TransSearch: A Free Translation Memory on the World Wide Web

Elliott Macklovitch, Michel Simard, Philippe Langlais

Laboratoire RALI
Université de Montréal, Montreal, Canada
{macklovi, simardm, felipe}@iro.umontreal.ca

Abstract

A translation memory is an archive of existing translations, structured in such a way as to promote translation re-use. Under this broad definition, an interactive bilingual concordancing tool like the RALI's *TransSearch* system certainly qualifies as a translation memory. This paper describes the Web-based version of *TransSearch*, which, for the last three years, has given Internet users access to a large English-French translation database made up of Canadian parliamentary debates. Despite the fact that the RALI has done very little to publicize the availability of *TransSearch* on the Web, the system has been attracting a growing and impressive number of users. We present some basic data on who is using *TransSearch* and how, data which was collected from the system's log file and by means of a questionnaire recently added to our Web site. We conclude with a call to the international community to help set up a network of bi-textual databases like *TransSearch*, which translators around the world could freely access over the Web.

1. Introduction

A translation memory (TM) is an archive of existing translations, structured in such a way as to promote translation re-use. Though rather vague, this definition is nevertheless somewhat contentious, in that it diverges from the widespread use of the term "translation memory" that has arisen as a result of the popularity of commercial translation support systems such as *Transit*, *Translator's WorkBench* and *Déjà-Vu*. These systems implement one application of TM, in the form of a full-sentence repetitions processor. Our point, however, is that other applications of the same underlying TM data structure are possible, and indeed desirable.[1]

TransSearch constitutes another application of TM technology. *TransSearch* is an interactive, bilingual concordancing tool intended not only for translators working in production mode, but also for lexicographers, terminologists, linguists, technical writers and editors, or anyone who is looking for a way to express an idea in a language that is not his or her mother tongue. In *TransSearch*, it is the user who takes the initiative in submitting queries to the system; and these queries may be single words, phrases, expressions or even full sentences. *TransSearch* exploits the same kind of bi-textual database as the commercial TM products mentioned above. What distinguishes it from these, however, is the manner in which the queries are submitted and the varying size of the units that can be queried.

For over three years now, a Web-based interface to the *TransSearch* system has given Internet users access to a large English-French TM made up of over seven years of Canadian parliamentary debates (better known as the Hansard). Considering how particular these texts are, and the fact that the RALI has made little effort to publicize their availability, the number of queries this site attracts every month is nothing short of amazing. In this paper, we present a brief overview of the *TransSearch* system: how its memory is structured (section 2) and how its content can be queried over the Internet (sections 3 and 4). We have also collected some basic data on who is using the Web-based version of the system and how. An analysis of this data is presented in section 5.

The considerable success of *TransSearch* on the Web demonstrates beyond a doubt that this application of TM corresponds to a real need. We conclude, therefore, with a call to the international community to help set up a network of bi-textual databases like *TransSearch*, which would cover a variety of language pairs in diverse domains.

2. The *TransBase*

The translation memory underlying *TransSearch* is called a *TransBase*. Essentially, it consists of a list of records called couples, each of which comprises a pair of text segments, usually sentences, which are mutual translations. In addition to the text, each couple includes information about the source of the segments (i.e. the name of the document from which the text originated), its position in this document, and the language of each segment. Optionally, the source language of the couple can also be recorded, if this information is available.

Typically, a *TransBase* is created from a collection of pairs of documents. To cast these documents into the *TransBase* mold, the files need to undergo a number of transformations. We review the main steps below.

Input files and formats

First, the collection must be organized into a sequence of pairs of matching files. This will often require renaming the files so that matching pairs of documents can easily be identified. A common convention is to have matching documents share the same prefix and distinguish versions with the use of standard two-letter language codes. This step is currently performed manually, although parts of it could very well be automated. See (Nie et al., 1999) for an example of how this could be done.

Second, all documents must be converted into plain ISO-Latin-1 text. Once again, this operation is not taken care of by *TransSearch* itself, and must be performed manually or otherwise. As a consequence of this format conversion, the *TransBase* does not preserve all the formatting information of the original text, other than spacing and blank lines present in the plain text versions. It also means

[1] Our definition, incidentally, concurs with that of the authors of the EAGLES NLP systems final report, who state that the notion of TM shouldn't be restricted to the systems currently available on the market (EAGLES, 1995).

that "extra-textual" units such as tables, captions, headers, footers, etc. are not dealt with in any special way; and graphics, such as figures and images, are discarded.

