Studying the Human Translation Process Through the *TransSearch* Log-Files

Michel Simard
Xerox Research Centre Europe (XRCE)
6, Chemin de Maupertuis
38240 Meylan France
michel.simard@xrce.xerox.com

Elliott Macklovitch
Laboratoire RALI
Université de Montréal
C.P. 6128, succursale Centre-ville, Local 2241
Montréal (Québec) Canada H3C 3J7
macklovi@iro.umontreal.ca

Abstract

This paper presents the *TransSearch* log-files. These are records of interactions between human translators and *TransSearch*, a bilingual concordancing system. The authors show how this data can be used as experimental evidence to study the translation process. This is exemplified by the results of a study on the nature of the text units on which human translators operate, based on this data. Finally, some enhancements to the *TransSearch* system are proposed, aiming both at improving its usefulness for the end-users and the quality of the data that can be collected from its log-files.

Introduction

TransSearch is a bilingual concordancer: it allows one to query a large corpus of French-English texts, so as to view occurrences of specific words or expressions within their bilingual context. The system is accessed over the Internet; the database is centralized, and users submit queries using a Web browser. *TransSearch* processes thousands of such queries every day, submitted by professional translators, looking for solutions to specific, real-life translation problems.

All of these queries are recorded, along with some related information: who submitted it, when, from what computer, how many matches were found and displayed, etc. The *TransSearch* log-files, as they are called, literally contain millions of such queries. It occurred to us that this data offered the chance for a new and original perspective on how translators operate. The goal of this paper is to present this exciting new data, and to show how it can be used to learn things about translation and translators.

In the first section, we give an overview of *TransSearch*: what it is, where it comes from, how it is used, and what its log-files contain. Then in the second section, we present the results of a study about the linguistic nature of *translation units* which was conducted using this data. Finally, in the last section, we discuss how an improved *TransSearch* system could both better fill the needs of its users and provide us with richer data to study the translation process.

TransSearch

The System

TransSearch is a bilingual concordancer. It allows one to query a large database of bilingual text, aligned at the sentence level (i.e., *bitext*). Queries can be single words or more complex expressions, referring to groups of words, contiguous or not. Figure 1 summarizes the various types of queries allowed. The user is not required to specify the language of the query, but it is possible to do so, as it is possible to submit “bilingual” queries, i.e. queries formed of a pair of expressions, one in each language, both of which must be matched for the query to succeed.

When a query is submitted, the system searches its database and displays each matched expression within its context (usually, a sentence), as well as the translation of this context. This way, users can see actual uses of the word or expression in context, and how they are translated.

A more complete description of the system can be found in Macklovitch, Simard, & Langlais (2000).

A Brief History

TransSearch is actually the most straightforward application of *translation analysis*, a concept that was initially proposed by Isabelle et al. (1993). The first implementation was done at the CITI, a research facility of the Canadian government. The original *TransSearch* was a standalone program: all processing was performed on the user’s machine, and the database was required to reside on a local server.

The advent of the Web gave *TransSearch* a second life. In the mid-nineties, a new version of the system was produced, this time using a Web browser as the basis for its user interface. In this new version, the bitext collection was stored on a single server, which handled all queries through an HTTP server. This version was made accessible to the public, as an on-line demo of the CITI’s research activities.

By the year 2000, *TransSearch* was still up and running, although by then it had moved to the RALI laboratory, at the University of Montreal. The creators of the system were then faced with something of a dilemma: on the one hand, they were being pressured to pull the plug on the system, for various practical and financial reasons; but on the other hand, every day, this “demo” was actually receiving thousands of queries to process. Obviously, *TransSearch* was

query	*example*	*match*
Single-word query: to look up a word *verbatim*	`dust`	*the **dust** settled down*
Word sequence query: to look up contiguous words	`bite the dust`	*I saw him **bite the dust***
Inflection match operator (`+`): to find all inflected forms of a word	`bite+ the dust`	*He really **bit the dust***
Long (`...`) and short (`..`) ellipsis operators: to find non-contiguous sequences of words	`bite+ ..  dust`	*He **bit** the proverbial **dust***

Figure 1: *TransSearch* queries

match	*source*	*target*
1.	Members on that side of the House started **ragging the puck**.	Les députés d'en face ont commencé à tricoter avec la rondelle.
2.	Mr. Speaker, being a former hockey player I was used to **ragging the puck** whenever I was able to get it.	Monsieur le Président, en tant qu'ancien joueur de hockey, j'ai l'habitude de taquiner la rondelle chaque fois que j'en ai la chance.
3.	They are trying to rag the puck just as the Detroit Red Wings tried to **rag the puck**.	Nos vis-à-vis tricotent avec la rondelle en quelque sorte à l'instar des Red Wings de Détroit.
4.	...	...

