[*Translating and the Computer 22*: Proceedings of the Twenty-second international conference…16-17 November 2000. (London: Aslib, 2000)]

Two Types of Translation Memory

Elliott Macklovitch[1]
Laboratoire RALI
Université de Montréal

Abstract. Defined most generally, a translation memory is a computerised archive of existing translations, structured in such a way as to promote translation re-use. In this paper, we contrast two types of translation memory - an interactive bilingual concordancer, like the RALI's *TransSearch* system, and a full-sentence repetitions processor, like Trados's *Translator's Workbench* - bringing out the strengths and weaknesses of each. We conclude by examining some of the challenges that will have to met in order to produce more powerful translation memory technology, systems capable of exploiting a larger portion of the knowledge lying dormant in translators' past production.

1 Introduction

The term "translation memory" admits of at least two different definitions, one broad and one narrow. The narrower but more widely used definition corresponds to the characteristics of a popular set of commercial products that includes *Translator's Workbench* from Trados, *Transit* from Star AG, *Déjà-Vu* from Atril and IBM's *TranslationManager/2*. According to this definition, a translation memory (abbreviated henceforth as TM) is a particular type of translation support tool that maintains a database of source and target-language sentence pairs, and automatically retrieves the translation of those sentences in a new text which occur in the database.

The broader definition regards TM simply as a computerised archive of past translations, structured in such way as to promote translation reuse.[2] This definition, notice, makes no assumptions about the manner in which the archive is queried, nor about the linguistic units that are to be searched for in the archive. The narrower definition, by contrast, fixes the sentence as the privileged processing unit of TM systems and presumes automatic look-up as the privileged processing mode. It would thus exclude from the class of translation memories an interactive bilingual concordancing tool like the RALI's *TransSearch* system, where the initiative for querying the archive resides with the user and not the system, and where any linguistic unit - full sentence, word or expression - may be submitted to the system's bi-textual database.

The exclusion of interactive bilingual concordancers from the class of translation memories is entirely unwarranted, in our view. As we will show below, both types of systems exploit the same kind of bi-textual database; all that distinguishes them is the manner in which this database is queried. Indeed, some of the commercial products mentioned above actually include an interactive concordancer alongside their automatic repetitions processor. In this way, when the system finds no match for a complete sentence, the user can manually select a sub-segment of the input sentence and submit it via the concordancer to the database, thereby allowing for greater flexibility and increasing the chances of finding within the archive an already existing solution to a problematic element in a new text.

---

[1] I am indebted to my colleague Graham Russell for input and comments on an earlier version of this paper.

[2] This generic definition of TM is quite similar to that provided in the final report of the EAGLES Evaluation of Natural Language Processing Systems (EAGLES 1995).

In this paper, we take a closer look at the two types of TM technology, in an effort to bring out their respective strengths and weaknesses. We also attempt to elucidate some of the challenges that will have to be met if we are to obtain more powerful and more broadly applicable translation memories.

2 Interactive bilingual concordancers

From the user's point of view, an interactive bilingual concordancer is essentially composed of two components: an interface, which allows the user to interact with the system, formulating and submitting queries and then inspecting the retrieved results; and a database of texts, which will hopefully contain matches to the queries that the user submits. What distinguishes a *bilingual* concordancer from other types of concordance programs is precisely the bilingual nature of this database: it is made up of paired segments of texts in two languages, such that when the user asks to see all the occurrences of a given word or expression in one language, the system can retrieve and display the segments (generally sentences) that contain those words as well as the translation of each corresponding segment in the other language. Brian Harris (1988) has proposed the term "bi-textual" to describe a database that is structured in this way, i.e. a database which explicitly links the corresponding segments of two texts that are mutual translations.

In this paper, we will not be overly concerned with the detailed internal structure of such bi-textual databases, nor with the precise nature of the automatic alignment algorithms that generally serve to populate them.[3] Suffice it to say that the bi-textual databases queried by concordancers are generally created from large collections of pairs of documents that are mutual translations by means of programs that automatically calculate which segments in one text correspond to which segments in the other. To produce a complete and fully explicit bi-text, these programs would have to calculate all the translation correspondences between two texts, down to the level of the word and perhaps even the morpheme. As it turns out, this is an extremely difficult problem, one on which skilled humans are not always able to agree. However, algorithms do exist today which can reliably establish the translational correspondences between two texts at the sentence level, where a given sentence in one text may of course be translated by zero, one or more sentences in the other text. (C.f. Langlais et al. (1998); Simard et al. (1992)).

The results of this alignment operation are generally fed into a database management system which indexes all the forms and stores the data in an efficient manner. In order that previous translations may be recalled from this database, an interactive bilingual concordancer also provides an interface in which the user may formulate and submit queries to the database. What kinds of queries are the users of a bilingual concordance apt to submit? If they are translators, their queries will generally seek to answer the following question: "What translation(s) in target language T have already been proposed for X?", where X corresponds to a particular word, a multiword term or a (possibly discontinuous) expression in a source language S. Otherwise, translators may already have an idea for the translation of X and may wish to know whether that translation Y is attested or frequent.[4] Now for some of these questions, a translator would normally turn to resources other than a bilingual

[3] For the *TransSearch* system, these questions have been dealt with in other papers; see Simard et al. (1993) and Macklovitch et al. (2000).

[4] A bilingual concordancer can also be of help to non-translators who draft texts in a language that is not their mother tongue by providing information on collocations; e.g. which preposition does the verb "consist" govern? For this purpose, users simply ignore the translations in the other language.

concordancer; in particular, where X is a single word or a term, the translator would likely first consult a bilingual dictionary or a term bank. These resources, however, are not always up to date. And even when they are, standard reference works do not systematically catalogue the many figurative expressions that abound in natural language; nor do they reflect the preferences of a particular client, which is one reason why translators often like to examine previous documents translated for that client before undertaking a new text. More generally, we subscribe to Pierre Isabelle's oft-cited observation that "existing translations contain more solutions to more translation problems than any other available resource." (Isabelle et al. 1993) The challenge, of course, is how make all the information lying dormant in past translations readily and easily accessible to translators. It is our contention that an interactive bilingual concordancer like the RALI's *TransSearch* system provides a powerful and flexible way of tapping into the richness of translation archives, converting past translations into an enormous virtual example-based dictionary.