Segmentation

Third, each file must be segmented into sentences, or whatever units are deemed appropriate. (As we will see below, limitations on the accuracy of current alignment algorithms makes segmentation into words inappropriate.) This step is not as trivial as it may appear, in part because, there is no real consensus as to what constitutes a sentence. Are titles sentences? What about items in a list? Can sentences be embedded, as "All sentences end with a period." appears to be?

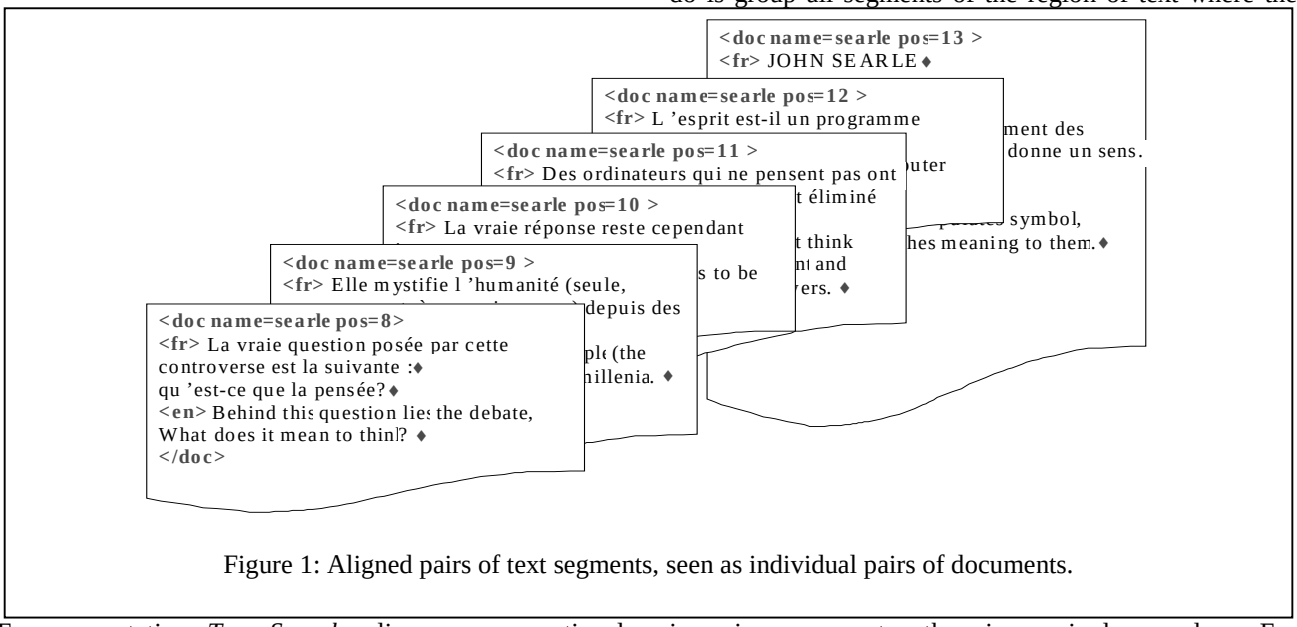

Figure 1: Aligned pairs of text segments, seen as individual pairs of documents.

For segmentation, *TransSearch* relies on an operational definition which, in the majority of situations, seems to coincide with most people's intuitive notion of what sentences are. Without going into the details, we basically treat all periods, ellipsis, exclamation marks, question marks and paragraph boundaries as end-of-sentence markers. Paragraph boundaries are identified by sequences of blank lines (the minimum number of which is an adjustable parameter). The system also relies on a simplistic number grammar, lists of abbreviations and various heuristics to distinguish between full stops and other periods. Finally, the user can specify alternate or additional end-of-sentence markers if, for example, he wishes to consider semicolons as such.

The output of this step is a version of each text file in which sentence boundaries are explicitly marked-up, and where each sentence is assigned a unique identifier (e.g. "`en53`" or "`fr48`").

