if that was the case, it might still be possible to build a useful Web TM by pulling together parallel content from several Internet domains for a given language. For example, if we were to pool all English-Spanish parallel content that is available on the Web, from countries like Spain, Latin American states, and even the US, we might end up with a corpus that is even bigger and better than gc.ca.

A related question is the extent to which coverage of even WeBiText-GC could be improved by crawling additional sites outside gc.ca. Although the Government of Canada domain is fairly large and rich (19 million pages covering most aspects of Canadian society), it probably represent only a fraction of parallel Web content for the English-French pair. It could be that other French-English sites could help increase coverage for specialized terminology. For example, much medical terminology might be found on sites of organizations like the Canadian Medical Association, Canadian Lung Association, and Canadian Heart and Stroke Foundation (all of which are at least partially bilingual).

In order to answer the two questions above, we need to design tools to help us find Web sites with good parallel content. In Section 4.1, we talked about the possibility of allowing users to specify an arbitrary Internet domain on which to carry out the search using an on-demand approach. Given a sufficiently large population of users, this simple technique might turn out to be a very efficient way of finding out about large domains that contain good parallel content. Another approach would be to use automated techniques such as the ones described in Nie et al (1999). One would send queries to a Web search engine to find pages in say, English, that have a link whose anchor text contains the word “français” (French). One could then identify sites that contain many such pages, then sample a small number of pages from each site, to determine if a large proportion of those actually have a parallel page in French. When that is the case, the system could then crawl that site in order to harvest its parallel content.

Contextual Inquiries with translators outside of Canada may also be warranted to assess the extent to which searching for parallel Web text is consistent with the work practices and values of translators in other parts of the world. We plan to do this in the coming year.

An altogether different research axis is to evaluate usefulness of the WeBiText concept for French-English, through extensive testing with pilot users. We plan to do this with translation students, in the context of a course on Computer Assisted Translation tools at the University of Ottawa. We expect that this will uncover the need for new features which, while not critical for conventional TMs, become invaluable with a large TM developed from uncontrolled heterogeneous sources. For example, it may be a WeBiText TM returns results from a much wider range of topical domains than a typical TM. Consequently, users may need ways of filtering or re-ordering results based on a specification of a topical domain. WeBiText already provides a lightweight way of doing this, by allowing the user to specify an Internet domain for the search. This could be used for example, to narrow search down to sites in the ic.gc.ca domain (Canadian Ministry of Industry). However, it could be that more sophisticated features need to be developed, using content-based,

automatic text classification algorithms. As we alluded to in Section 3, users may also need tools to rapidly assess the reliability of the source for a particular sentence pair. For example, the system might display a reliability scores based on something like Google's PageRank. An alternative would be for the system to use a massive online collaboration approach (Désilets, 2007) and allow the community of users to collectively rate sources and solutions they proposed.

In this paper, we only evaluated the various WeBiText instances in terms of problem coverage, that is, the extent to which they can provide at least one relevant solution to each translation problem in a given set. It would be interesting to also evaluate those systems in terms of solutions recall, that is, the extent to which they are able to suggest a variety of different unique solutions for each problem. In particular, it would be interesting to know to what extent a larger Web-based corpus like gc.ca provides more variety that a more limited corpus like the Canadian Hansard.

In the longer term, it might also be interesting to investigate the use of *comparable* (as opposed to parallel) Web corpora. This is particularly interesting since the amount of comparable content on the Web is probably much larger than the amount of strictly parallel content (this however still remains to be shown empirically). For example, all large multinational corporations have sites in different languages which, while not parallel, talk about more or less the same things. Wikipedia is another example, where in addition, there is a clear mapping between pages in different languages, even if the actual content of those pages is not strictly parallel. Therefore, it would be interesting to develop and evaluate algorithms that can carry out searches in comparable Web corpora. Note however that technologies for searching in comparable corpora are much less mature than for parallel texts (Hartley et al., 2007), and therefore, one should expect much lower problem coverage there.

8. ACKNOWLEDGMENTS

The authors are indebted to the following people, for stimulating corridor conversations and feedback on the WeBiText project. From National Research Council of Canada: Michel Simard, Caroline Barrière, George Foster, Joel Martin and Pierre Isabelle. From RALI (U. of Montreal): Elliott Macklovitch and Guy Lapalme. From Université du Québec en Outaouais: Louise Brunette and Christiane Melançon. From Ottawa U: Lynn Bowker and Elizabeth Marshman. Also, the 11 subjects who participated in our Contextual Inquiry, who, for obvious reasons, must remain anonymous, as well as their managers who allowed them to participate in the study during working hours.

9. REFERENCES

Désilets, A. Brunette, L. Melançon, C. Patenaude, G. (2008), *"Reliable Innovation: a Tecchie's Travels in the Land of Translators."* In Proc. AMTA'08, Waikiki, Hawaii, USA, October 21-25, 2008.

Désilets, A., *"Translation Wikified: How will Massive Online Collaboration Impact the World of Translation?"*, in Proceedings of Translating and the Computer (29). November 29-30, 2007. London, United Kingdom.

Fattah, M., A., Ren, F., Shingo, K., (2004), *"Internet Archive as a Source of Bilingual Dictionary"*, in Proc. ITCC'04, Las Vegas, Nevada, USA, April 5-7, 2004.

Gale, William A. & Church, Kenneth W. (1993), *"A Program for Aligning Sentences in Bilingual Corpora"*, Computational Linguistics 19(1): 75-102

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

Macklovitch, E., Simard, M., Langlais, P, (2000). *"TransSearch: A Free Translation Memory on the World Wide Web."* In Proceedings of LREC 2000, Athens, Greece.

McEwan, C., J., A., Ounis, I., Ruthven, I., (2002) *"Building Bilingual Dictionaries from Parallel Web Documents.",* in Advances in Information Retrieval, Springer, ISBN 978-3-540-43343-9, 2002.

Nie, J. Y., Simard, M., Isabelle, P., Durand, R., (1999) *"Cross Language Information Retrieval based on Parallel Texts and Automatic Mining of Parallel Texts from the Web"*, in Proc. SIGIR'99, Berkeley, California, USA, August 15-19, 1999.

Resnick, P., Noah, A., S., (2003), *"The Web as a parallel corpus.", Computational Linguistics"*, Volume 29 , Issue 3, 2003.

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.

Ueffing, N., Simard, M., Larkin, S., Johnson, J. H., (2007), *"NRC's PORTAGE system for WMT 2007"*,ACL-2007 Workshop on SMT, Prague, Czech Republic 2007