2.1 The Web-based version of *TransSearch*

The RALI has developed a number of distinct interfaces for its *TransSearch* system, including a command line interface, an interface that runs under X-Windows and an html-based interface. These interfaces differ somewhat in the expressive power of the queries they allow, but all can be used to access the same bi-textual databases. In this paper, we will focus on the html-based interface, since this is the best-known version of the system, being freely available to Internet users over the Web.[5]

The RALI has also compiled numerous bi-textual databases for *TransSearch.* The one we make publicly available over the Internet is composed of seven years of Canadian parliamentary debates (commonly known as the Hansard) and totals approximately 70 million words of English and French. The user interface to this version of the system is embedded within a standard Web browser, and so we have somewhat simplified the syntax of its query language in comparison to that of the other interfaces. Figures 1-5 at the end of this paper are snapshots of the Web-based interface and may serve to illustrate the basic functionality of the system.

In Figure 1, the user has submitted a query asking to see all occurrences of the word "take" followed by the word "swipe": the order of two words is important and the restricted ellipsis operator signifies that they may be separated by up to 25 characters. The plus sign appended to the two words is an indicator of morphological expansion, i.e. the user is interested not just in the literal forms "take" and "swipe" but in all their morphological variants. In order to be able to expand the query in this way, *TransSearch* incorporates a complete morphological grammar for both English and French. Notice too that in this simple query interface the user need not indicate the language of the query being submitted.

The results of this query are given in Figure 2, where the expressions in bold in the right-hand column correspond to the portions of the source language segment that match the query. As we can see, these include not just various forms of the words "take" and "swipe" but intervening and unpredictable material between the two. In the left-hand column are found the translations of the sentences *TransSearch* has retrieved containing the expression "take .. swipe"; here we see just a few of the ways that the Parliamentary translators have

[5] http://www-rali.iro.umontreal/TransSearch

come up with to render this expression in French.[6] Hence, for someone who was puzzled by the meaning of this non-literal collocation or was simply seeking a little inspiration for his or her translation, the results in the left-hand column would normally be quite useful. (The reader is invited to compare the entry for "swipe" in a standard bilingual dictionary.)

The query submitted in Figure 3 corresponds to a different type of question: here, what the user wants to know is whether "parler à travers son chapeau" is an attested translation of the English expression "to talk through one's hat". As the results in Figure 4 show, the database does indeed contain paired occurrences of the two expressions; although the user needs to be wary, for on their own such co-occurrences are not sufficient to establish whether this French expression constitutes correct usage. This allows us to illustrate one final feature of the Web-based version of *TransSearch.* Should the user want to verify the larger context in which the retrieved expression occurs, he can click on the CONTEXT link provided with each match. This will bring up a new display in which the English and French versions of that day's debates are displayed in full, in side-by-side format; see Figure 5. On the screen, the alignments between corresponding segments are indicated by matching colours, with the third match in the list now in black italics at the bottom of the page. Inspecting this larger context, we observe that the deputy's English intervention actually contains two errors, one of grammar and one of geography, both of which have been corrected in the translated version.

2.2 Who uses *TransSearch* and how?

The *TransSearch* Web page was opened to the public in 1996, essentially as a demonstration of one possible application of the RALI's research in alignment technology. But in fact, the RALI did very little to publicize the availability of the system, other than mentioning it at various conferences or presentations. Nevertheless, *TransSearch* 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.

This 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 well as the approximate number of system users.[7] As mentioned above, these numbers have been growing steadily, although they do drop predictably during weekends and vacation periods. Currently, *TransSearch* processes over 20 thousand queries a month, originating from about 1500 different users around the world. When we recall how particular the texts are that make up the Web site database - outdated Canadian parliamentary debates! - these numbers are certainly very impressive. In our view, they constitute convincing evidence that this application of TM is responding to a very real need.

The log file also indicates the number of hits *TransSearch* finds for every query submitted. Given the size of our Hansard database - around 70 million words in the two languages - we were somewhat surprised to discover that nearly 39% of all the queries in the log file returned no match. Subjecting these queries to closer scrutiny, we found that a good

[6] This is not quite exact, since we don't know which of the two languages is the source and which is translation. This information is available in the underlying database and appears in other interfaces to *TransSearch.*

[7] Approximate, because it is quite possible for many users (particularly in a network) to channel their queries to the system via one machine; in this case, the log file will only record the IP address of the gateway machine.

number are caused by typos of one sort or another, many involving missing accents.[8] 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 1 below. 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 4 below.

Table 1: Length of *TransSearch* queries and non-response rate

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

3 Automatic repetitions processing

I employ the term "automatic repetitions processing" to refer to the better-known commercial TM products, all of which basically function in the same manner. A new text to be translated is first segmented into units, which are generally sentences but may also include titles, headings, table cells, and other "stand-alone" elements. As the translator works his way through the new text, each successive segment is automatically looked up in a bi-textual database, whose structure is identical in all important respects to that of a bilingual concordancer. When a match is found for a new source language (SL) segment, the system retrieves the associated target language (TL) segment from the database, which the translator

[8] Of the one-word queries that returned no match, two thirds were forms not recognised by either our English or French dictionary. Of course, this doesn't mean that all these are typos; our dictionaries are large but certainly not exhaustive.

may accept as is or alter as necessary. In this way, the vendors of commercial TM systems claim, the translator never has to retranslate the same sentence twice.

3.1 Identical, similar and fuzzy matches

What exactly is meant by the expression "same sentence" in this context? That is, what qualifies as an exact match between a new SL segment and the contents of the TM database? The answer is not as obvious as one might think. For example, are two SL units considered identical if they contain exactly the same wording but differ in their formatting attributes? Some TM systems discard all formatting and store only the plain text content, while others claim to offer the user the choice of whether or not to match on formatting attributes. Does a new sentence match a stored sentence if the wording of the two is identical except for certain non-translatables, e.g. proper names, dates or other types of numerical expressions? Trados' *Translator's Workbench* (henceforth TWB) will in fact treat the two sentences as identical and can, moreover, automatically replace the values of certain non-translatables in the retrieved TL sentence with the appropriate values from the new source sentence.[9] What about two SL sentences that are composed of the same lexical units, although some of these are inflected differently, say, for tense or number? In this case, few of the major TM systems will recognise the two sentence as constituting an exact match. Indeed, as Planas & Furuse (1999) point out, unless a TM system can do morphological analysis, it will have difficulty recognising that sentence (iii) below is more similar to input sentence (i) than sentence (ii) is.[10]

(i) The wild child is destroying his new toy.