Alignment

Next, parallel documents need to be aligned. This means that for each pair of files, we must explicitly match each segment (i.e. sentence, title, etc.) in one text with its translation in the other text. As in the case of segmentation, this process is not as trivial as it may appear, even if we restrict ourselves to aligning at the level of sentences. In fact, this problem has been the subject of an abundant scientific literature, beginning with the early works of (Catizone et al., 1989), (Brown et al., 1991), (Gale & Church, 1991), (Kay & Röscheisen, 1993), and the latest evaluation campaigns of the ARCADE project (Langlais et al., 1998).

Once again, without going into the details, *TransSearch's* alignments are produced by a program called *sfial*, which implements a slightly modified version of the method elaborated in (Simard et al., 1992). This simple yet effective method exploits the natural correlation between the lengths of translated segments, as well as the existence of cognate words in pairs of related languages like English and French. One of its limitations, however, is that it cannot explicitly account for inversions in translation, i.e. situations where the order of propositions is not the same in the source and target. In this case, the best that it can do is group all segments of the region of text where the inversion occurs together in a single couple. For sentences, however, such phenomena are quite rare, so that this limitation is tolerable in practice.[2]

The output of this step, for each bilingual document, is a sequence of pairs of sentence identifier lists. For example, one such pair might look like this:

`(en53; fr48, fr49)`

meaning that in some pair of documents, the English sentence with identifier `en53` is translated into French by two sentences, namely `fr48` and `fr49`.

Database

Once the alignments have been produced, it is a trivial matter to convert the documents into couples, as described above, which are then ready to be stored in a textual database. Currently, the *TransSearch* database is implemented using the *MG* (*Managing Gigabytes)* document retrieval system. *MG* is a freely-available indexing and retrieval system for text and images.[3] It basically takes as input an arbitrarily large collection of documents, which it indexes and organizes in a highly compressed form, thereby allowing quick retrievals to be

[2] If near-perfect alignments are required, one can review and manually correct both the automatic alignment and the segmentation using a tool like the RALI's *Aladin* (Lokbani, 2000).

[3] *MG* is covered by a GNU public licence. The current version of *MG* is available via ftp from http://www.cs.mu.oz.au/mg/.

made using various types of queries. *MG* was chosen because of its ability to deal with large quantities of data, and because it is free; but in theory, any document retrieval system with similar functionalities could do the job.

From the point of view of the document retrieval system, each couple produced by the alignment algorithm is considered a small bilingual document (see Figure 1 above). Once all pairs of input documents have been converted into such couples and inserted into the MG database, it is possible to submit standard Boolean, or so-called ranked queries, to retrieve pairs of matching segments. *TransSearch* builds on top of this capability, to allow more sophisticated interactions, as described below.

3. The Queries

The goal of the *TransSearch* system is to allow a user to look for instances of specific words or expressions as they appear in context, and to retrieve their translations as well. For this purpose, standard Boolean or ranked query languages are not enough, because they do not take word order into account. *TransSearch* queries are made up of one or several expressions, intended to match specific portions of text. Each expression is preceded by a language specifier which, as its name suggests, determines which "half" of a couple it can match. The expressions that make up a query are then taken conjunctively; i.e. all of them must match for a couple to be retrieved. Optionally, some expressions (but not all) may be globally negated, so as to act as filtering constraints. Table 1 shows the basic syntax of query expressions, while Table 2 provides some examples.

	expression	matches
1.	w	word-form w exactly
2.	w+	any flexional variant of word-form w
3.	(x_1)	whatever expression x_1 matches
4.	x_1 x_2	expression x_1 immediately followed by expression x_2
5.	x_1… x_2	expression x_1 eventually followed by expression x_2
6.	x_1 .. x_2	expression x_1 followed by x_2, separated by at most 25 characters
7.	x_1 | x_2	expression x_1 or expression x_2

Table 1: Syntax of *TransSearch* Expressions

Several interfaces have been developed for submitting queries to *TransSearch*, including a command-line interface and an X-Windows interface. In this paper, we are focussing on the Web-based version of *TransSearch*, which means that the user interface is a standard Web browser. In this interface, users have access to a subset of the query language described in Table 1: negations and disjunctions are not available, and only one expression per language is allowed; in other words, only the juxtaposition, ellipsis and flexional variants operators are provided. In fact, when the user first accesses the system, the interface only makes available a single, language-independent expression, which *TransSearch* then expands into a pair of identical, language-dependent expressions. To submit his own language-dependent query, the user must explicitly call up the bilingual version of the interface. These restrictions were initially motivated by computational considerations. But, as we will see below, and as has been observed in most public-access information retrieval systems, a large majority of users avail themselves of only a fraction of the system's features.