Figure 2: Results for the *TransSearch* query "`rag+..puck`"

filling a need for a good number of people. Maybe it would be possible to ask these people for help in maintening the service? Thus was born the idea of turning *TransSearch* into a commercial service.

And so, in April 2001, *TransSearch* became something of a business[1]. Since that date, users have been required to pay an annual subscription fee, in exchange for which they can query the system's databases as much they want. The two main databases are made up of French-English parliamentary debates (the *Canadian Hansard*) and of a collection of legal documents (the *Canadian Court Rulings*). Both databases are periodically updated with recent documents.

In late 2003, the management of *TransSearch* was finally transferred to a private-sector partner, *Terminotix*[2], an Ottawa-based company that specializes in computer-assisted translation software.

The *TransSearch* Log-Files

Although the switch to a subscription-based service was undoubtedly a disappointment for many faithful users, *TransSearch* still manages to attract a loyal and reasonably numerous clientele. There are no official statistics, but it is quite safe to assume that the vast majority of users are professional English-French translators, who see *TransSearch* as an essential component of their toolbox. Current clients include the Canadian government's Translation Bureau (one of the largest translation services in the world), as well as many of the most important translation services in Canada. Some individual users just submit a few queries every now and then, while others literally bombard the system with dozens and sometimes hundreds of queries every day. As far as we can guess, users submit their queries in the natural course of their work, as they encounter translation difficulties.

As it turns out, each one of these queries is recorded in a log-file, along with a number of information items: when it was submitted, by whom (because the service is subscription-based, it is possible to know "who" the users are, insofar as this makes sense in this world of electronic business), from where, how many results were returned, etc. The actual results of each query are not recorded, but given all the logged information, these can easily be reproduced. Figure 3 shows a snippet from one of these files.

Each month sees tens of thousands of queries recorded this way. As a result, the log-files collected over the past few years make up an impressive collection of past user interactions. Initially, this data had mostly been used for administrative purposes (e.g. tracking fraudulent users). But it eventually occurred to us that this data might offer a fresh perspective on the kinds of problems which human translators are routinely faced with and on how they formulate these problems. In other words, a new way of looking at the human translation process.

Case Study: Translation Units

As an example of how data from the *TransSearch* log-files can be used, we describe here a study which we conducted, in order to verify some hypotheses about the notion of *translation units*[3]. That human translators do not translate text word-by-word is something of a truism; however, the actual nature of the units on which translators operate is not so easily pinned down. This question bears theoretical importance from the point of view of translation studies and the cognitive sciences. But it also has practical implications, as it hints at the type of units on which computer-assisted and automatic translation systems should operate.

For this study, we focused our attention on queries submitted during the week beginning November 3, 2002. Table 1 gives some overall statistics about the contents of the log-file for this period.

[1] www.tsrali.com

[2] www.terminotix.com

[3] A more detailed account of this study can be found in the fist author's doctoral thesis (Simard 2003)

```
2002/11/03-15h00:51 16278 <user-info> Submit:  hansard (min = ) "e:", "f:", "x:which will not go away"...
2002/11/03-15h00:52 16278 <user-info> Submit:  ...1 matches.
2002/11/03-15h00:58 16284 <user-info> Submit:  juridique (min = ) "e:", "f:", "x:which will not go away"...
2002/11/03-15h00:58 16284 <user-info> Submit:  ...0 matches.
2002/11/03-15h01:01 16292 <user-info> Submit:  hansard (min = ) "e:", "f:", "x:which will not go away"...
2002/11/03-15h01:02 16292 <user-info> Submit:  ...1 matches.
2002/11/03-15h01:10 16300 <user-info> Submit:  hansard (min = ) "e:", "f:", "x:as such"...
2002/11/03-15h01:28 16306 <user-info> Submit:  juridique (min = ) "e:", "f:", "x:settlement arrangements"...
2002/11/03-15h01:28 16306 <user-info> Submit:  ...1 matches.
2002/11/03-15h01:35 16300 <user-info> Submit:  ...10 matches.
2002/11/03-15h01:46 16310 <user-info> Submit:  hansard (min = ) "e:", "f:", "x:consisting of"...
2002/11/03-15h01:47 16310 <user-info> Submit:  ...10 matches.
2002/11/03-15h02:31 16314 <user-info> Submit:  hansard (min = ) "e:", "f:", "x:along with"...
2002/11/03-15h02:32 16314 <user-info> Submit:  ...25 matches.
2002/11/03-15h02:36 16318 <user-info> Submit:  hansard (min = ) "e:", "f:", "x:accorder+..parole"...
2002/11/03-15h02:37 16318 <user-info> Submit:  ...10 matches.
2002/11/03-15h03:03 16325 <user-info> Submit:  hansard (min = 60408) "e:", "f:", "x:accorder+..parole"...
2002/11/03-15h03:06 16325 <user-info> Submit:  ...10 matches.
2002/11/03-15h03:09 16331 <user-info> Submit:  hansard (min = ) "e:", "f:", "x:questioned whether"...
2002/11/03-15h03:10 16331 <user-info> Submit:  ...10 matches.
2002/11/03-15h03:38 16335 <user-info> Submit:  hansard (min = 122389) "e:", "f:", "x:accorder+..parole"...
```