(ii) The wild chief is destroying his new tool.

(iii) The wild children are destroying their new toy.

In a sense, such qualifications to the notion of "same sentence" can be seen as attempts by TM developers to come to grips with a fundamental problem faced by this type of repetitions processing technology, and that is that, outside the particular context of document revisions or updates, and perhaps certain types of technical maintenance manuals, the verbatim repetition of complete sentences is relatively rare in natural language texts. Given that the overwhelming demand for translation today is *not* made up of revisions and updates, this imposes a serious limit on the applicability of these systems.

Why can't existing systems retrieve repetitions *below* the level of the full sentence? As the discussion of examples (i-iii) suggests, the bi-textual databases underlying these systems are composed of essentially unanalysed sentences strings. Rather than parsing a sentence into units at a finer level of granularity and attempting to align those units across the two languages, today's TM systems typically accommodate non-identical sentences within the input text by means of some notion of "fuzzy" or approximate matching. How exactly do these fuzzy matching algorithms work? It is difficult to say with certainty because TM vendors, although they do illustrate the concept in their promotional literature and demos, do not generally provide a formal definition of the similarity coefficient that users may specify in order to constrain the search for approximate matches. Hence, it is not at all obvious just how the results of a 70% match will differ, say, from a 74% match or an 81% match. According to Planas & Furuse (1999), "the notion of similarity ... in Trados [is] based on the

[9] Other TM products may be able to do so as well, but we are less familiar with these systems than we are with TWB.

[10] Because (ii) differs from (i) by only 4 letters while (iii) differs from (i) by 9 letters.

number of similar characters" (p.338). While this is undoubtedly true, it is not the whole story, for systems like TWB may lower the value of a match when the stored translation unit has been produced by an automatic alignment program or by a machine translation system, or when the source segment has multiple target equivalents; not to mention the opaque effects of word-order differences on the matching score. Combining several distinct and incomparable factors into a single numerical measure may appear to simplify things for the user; on the other hand, it leaves the user with a vague and ill-defined comprehension of a parameter that is central to the system.

In any event, the important point to underline is that in all cases, what these fuzzy matching algorithms are evaluating is the degree of similarity between **complete sentences.** When no sufficiently close match can be found for a new input sentence, current TM systems are unable to 'back off and retrieve examples of clauses or other major phrases, even though such units may well be present in the database. Allow us illustrate with a simplified, schematised example. Suppose that example (iv) below is a new input sentence made up of twenty words, each five characters long. The TM database contains no exact match for (iv) but does contain the SL sentence in (v). The two sentences, notice, share an identical sub-string, w1-w_5, which in both cases is marked off from the rest of the sentence by a comma. However, since this sub-string contains only 25% of the sentence's total number of characters, it is doubtful that any current TM system would be able to retrieve it among its a fuzzy matches; for users are generally advised not to set the similarity coefficient too low, to avoid being swamped by dissimilar and irrelevant examples.

(iv) $w_1\ w_2\ w_3\ w_4\ w_5, w_6\ .\ .\ .\ w_{20}$

(v) $w_1\ w_2\ w_3\ w_4\ w_5, w_{21}\ .\ .\ .\ w_{35}.$

3.2 Repetitions above the sentence level

Another weakness in current TM systems that can be traced to the nature of the underlying database structure is the fact that in these systems, the very notion of a document is lost. Not only are the segmented units in a new text extracted from their context and submitted to the database in isolation, but the contents of the database are also stored as isolated sentences, with no indication of their place in the original document. As every competent translator knows, however, it is not always possible to translate a sentence in isolation; the same sentence may have to be rendered differently in different documents, or even within the same document, as Bédard (1998) convincingly argues. It is not hard to come up with examples of phenomena that are simply not amenable to translation in isolation: cross-sentence anaphora is one obvious example, but there are many others. Sceptics may argue that such problems are relatively rare, but they are missing the point. In order to evaluate a translation retrieved from memory, translators routinely need to situate that target sentence in its larger context. Current TM systems offer no straightforward of doing this because, unlike full document archiving systems, they archive isolated sentences.

The above-mentioned article by Bédard also contains an interesting analysis of different configurations of repetition, not all of which, he maintains, warrant recourse to a TM system. In particular, if all the repetitions in a text are grouped together in a readily identifiable block, e.g. a page of introduction or the numbered clauses of a boiler-plate contract, or if the repetitions are limited to a small number of sentences each of which reoccurs very often, then there may be more efficient ways to proceed than strict successive sentence-by-sentence processing. Similarly, when an updated document has undergone only a few changes, it will

often prove simpler to use a "diff' or a document comparison program to locate those changes and then modify only the corresponding sentences in the previous translation rather than to resubmit the full document to TM. On the other hand, when the repetitions range over a large number of different sentences and these are dispersed unpredictably throughout the text, the type of repetitions processing that current TM products offer may well constitute the best solution.

To summarise: There is no denying the usefulness of current commercial TM systems, particularly for texts that display a high degree of sentence-level repetition. On the other hand, existing TM systems are certainly far from optimal; in particular, their restriction to complete sentences as the sole processing unit, and their rudimentary character-based algorithms for locating approximate matches means that these systems can exploit only a small part of the translational knowledge lying dormant in past translations.