	expression	matches
1.	en: (banana republic+)	"banana republic" "banana republics"
2.	en: (cut .. bone+)	"cut into the bones" "cut to the bone" "cut the program to the bone" etc.
3.	fr:(déception) en: (deception)	potential examples of this deceptive cognate

Table 2: Examples of *TransSearch* Queries

A snapshot of the Web-based query interface is given in Figure 2 at the end of the paper.

Once a query is submitted, *TransSearch* displays the results on separate Web pages, listing 10 results per page. Each result appears as a pair of segments, in side-by-side format, as in Figure 3. The words matching the query are highlighted, and a "context" hypertext link is included in each row. If this link is selected, a new page appears displaying the same pair of segments but in a larger context (approx. a dozen segments before and after). If the user so wishes, he can scroll through the following or preceding pages of context in the original document.

4. The Corpus

The RALI has produced *TransBases* from various document collections; but in the publicly available, Web-based version of *TransSearch*, the only database that users can query is the Hansard. This *TransBase* is made up of Canadian parliamentary debates covering the period from 1986 to 1993 and totals approximately 50 million words. As for the accuracy of the automatically-produced alignments, our estimates indicate that over 99% of all couples in this *TransBase* are correct. And in our own experience, many of the remaining alignment errors turn out to be inconsequential for this application, involving either under-segmentations (too many sentences are grouped together) or over-segmentations (sentences were accidentally chopped between real sentence boundaries). Only in the latter case may users fail to find the translation of their query.

5. The Users

The *TransSearch* Web page was opened to the public in 1996 as a demonstration of one possible application of our lab's alignment technology. But in fact, the RALI did very little to publicize the availability of *TransSearch*, other than mentioning it at various conferences or presentations. As it turned out, however, the system gradually began attracting an increasing number of users, to the point that we started to worry about the growing burden on our Web server. We therefore decided to add a log file to *TransSearch* in 1997, in order to collect some basic data on who was using the system and how.

The log file

The log file records all the queries submitted to *TransSearch*, along with the number of hits that each query produces. In addition, each log file entry specifies the date and time the query was submitted and the IP address of the machine it was received from. Hence, the log file allows us to keep track of the number of queries processed by the system over time. As can be seen from the graph in Figure 4, this number has been growing steadily, despite occasional and predictable drops around vacation periods. The current peak was reached in November of last year, when *TransSearch* processed over twenty thousand requests in a single month.

The graph in Figure 4 also shows the number of machines from which these requests originated, or, roughly speaking the number of *TransSearch* users.[4] In the same month of November, for example, *TransSearch* processed requests from over 1500 different machines. Of course, the IP address does not allow us to identify the names of these users; nor are we interested in doing so. But by sorting the log file on these addresses, we can tell which users consult the system most frequently, and often what organization and country they come from. As one would expect, given the English-French content of the Hansard database, those who make the most intensive use of the system come from French speaking or bilingual countries like Canada, France and Belgium; but there are also many users from the UK and the former Soviet Union, although users from educational institutions (the "edu" domain name) and various organizations (the "org" domain name) are even more numerous.

None of this is terribly startling. What is somewhat surprising, perhaps, given that our Hansard database contains about 50 million words of English and French text, is the fact that approximately 39% of all the queries submitted return no match. We have subjected these queries to closer scrutiny and found that a good number result from typos of one sort or another, many involving missing accents.[5] Hence, one relatively simple way to improve the system would be to add a language-sensitive spelling checker which would inform the user that the query he has submitted is orthographically ill-formed. Currently, the system responds with an uninformative "no match".

We have also correlated the unsuccessful queries submitted to *TransSearch* with their length in number of words; and what we found, again unsurprisingly, is that the more words a query contains, the more likely it is to come up empty. The figures appear in Table 3 at the end of the paper. What the Table shows, first, is that most queries submitted to *TransSearch* are comprised of two words, followed by 1-word queries, and then 3- and 4-word queries; after this, the numbers begin to drop off quite dramatically. Furthermore, between 1- and 2-word queries, the non-response rate nearly doubles; it then continues to gradually climb until it reaches 100% with the 14- and 16-word queries. We will discuss possible implications of this correlation between query length and the non-response rate in Section 6.