Figure 3: A short extract from a *TransSearch* log-file

Total nb. of queries	29,350	
Hansard	24,937	(84.96%)
Court Rulings	4413	(15.04%)
queries :		
contiguous sequences	27,566	(94.72%)
non-contiguous sequences (ellipsis operators “...” and “..”)	1549	(5.27%)
inflection match operator (+)	983	(3.34%)
language:		
unspecified language	28,097	(95.73%)
English	826	(2.81%)
French	191	(0.65%)
bilingual (French and English)	236	(0.80%)
productive queries	17,995	(61.31%)
results per query	5.22	

Table 1: Statistics on *TransSearch* queries (week of November 3, 2002)

General Observations

As can be seen, most queries were submitted to the *Hansard* database. Indeed, this database is much larger and more general than the other. Furthermore, the vast majority of queries (95,5%) were submitted in a language-independant fashion, i.e. without the user specifying whether the query was in French or English. Both of these aspects (database and language) reflect the default settings in the user-interface.

For our purposes, perhaps the most striking result lies in the queries themselves: over 85% of all queries were composed of expressions of two or more words. Figure 4 shows the distribution of queries as a function of the number of words they contain. Queries consisting of two words are by far the most frequent, followed by 3-words, 1-word and 4-word queries. There are two possible interpretations of this: either *TransSearch* users rarely encounter “lexical” problems (i.e. concerning a single, isolated word), or they turn to other resources when this happens, e.g. dictionaries, glossaries, thesauri, terminology banks, etc. What is clear, however, is that multi-word translation problems are the number one motivation for using *TransSearch*.

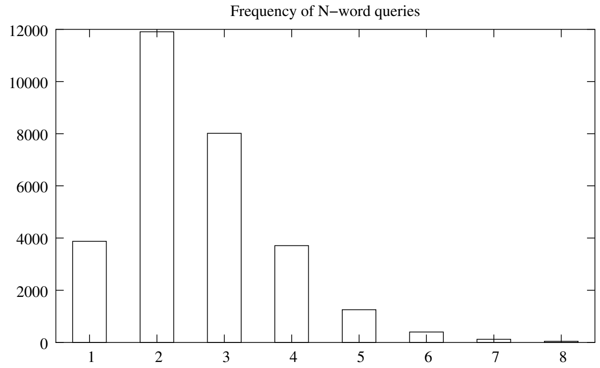

Figure 4: Distribution of queries as a function of their size

Also worth noting is that language-specific and bilingual queries are very seldom used (less than 5% of all queries). The same holds for the ellipses and inflection-match operators, which can be seen as *generalization devices*, i.e. they allow users to formulate a problem in ways less specific than how it originally appeared. For example, a translator stumbling upon a text containing the expression *reap the benefits*, in the following context:

> *...we will not be able to* ***reap the*** *kind of economic* ***benefits*** *that we wish...*

has much to gain by expressing his problem in more general terms, for example with a query like “`reap+..benefits`”. As it turns out, they very seldom do.

Here, again, we can find many explanations. The simplest would be that most problems that translators encounter are not of this form. Or maybe queries are submitted in a very spontaneous way, under pressure, so that most users simply do not have the time or freedom to perform this sort of mental generalization. Queries would then be submitted in their simplest and most direct way, maybe even by cutting-and-pasting portions of a document into the *TransSearch* query field. Finally, one suspects that many users don’t know or fully understand how to use the more elaborate search mechanisms. This is a factor that should not be underestimated.