4 Comparing the two types of translation memory

Before we consider some ways in which current TM technology could be improved, let us step back and attempt to compare the two applications of translation memory that have been described in the preceding sections. One obvious way in which commercial TM systems differ from bilingual concordancers is that the former come embedded within complete translation production environments, either integrated with a word processor or a DTP package, or offering a substitute for the translator's word processor. A bilingual concordancer, on the other hand, is simply a reference tool, functionally on a par with a term bank, which the translator may or may not consult while producing his translation in any environment. This difference in turn correlates with the manner in which the bi-textual databases underlying the two applications are generally built up. As we have seen, the concordancer's databases are normally created in batch mode by automatically aligning large collections of already translated documents. Such alignment programs can also be used to populate the databases of commercial TM systems, but this is not the method that is generally preferred.[11] Rather, it is the translator who usually builds up the database, manually and sentence-by-sentence: whenever he or she is required to furnish a TL equivalent for a novel input sentence, the system links this translation to the SL sentence and stores the combined translation unit in the database as soon as the translator decides to move on to the next segment. This has the obvious advantage of providing highly reliable alignments, and the obvious disadvantage of restricting the size of the memory, at least initially. On the other hand, one shouldn't exaggerate the differences between the two types of TM, which may in certain respects be more apparent than real. One indication of their underlying similarity can be seen from the fact some commercial TM systems also offer a batch mode alternative, in the form of a pre-translate function which automatically inserts TL sentences into a SL document. In this mode, the system is not functioning as a full production environment - no new units are being interactively added to the memory - and its database is being consulted much like that of a term bank or a concordancer.[12]

A different, perhaps more abstract way of comparing the two types of TM is to consider the trade-off each effects between automation and flexibility. Repetitions processing provides

[11] Among other reasons, owing to problems with proprietary formats both of the texts to be aligned and of the database in which they are to be stored.

[12] Nor should too much be made of the automatic insertion of TL sentences into the SL document. Similar proposals have been made in the past for automatically inserting TL terminology into a SL document, c.f. Bédard (1990).

a higher level of automation, albeit at the expense of a certain rigidity. A system like TWB automatically submits each successive segment in a new text to the bi-textual database, thereby ensuring that no repeated sentences are overlooked; however, because only complete sentences are submitted, repetitions of segments below the sentence level will often not be brought to the user's attention. A bilingual concordancer like *TransSearch* offers greater flexibility in the units that can be submitted to the system - provided, that is, that the user manually selects and submits the appropriate segments as queries. With these systems too, therefore, it is almost certain that pertinent information in the database is not being recycled. Invoking the information retrieval concepts of recall and precision, one could say that automatic repetitions processing yields high precision but low recall: high precision, since any result means that the entire query sentence exists (nearly) verbatim in the database and must therefore be relevant; low recall, since other relevant sentences will not be extracted owing to the low rate of full-sentence repetition.[13] The challenge for those interested in developing better TM technology is how to improve recall without significantly diminishing precision.

In section 3.1 above, we mentioned two techniques that already allow some of today's TM systems to enlarge the class of sentences that are recognised as an exact match for a given sentence in the database: ignoring the variants of inflectional morphology and conflating various instances of certain types of named entities.[14] Neither strategy, however, addresses the problem schematised in (iv-v) above, where the repetition is on the **sub-**sentential level. It is our contention that, outside the special context of document updates, most of the useful repetitions in a new text will be found on a sub-sentential level, in the form of phrases or even clauses. The fundamental problem for current commercial TM systems is that they have no conception of syntactic constituency (or even adjacency), employing as they do the character-based notion of edit distance to establish similarity between SL segments.

How would an ideal TM system respond in cases like (iv-v)? When faced with a novel input sentence for which no complete match can be found in memory, an ideal TM system would perform a full syntactic parse of that sentence in order to identify its major constituents; these would then be re-submitted to the database in a second search cycle. What this suggestion amounts to, notice, is an automated and systematic variant of what a human translator does in using an interactive concordancer. Just as with the concordancer, the sentential units of the stored database needn't themselves be parsed; in the event of a match on a SL sub-constituent, the system would still display the complete TL sentence.[15] However, achieving a full and accurate parse of unrestricted text remains a difficult problem, and even when it can be done, it may take a prohibitive amount of time. For this reason, we believe that the most promising strategy for the next generation of TM systems will be to employ various partial parsing or "chunking" techniques (c.f. Skut & Brants 1998; Abney 1996). The input sentence would then be broken down into phrases or simplex clauses which, as above, would be resubmitted to the database, beginning with the longest sub-constituent. Because these are still shorter and inherently less variable than the entire sentence, they are more likely to be present in the TM; and because they correspond to syntactically well-defined

[13] Our analysis of the data in the *TransSearch* log file can be interpreted as indirect evidence of the rarity of full-sentence repetition. As we saw in section 2.2, the longer the query submitted to *TransSearch*, the more likely it is to come up empty.

[14] In many respects, the latter technique resembles the proposal for a Generalized Example-Based MT found in Carbonell and Brown (1999).

[15] Although ultimately, research on finer grained alignments should allow the system to reliably highlight just that portion of the TL sentence which corresponds to the translation of the SL sub-constituent.

expressions, they are more likely than a random sequence of the same length to yield a relevant translation. How the (possibly multiple) results should best be presented to the user is something that remains to be worked out. But in combining the strengths of repetitions processing - i.e. automated and systematic searches - with the principal advantage of interactive concordancing - flexible and variable search units - this proposal does have the potential to extract far more useful information from translators' archives.

5 Conclusion

We have drawn attention to some of the limitations of current TM products, which stem, as we've seen, from the restriction to the complete sentence as the privileged processing unit and the use of edit distance as the main matching criterion. There are good reasons, of course, that justify these choices. Calculating the edit distance between two strings of characters is something that can be implemented rapidly and efficiently (as opposed, say, to producing and comparing two complete parse trees). What is more, accurate alignments of complete sentences can be obtained from translators (almost unwittingly), without doing violence to their normal working habits.[16] Nevertheless, we believe that there is room for more advanced TM technology, particularly by allowing for more linguistically informed search techniques that will facilitate access to past translations at levels other than the full sentence. The pay-offs for such improved TM systems promise to be substantial.

6 References

S. Abney, "Part-of-speech Tagging and Partial Parsing", in K. Church et al. (eds.) **Corpus-Based Methods in Language and Speech,** Kluwer, Dordrecht, 1996.

C. Bédard, "Les mémoires de traduction: une tendance lourde", in **Circuit,** no. 60,1998, pp.25-26.

C. Bédard, "La Pré-traduction automatique (PTA) : un pas en arrière dans la bonne direction?" in proceedings of **Les industries de la langue - Perspectives des années 1990,** Gouvernement du Québec, Montréal, 1991.

J. Carbonell & R. Brown, "Generalized Example-Based Machine Translation", 1999, available online at http://www.cs.cmu.edu/afs/cs.cmu.edu/user/ralf/pub/WWW/ebmt/general.html.