The questionnaire

The log file provides some useful but rather rudimentary data on the queries submitted to *TransSearch*. In order to complete this picture, we decided in late 1999 to add a short questionnaire to the Web site, in which we asked the users a few pointed and more personal questions. However, the questionnaire was entirely anonymous and also optional; people could continue using the system without responding to it. But in order to help us find ways to maintain this service, we felt it was important to obtain more detailed information on our clientele and to elicit as well the users' comments on the system. At the time of this writing, 119 completed questionnaires had been received.[6] Here is what emerged from our analysis of their responses.

The majority of *TransSearch* users are translators (51%) or students (32%), presumably of translation; but there are also a fair number of linguists, terminologists and professional writers (12%). French is the mother tongue of 73% of the users who responded to the questionnaire, and English, 21%. Asked how they had learned about *TransSearch*, 42% answered by word of mouth and 19%, via a Web search engine. But in their comments, several respondents mentioned that they had heard about the system either from their professors or from a professional translators' association.

The questionnaire asks users what they use *TransSearch* for. Among the multiple choice answers, 75% of the respondents selected either to find a translation solution or to verify a translation – again, not very surprising, given the profile of the majority of users. But 10% of the respondents indicated that they also use the system to find monolingual information, such as collocations. This is particularly helpful for people who have to write in a language that is not their mother tongue. And then there are users like the following: "I use your site for spelling, synonyms, wording, cross-checking references to words or expressions, blatant translating, ideas for wordings, definitions, sentence and phrase meanings and probably more that I can't think of right now... "

Not all *TransSearch* users are quite so exuberant, but those who took the trouble of responding to the questionnaire do seem very satisfied with the system: 94% find it very useful or indispensable, only 8% find it adequate, and no one who answered found the system not very useful.[7] On the other hand, 61% of the respondents said they had never consulted the on-line help, suggesting that many users may not be taking advantage of some of the system's more advanced features, such morphological expansion or ellipses. In fact, this is confirmed by the data in the log file: queries that include the "+" operator make up less than 3% of the total, and those with ellipsis (the "…" or "..") less than 6.5%. Moreover, only 0.81% of the

[4] We assume that multiple users on a single machine or single users with multiple machines are more or less exceptional. We also ignore such complicated issues as Internet providers that assign IP addresses dynamically, and so on.

[5] Of the one-word queries that returned no match, two thirds were forms not recognized by either our English or French dictionary.

[6] Fifteen of these, however, originated from an IP address that had already responded to the questionnaire. Either the same user responded twice or two or more people are using the same address. Since the survey was anonymous, it was impossible to tell.

[7] Obviously, this is not a representative sampling. People who have little use for such a system probably would not bother to answer the questionnaire.

total queries are submitted via the bilingual interface, which allows the user to specify a separate expression in each language.

The questionnaire also asked users what improvements they would like to see made to *TransSearch*. Here, 58% of the respondents selected the possibility of consulting databases in other domains, while 27% selected additional databases involving other language pairs. Asked to specify which new domains they would prefer, respondents provided a wide variety of answers, none of which attracted a clear majority – unless informatics is lumped together with the scientific and technical domains (yielding 45%), followed by the financial and economic domains (25%). As for new language pairs, respondents requested *TransBases* that would provide access to Spanish and French translations and Spanish and English translations, followed by translations between German and either French or English.

Implementing any of these improvements to *TransSearch* would cost money. One possible way of financing this work would be to sell subscriptions to the service. We therefore asked users whether they would be prepared to pay in order to retain access to *TransSearch*. Much to our surprise, 61% answered in the affirmative (although it should be mentioned that no subscription rates were specified.) Another way to finance *TransSearch* would be to help large translation services convert their own archives into bi-textual databases. Here, only 11% of the respondents expressed an interest, and of those, only a few have bothered to contact us so far.