Translation Units

A number of recent psycholinguistic studies have focused on the question of the linguistic nature of translation units on which human translators naturally operate. These studies tend to show that these units coincide more or less with syntactic constituents (Bernardini 2001). Without going as far as that, we have attempted to measure to what extent *TransSearch* queries match *syntactic chunks*, as defined by Abney (Abney 1991; 1992), or sequences thereof. According to Abney, chunks play an important role in establishing a link between classical constituent structure and prosodic phenomena; but more generally, they make up highly cohesive elements in human discourse. Therefore the question arises: if this cohesion phenomenon is observable in speech, can we also observe it in the translation process?

Of course, chunking is naturally observed over complete linguistic utterances, such as sentences. As a result, it is not possible to establish *a priori* if a certain query is or is not a chunk: for this, it must be seen within its context. “*The great outdoors*” may look like a nominal chunk at first sight, until you see it in this context:

> *...So I sent Alexander the great outdoors.*

The only way to establish with certainty the linguistic nature of a query submitted to *TransSearch* would be to examine the context in which it originally appeared to the user. Unfortunately, we have no way to access this information, because we don’t have access to the source-language text that the translator is working on. In fact, for all we know, some queries might not even appear in any specific context: maybe some users just spontaneously invent queries. What we *can* do however is examine the context within which these queries appear in the *TransSearch* databases.

We conducted a study along these lines: we resubmitted a large number of English queries from the log-files and examined the context within which matches occurred, to verify whether or not these matches coincided with chunks or sequences of chunks. Here in more detail is how we proceeded:

- We considered only the 20,000 first queries submitted during the week of November 3, 2002. We felt that this made up a large enough sample;[4]
- We then excluded all queries that contained only a single word;
- The remaining queries were resubmitted to *TransSearch*, each on the database that was initially specified by the user (*Hansard* ou *Court Decisions*);
- For each query, this process returned a number of matching pairs of sentences. From these, we retained only the 10 first results for which the query matched the English “side”.
- We then proceeded to automatically chunk these English sentences. This was done using a statistical chunker developed at the RALI laboratory, following the method proposed in (Osborne 2000);

We report the main results from this experience in Table 2.

Queries:		
considered	20,000	
submitted (2+ words)	16,230	(81.15%)
results	63,352	(3.9 per query)
Average match size:		
number of words	2.82	
number of chunks	2.00	
Boundary matches:		
on the left	34,685	(54.75%)
on the right	53,399	(84.29%)
both	28,838	(45.52%)

Table 2: Results of chunking *TransSearch* matches

The most important results in this table appear under the heading *boundary matches*. These figures refer to situations where the beginning (left-hand side) or the end (right-hand side) of a query coincide with the boundary between two syntactic chunks. Figure 5 gives some examples of such matches. (Table 3 gives the meaning of chunk labels, as given in Bies *et al.* 1995.)

responsible and accountable:

[$_{NP}$ Exempt] [$_{NP}$ it] [$_{PP}$ from] [$_{NP}$ the things] [$_{NP}$ that] [$_{VP}$ are appropriate and leave] [$_{NP}$ it] [$_{ADJP}$ **responsible and accountable**] [$_{PP}$ for] [$_{NP}$ the rest]

full marks:

[$_{NP}$ **Full marks**] [$_{PP}$ for] [$_{NP}$ that lady] and [$_{NP}$ **full marks**] [$_{PP}$ for] [$_{NP}$ the business]

satisfy the condition:

[$_{NP}$ Each new member] [$_{VP}$ must] [$_{VP}$ **satisfy**] [$_{NP}$ **the condition**] [$_{PP}$ of] [$_{VP}$ carrying] [$_{PP}$ on] [$_{NP}$ business] [$_{PP}$ in] [$_{NP}$ common] [$_{PP}$ with] [$_{NP}$ a view] [$_{PP}$ to] [$_{NP}$ profit]

there is evidence of:

[$_{NP}$ **There**] [$_{VP}$ **is**] [$_{NP}$ **evidence**] [$_{PP}$ **of**] [$_{NP}$ substantial implementation failure]

Figure 5: Examples of query matches that coincide with chunk boundaries

The most intriguing aspect of these results is the apparent asymmetry between boundary matches to the right and to the left of the query. While the end of a query almost always matches a chunk boundary (close to 85%), the proportion for the beginnings of queries drops below 55%. This difference is even more surprising if one takes into account the average length of the chunks, which is 1.55 words in our sample. Under these conditions, the probability that either end of an arbitrary subsequence coincide with a chunk boundary

[4] In fact, separate evaluations on much smaller samples yielded comparable results.