EAGLES Evaluation of Natural Language Processing Systems, Final Report, "Featurization: Design and function of translation memory", available online at http://issco-www.unige.ch/ewg95/nodel40.html, September 1995.

B. Harris, "Bi-text: A New Concept in Translation Theory", **Language Monthly,** no. 54, 1988,pp.8-10.

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

[16] Not to mention the fact that a new text can be automatically segmented into sentential units fairly reliably, using essentially no more than punctuation marks and an extendable list of abbreviations.

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

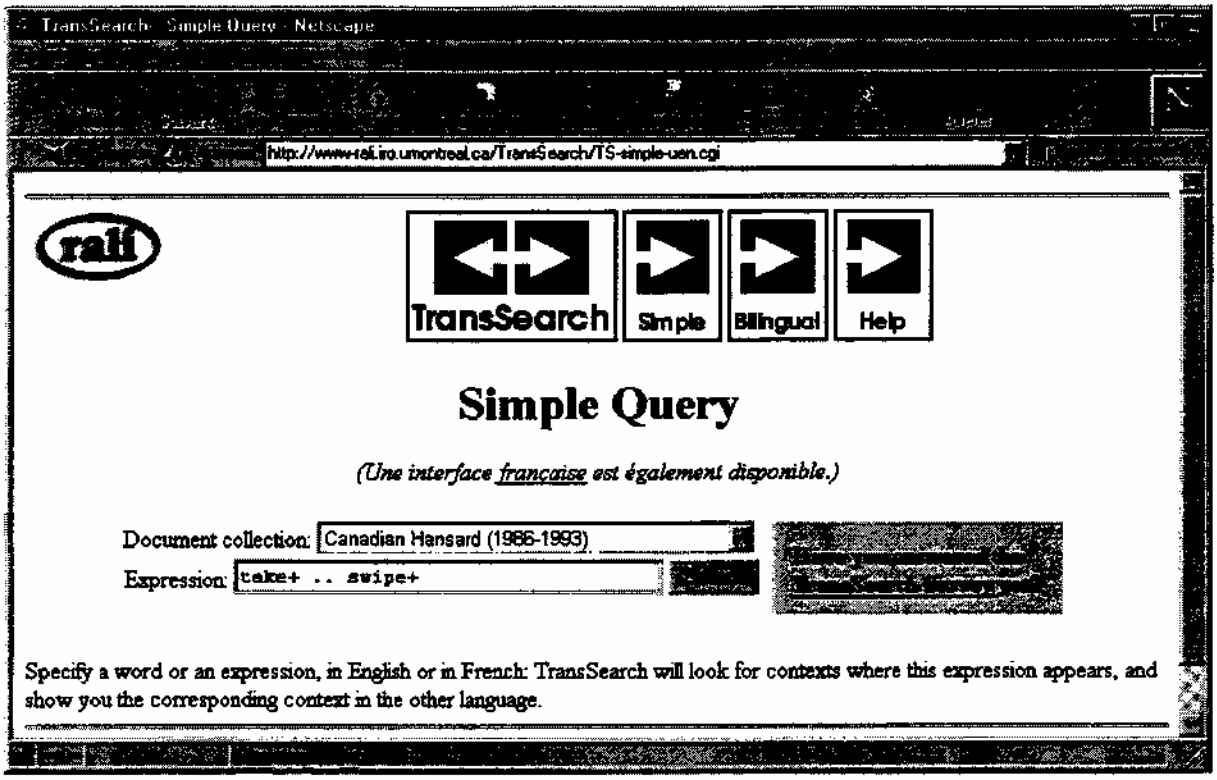
E. Macklovitch, M. Simard, P. Langlais, "TransSearch: A Free Translation Memory on the World Wide Web", in the **Proceedings of LREC-2000,** Athens, May 31 - June 2, 2000.

E. Planas & O. Furuse, "Formalizing Translation Memories", in **Proceedings of MT Summit VII**, Singapore, September 13-17,1999, pp.331-339.

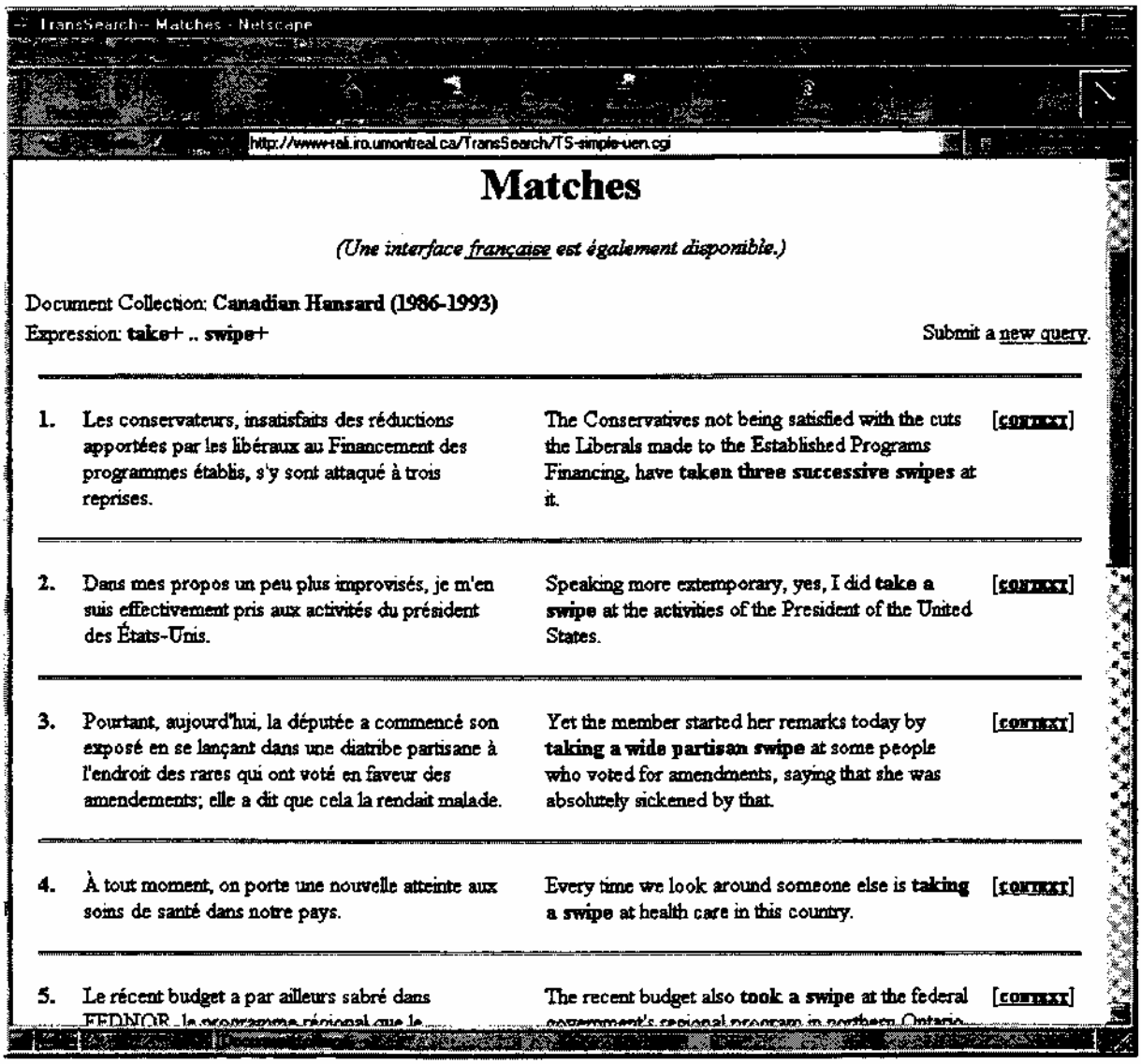
M. Simard, G. Foster, P. Isabelle, "Using Cognates to Align Sentences in Parallel Corpora", in **Proceedings of TMI-92,** Montreal, Canada, 1992