Finally, respondents were encouraged to send us their comments and suggestions on the system. Among the suggestions, the request that reoccurred most frequently was to have the Hansard database updated. Other more technical suggestions included displaying the total number of hits for each query; introducing explicit Boolean operators, particularly negation; offering alternate ways of sorting the results; and permitting category matching.[8] Among the comments, the most frequent by far were kudos and messages of gratitude, much too embarrassing for us to repeat here.

6. Discussion

If we return to the distinction made in the Introduction between the two applications of TM – bilingual concordancing and full-sentence repetitions processing – we observe that each effects a certain trade-off between automation and flexibility. Repetitions processing provides a higher level of automation at the expense of a certain rigidity; bilingual concordancing offers greater flexibility in the units that can be submitted to the system, but the user must formulate and submit the queries manually. Furthermore, as we saw in section 5, our data suggests that there is a direct correlation between the length of queries submitted to a TM and the system's non-response rate. On the basis of these findings, one might consider constructing a plausibility argument on the potential usefulness of the two types of TM. The argument would go roughly as follows: Excluding the very special context of document updates, a TM that only operates on complete sentence units will necessarily be of limited utility to translators. For the great majority of translation situations where updates are not involved, a more flexible translation memory that allows users to submit units of any size is likely to prove much more useful. No doubt, this largely accounts for the remarkable success of *TransSearch* on the Web.

Of course, proponents of repetitions processing will respond to this argument by invoking the so-called fuzzy matching capability which allows certain systems to retrieve sentences that are similar (and not just identical) to the input sentence. We have no desire, however, to engage in a polemic on the notion of fuzzy matching;[9] because, ultimately, it may not be necessary to choose one variety of TM over another. For certain types of translations – updates, for example – repetitions processing may well prove very cost effective. For texts that do not contain a high level of repetition, bilingual concordancing will probably prove more productive. That the two are indeed complementary is confirmed by the fact that certain translation support tools, e.g. *Translator's WorkBench*, actually offer both.

The challenge for the user, then, is to learn how to exploit the strong points of each of these and other translation support tools. For example, *TransSearch* is not intended to replace a bona fide terminology bank. The system merely retrieves previous solutions that translators have devised for any number of translation conundrums; unlike a term bank, however, these are not evaluated or commented on by usage experts. On the other hand, translators are likely to find within the enormous databases that *TransSearch* makes available answers to many problems that often aren't catalogued in either term banks or bilingual dictionaries.[10] For, as Pierre Isabelle observed in 1993: "existing translations contain more solutions to more translation problems than any other existing resource."

One of the reasons that translators find the content of the Hansard so useful is that these parliamentary debates range over a wide variety of topics, covering nearly every aspect of Canadian life. This is another respect in which *TransSearch* differs significantly from the TM's of private translation services: the latter tend to be restricted to a more narrow range of topics. Of course, the proceedings of the Canadian Parliament are not unique in this regard. The proceedings of the European Parliament and the European Commission are also systematically translated in multiple languages; indeed, substantial samples of these multilingual parallel corpora are available for research purposes through ELRA.[11] We believe that these texts too would be of great benefit to translators, if they were converted into translation memories on the model of *TransSearch*. In fact, what we would like to see – and we take advantage of this tribune to publicize this call – is the creation of an international network of bi-textual databases like that which *TransSearch* currently offers; a

[8] All but the third of these suggestions have been available at one time or another in previous versions of the system. On the other hand, the kind of sorting suggested – by target equivalent – would not be obvious to implement, requiring, among other things, accurate and reliable word-level alignment.

[9] For an interesting discussion of fuzzy matching and the limits of translation memories that store only strings without performing any linguistic analysis on them, see (Planas & Furuse, 1999).

[10] In particular, translations of the many figurative expressions that abound in natural language.

[11] See http://www.icp.fr/ELRA/ cata/tabtext.html .

network which would cover a variety of language pairs in diverse domains, and which translators around the world could freely access over the Web. The RALI would be most eager to share its expertise with other research groups and international funding agencies who would be interested in launching such a project.

References

Brown, P., Lai, J., Mercer, R. (1991) Aligning Sentences in Parallel Corpora. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley CA.

Catizone, R., Russell, G., Warwick, S. (1989). Deriving Translation Data from Bilingual Texts. In Proceedings of the First International Lexical Acquisition Workshop, Detroit MI.

EAGLES Evaluation of Natural Language Processing Systems, Final Report. (1995) Available online at http://issco-www.unige.ch/ewg95/node140.html

Gale, W. & Church, K. (1991) A Program for Aligning Sentences in Bilingual Corpora. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley CA.

Isabelle, P., Dymetman, M., Foster, G., Jutras, J-M., Macklovitch, E., Perrault, F., Ren, X., Simard, M. (1993). Translation Analysis and Translation Automation. In Proceedings of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation, Kyoto.

Kay, M. & Röscheisen, M. (1993). Text Translation Alignment. Computational Linguistics, 19(1).

Langlais, P., Simard, M., Véronis, J., Armstrong, S., Bonhomme, P., Debilil, F., Isabelle, P., Souissi, E., Théron, P. (1998) ARCADE: A Co-operative Research Project on Parallel Text Alignment. In Proceedings of the First International Conference on Language Resources and Evaluation, Grenada, Spain.

Lokbani, M. (2000). Aladin: An Alignment Management Tool for Translators. To appear in the Proceedings of RIAO-2000, Paris.

Nie, J.-Y., Simard, M., Isabelle, P. and Durand, R. (1999) Cross-Language Information Retrieval based on Parallel Texts and Automatic Mining of Parallel Texts in the Web. Proceedings of SIGIR '99, Berkeley, CA.

Planas, E. & Furuse, O. (1999) Formalizing Translation Memories. In Proceedings of MT Summit VII, Singapore.

Simard, M., Foster, G., Isabelle, P. (1992) Using Cognates to Align Sentences in Bilingual Corpora. In Proceedings of the Fourth Conference on Theoretical and Methodological Issues in Machine Translation, Montreal, Canada.

Number of words in query	Number of queries	Percentage of non-response
1	63978	21.31
2	76306	41.42
3	45152	47.02
4	18177	54.53
5	6139	64.51
6	2231	70.28
7	822	78.47
8	354	84.75
9	135	87.41
10	98	86.73
11	52	90.38
12	36	83.33
13	26	76.92
14	19	100.00
15	10	90.00
16	10	100.00
17	15	93.33
18	6	100.00
19	9	100.00
20	15	86.67