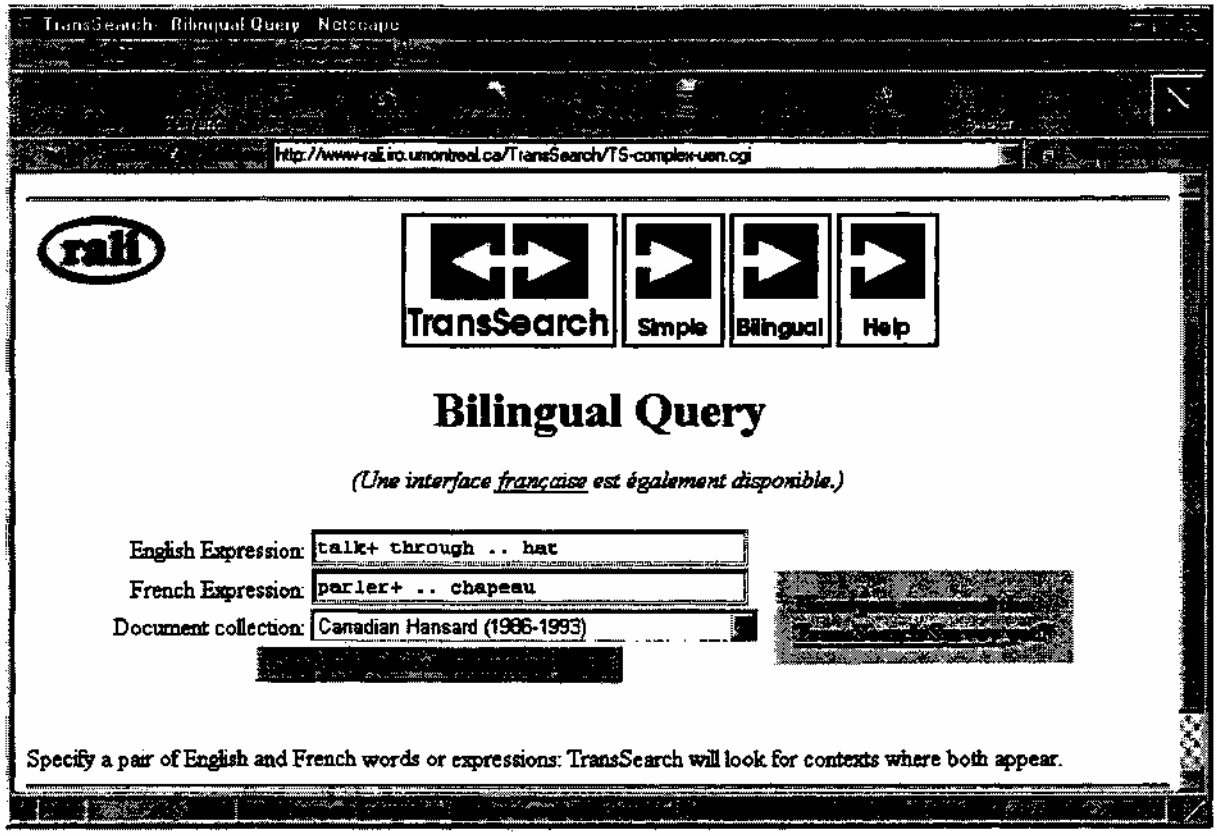
M. Simard, G. Foster, F. Perrault, "TransSearch: A Bilingual Concordance Tool," technical report, Centre for Information Technology Innovation, Industry Canada, 1993.

W. Skuts & T. Brandt, "A Maximum-Entropy Partial Parser for Unrestricted Text", in the **Proceedings of the Sixth Workshop on Very Large Corpora,** Montreal, Canada, 1998.