Table 3: Length of queries and non-response rate

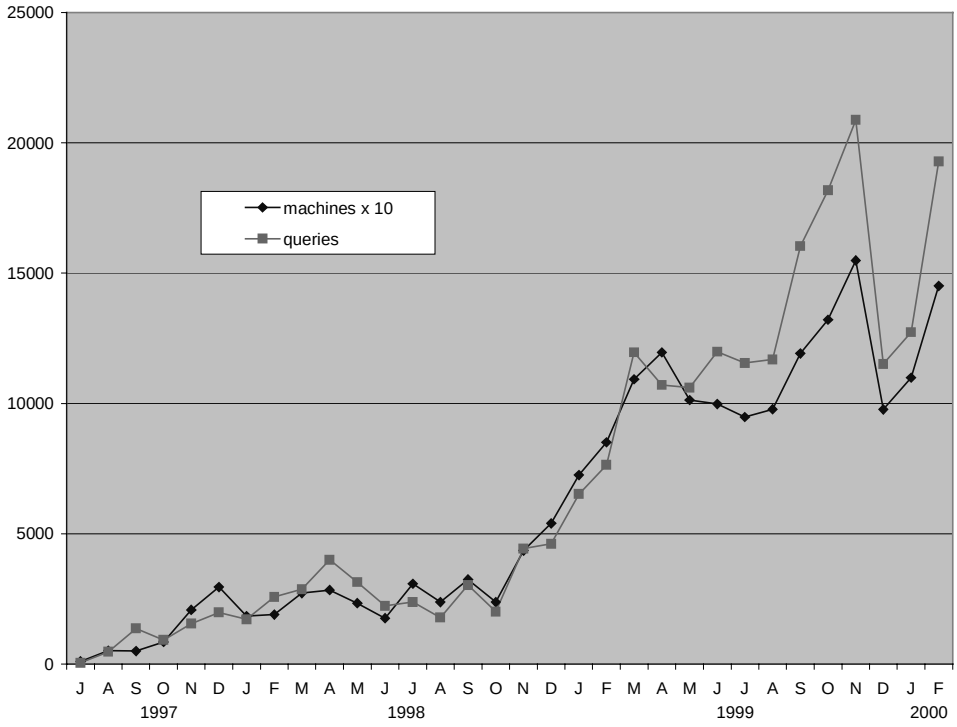

Figure 4: Number of TransSearch Users and Queries

Figure 2: *TransSearch* Query Interface

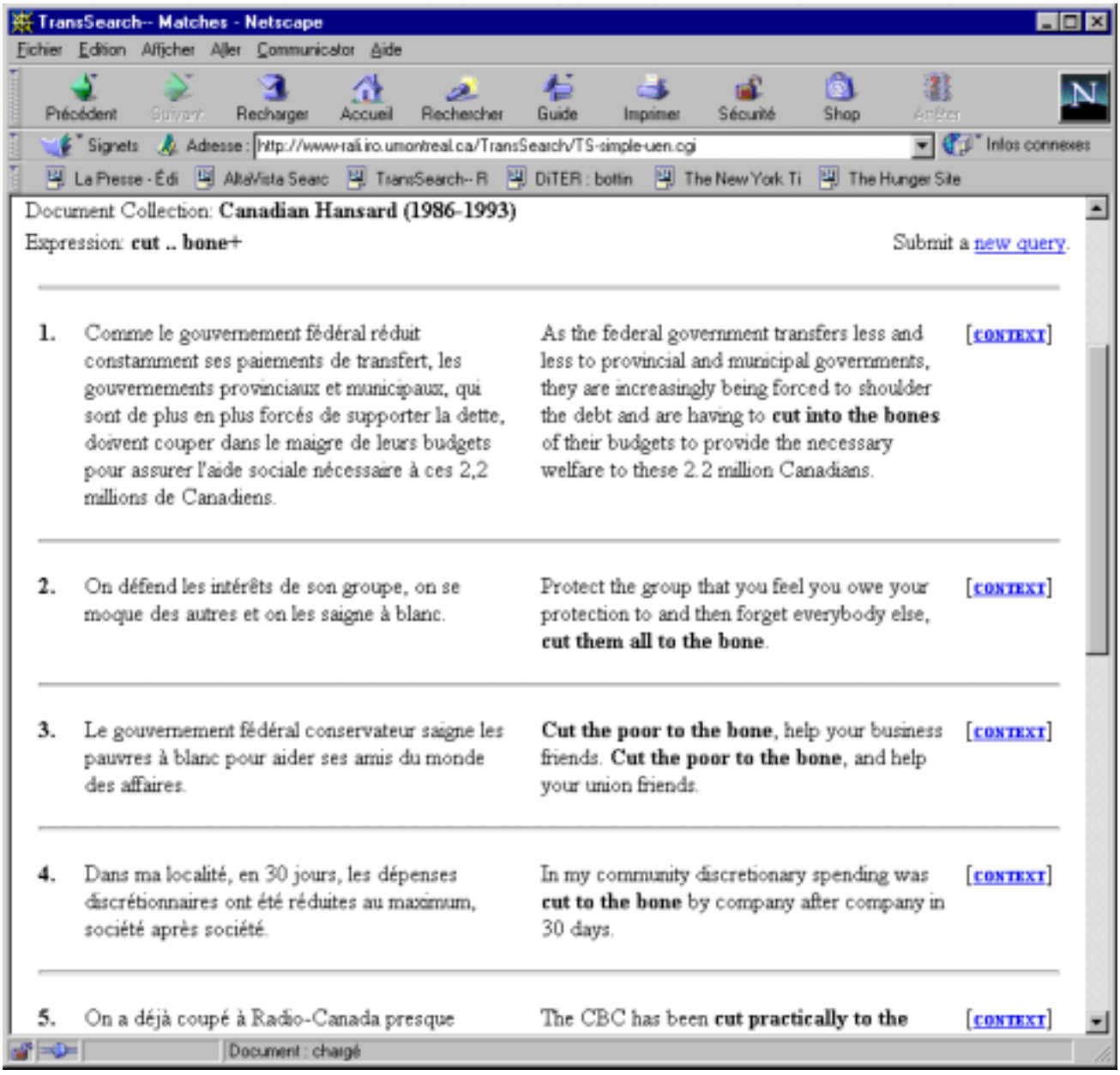

Figure 3: Results of *TransSearch* Query