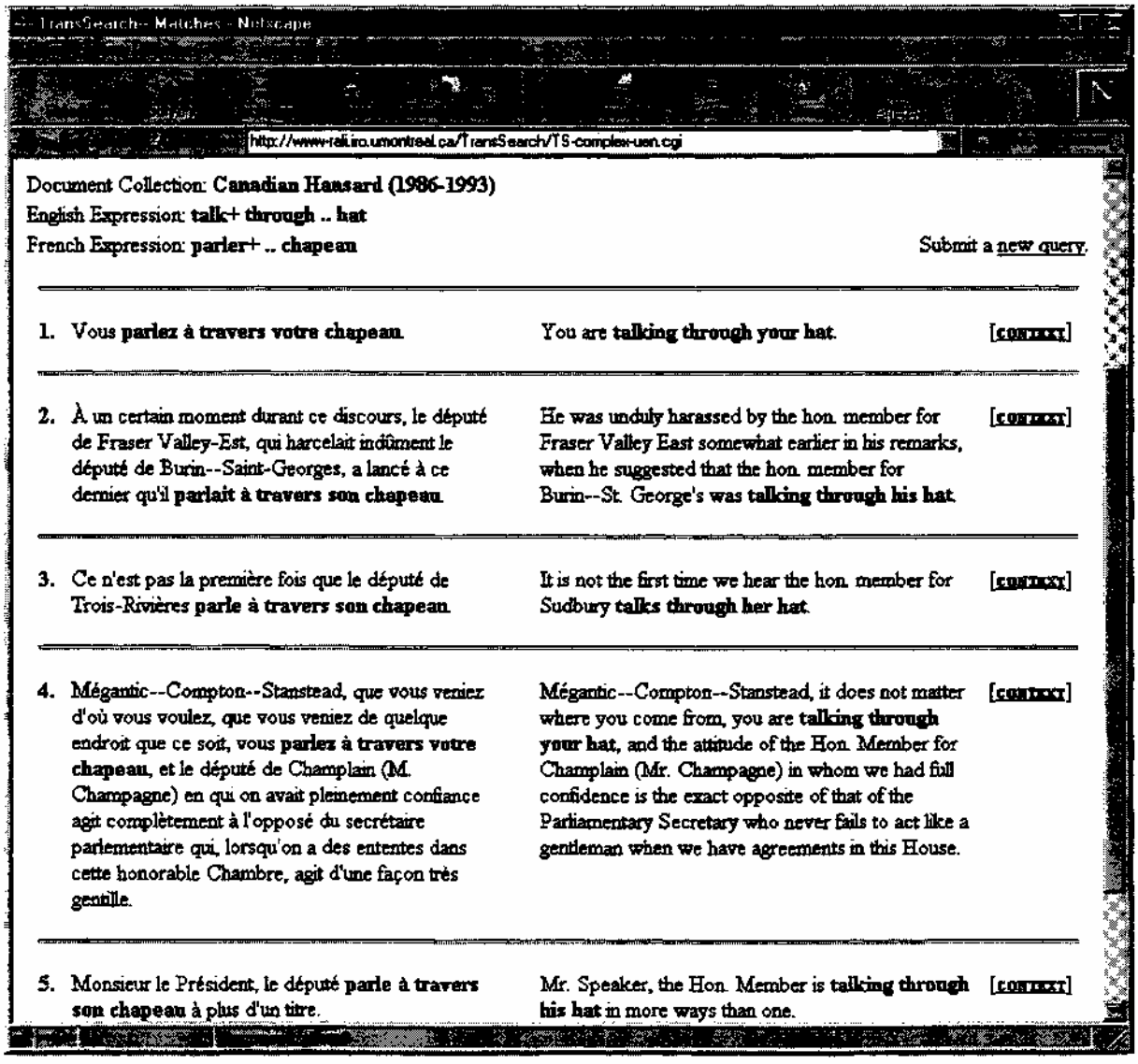

TransSearch - Simple Query - Netscape
http://www-rali.iro.umontreal.ca/TransSearch/TS-simple-uen.cgi
rali
TransSearch
Simple
Bilingual
Help
Simple Query
(Une interface française est également disponible.)
Document collection: Canadian Hansard (1986-1993)
Expression: take+ .. swipe+
Specify a word or an expression, in English or in French: TransSearch will look for contexts where this expression appears, and show you the corresponding context in the other language.

Figures 1 & 2


TransSearch -- Matches - Netscape
http://www-rali.iro.umontreal.ca/TransSearch/TS-simple-uen.cgi
Matches
(Une interface française est également disponible.)
Document Collection: Canadian Hansard (1986-1993)
Expression: take+ .. swipe+
Submit a new query.
1. Les conservateurs, insatisfaits des réductions apportées par les libéraux au Financement des programmes établis, s'y sont attaqué à trois reprises.
The Conservatives not being satisfied with the cuts the Liberals made to the Established Programs Financing, have taken three successive swipes at it.
[CONTEXT]
2. Dans mes propos un peu plus improvisés, je m'en suis effectivement pris aux activités du président des États-Unis.
Speaking more extemporary, yes, I did take a swipe at the activities of the President of the United States.
[CONTEXT]
3. Pourtant, aujourd'hui, la députée a commencé son exposé en se lançant dans une diatribe partisane à l'endroit des rares qui ont voté en faveur des amendements; elle a dit que cela la rendait malade.
Yet the member started her remarks today by taking a wide partisan swipe at some people who voted for amendments, saying that she was absolutely sickened by that.
[CONTEXT]
4. À tout moment, on porte une nouvelle atteinte aux soins de santé dans notre pays.
Every time we look around someone else is taking a swipe at health care in this country.
[CONTEXT]
5. Le récent budget a par ailleurs sabré dans
The recent budget also took a swipe at the federal
[CONTEXT]



Figures 3 & 4



Document Collection: **Canadian Hansard (1986-1993)**
English Expression: **talk+ through .. hat**
French Expression: **parler+ .. chapeau**

Submit a new query.

1. Vous **parlez à travers votre chapeau**.	You are **talking through your hat**.	[CONTEXT]
2. À un certain moment durant ce discours, le député de Fraser Valley-Est, qui harcelait indûment le député de Burin--Saint-Georges, a lancé à ce dernier qu'il **parlait à travers son chapeau**.	He was unduly harassed by the hon. member for Fraser Valley East somewhat earlier in his remarks, when he suggested that the hon. member for Burin--St. George's was **talking through his hat**.	[CONTEXT]
3. Ce n'est pas la première fois que le député de Trois-Rivières **parle à travers son chapeau**.	It is not the first time we hear the hon. member for Sudbury **talks through her hat**.	[CONTEXT]
4. Mégantic--Compton--Stanstead, que vous veniez d'où vous voulez, que vous veniez de quelque endroit que ce soit, vous **parlez à travers votre chapeau**, et le député de Champlain (M. Champagne) en qui on avait pleinement confiance agit complètement à l'opposé du secrétaire parlementaire qui, lorsqu'on a des ententes dans cette honorable Chambre, agit d'une façon très gentille.	Mégantic--Compton--Stanstead, it does not matter where you come from, you are **talking through your hat**, and the attitude of the Hon. Member for Champlain (Mr. Champagne) in whom we had full confidence is the exact opposite of that of the Parliamentary Secretary who never fails to act like a gentleman when we have agreements in this House.	[CONTEXT]
5. Monsieur le Président, le député **parle à travers son chapeau** à plus d'un titre.	Mr. Speaker, the Hon. Member is **talking through his hat** in more ways than one.	[CONTEXT]

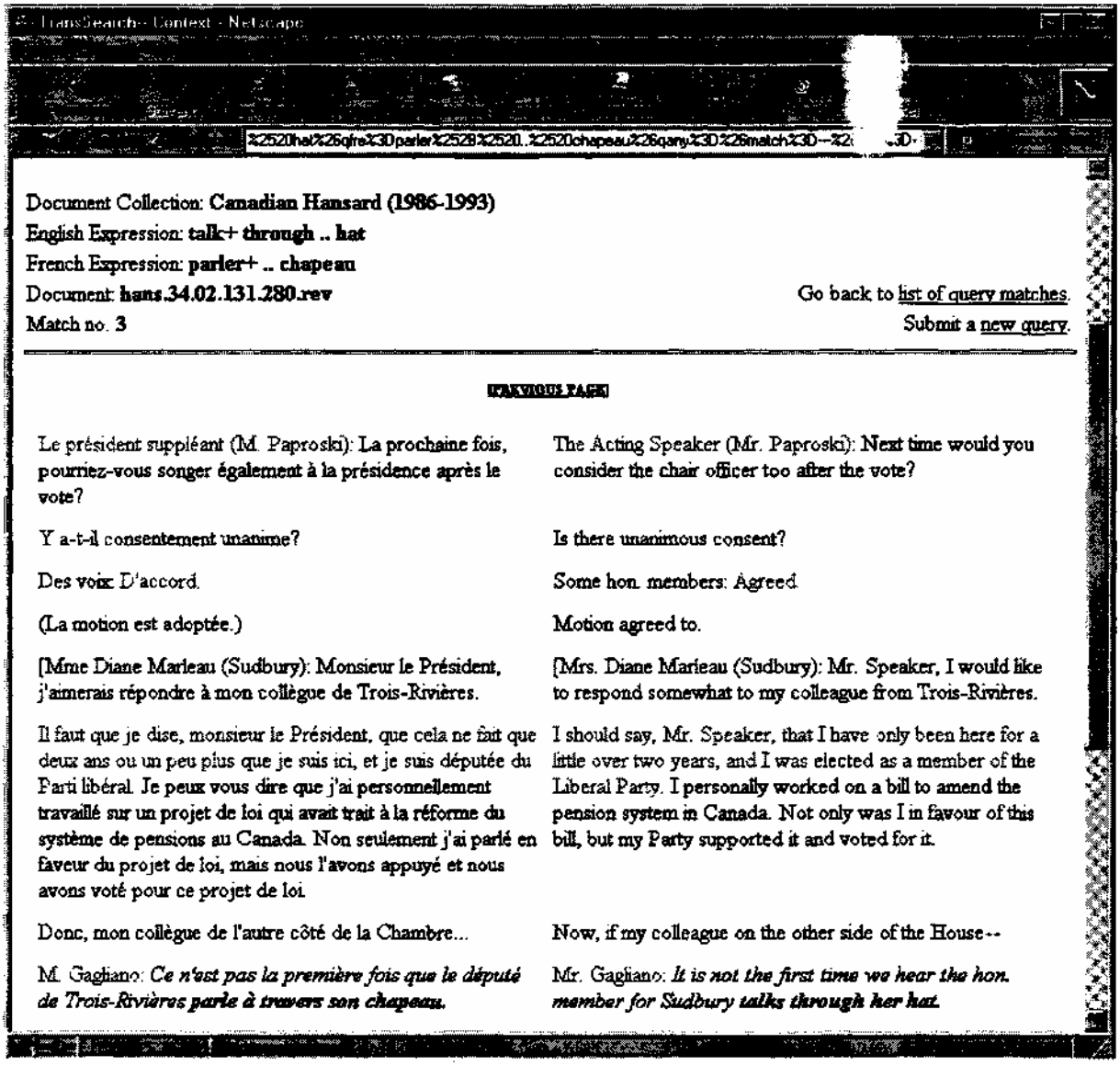

TransSearch -- Context - Netscape
%2520hat%26qfre%3Dparler%2528%2520..%2520chapeau%26qany%3D%26match%3D--%2
Document Collection: Canadian Hansard (1986-1993)
English Expression: talk+ through .. hat
French Expression: parler+ .. chapeau
Document: hans.34.02.131.280.rev
Match no. 3
Go back to list of query matches.
Submit a new query.
[PREVIOUS PAGE]
Le président suppléant (M. Paproski): La prochaine fois, pourriez-vous songer également à la présidence après le vote?
The Acting Speaker (Mr. Paproski): Next time would you consider the chair officer too after the vote?
Y a-t-il consentement unanime?
Is there unanimous consent?
Des voix: D'accord.
Some hon. members: Agreed.
(La motion est adoptée.)
Motion agreed to.
[Mme Diane Marleau (Sudbury): Monsieur le Président, j'aimerais répondre à mon collègue de Trois-Rivières.
[Mrs. Diane Marleau (Sudbury): Mr. Speaker, I would like to respond somewhat to my colleague from Trois-Rivières.
Il faut que je dise, monsieur le Président, que cela ne fait que deux ans ou un peu plus que je suis ici, et je suis députée du Parti libéral. Je peux vous dire que j'ai personnellement travaillé sur un projet de loi qui avait trait à la réforme du système de pensions au Canada. Non seulement j'ai parlé en faveur du projet de loi, mais nous l'avons appuyé et nous avons voté pour ce projet de loi.
I should say, Mr. Speaker, that I have only been here for a little over two years, and I was elected as a member of the Liberal Party. I personally worked on a bill to amend the pension system in Canada. Not only was I in favour of this bill, but my Party supported it and voted for it.
Donc, mon collègue de l'autre côté de la Chambre...
Now, if my colleague on the other side of the House--
M. Gagliano: Ce n'est pas la première fois que le député de Trois-Rivières parle à travers son chapeau.
Mr. Gagliano: It is not the first time we hear the hon. member for Sudbury talks through her hat.

Figure 5