Chunk label	meaning
ADJP	Adjectival phrase
ADVP	Adverbial phrase
CONJP	Conjunctival phrase
INTJ	Interjection
NP	Nominal phrase
PP	Prepositional phrase
PRT	Particle
SBAR	Relative or subordinate
UCP	Dissimilar coordinated phrase
VP	Verbal phrase
UNK	Unknown

Table 3: Chunk label meanings

is approximately 1/1.55 (64%). In other words, what the observed frequencies seem to indicate is that the beginning of queries tends *not* to coincide with chunk boundaries.

However, a quick visual examination of results allows us to formulate a relatively simple explanation for this apparent enigma. When the beginning of a query does coincide with that of a chunk, most often this is a nominal chunk (NP), as can be seen in Table 5. (This table shows the relative frequency of various types of chunks coinciding with the beginning and end of queries, while Table 6 shows the chunk sequences that most frequently match queries). In English, as it turns out, the lexical head of the chunk is usually found at the end, with modifiers preceding it. We can therefore surmise that many queries begin with the lexical head of a noun phrase, optionally preceded by some modifiers. If so, the beginning of such queries will rarely coincide with the beginning of chunks, because modifiers usually appear in front, the most frequent being a determiner.

For example, the query in Figure 6, which is an extract of the *TransSearch* log-files, is "`civil courts`". In all 10 sentences produced by *TransSearch* for this query, this two-word sequence appears within an NP chunk; but in 6 of these, it is preceded by at least one modifier (*the*, 5 times out of 6). Figure 6 shows some examples.

1. $[_{NP}$ Those matters $]$ $[_{VP}$ are best left $]$ $[_{PP}$ to $]$ $[_{NP}$ the **civil courts** $]$.
2. $[_{NP}$ It $]$ $[_{VP}$ mirrors $]$ $[_{NP}$ the legislation $]$ $[_{NP}$ that $]$ $[_{VP}$ was introduced $]$ $[_{PP}$ in $]$ $[_{NP}$ the House $]$ $[_{VP}$ dealing $]$ $[_{PP}$ with $]$ $[_{NP}$ the DNA identification data bank $]$ $[_{PP}$ in $]$ $[_{NP}$ the **civil courts** $]$.
3. $[_{NP}$ If $]$ $[_{ADVP}$ only $]$ $[_{NP}$ they $]$ $[_{VP}$ were being tried $]$ $[_{PP}$ in $]$ $[_{NP}$ American **civil courts** $]$.
4. $[_{PP}$ In $]$ $[_{NP}$ **civil courts** crown liability $]$ $[_{VP}$ exists $]$ $[_{PP}$ by $]$ $[_{NP}$ virtue $]$ $[_{PP}$ of $]$ $[_{NP}$ the Crown Liability and Proceedings Act $]$.

Figure 6: Match examples for the query `civil courts`

Example 4 in this figure is the only case where the match's right boundary does not coincide with the end of the chunk; as it turns out, this is a chunking error (possibly caused by the absence of a comma after *civil courts*).

To verify our hypothesis, according to which many queries would begin with a "truncated noun chunk", we counted the situations where the beginning of the match occurred on the second word of an NP chunk, the first word being either *the*, *a* or *an*. Table 4 shows the boundary matches statistics again, but this time taking into account this possibility.

Boundary Matches :		
to the left	43.821	(69.17%)
to the right	53.399	(84.29%)
both	36.655	(57.86%)

Table 4: Match - chunk boundary coincidence, allowing for a determiner before match.

As we can see, if we account for the possible omission of a determiner at the beginning of queries, left-boundary matches increase by over 25%, over the "random level". Clearly, admitting other types of pre-modifiers (adjectives, common nouns, etc.) would likely lead to similar observations. It also seems reasonable to believe that similar phenomena should be observed with verb groups (omitting modal or auxillary verbs), adjectival groups, etc.

If we suppose that *TransSearch* users mentally formulate their translation problems as sequences of chunks, how can we explain this tendency to truncate the beginning of the initial chunk? Different factors probably come into play. First, this may be simply a way of shortening queries: why include a determiner that adds nothing to the query? In the same vein, it is possible that *TransSearch* users "import" some of the reflexes they have developed while using Internet search engines (eliminate function words). But dropping initial words may also be a way for translators to generalize the initial problem: by eliminating irrelevant pre-modifiers, users increase their chances of finding reusable solutions. Experience possibly plays an important role in these mechanisms: after a while, users probably acquire intuitions about the "optimal" degree of specificity of a query, i.e. the degree of generalization that will strike the right balance between recall and precision. (Needless to say, most users don't apply these strategies consciously, and those that do would probably not formulate them in these terms!)

To summarize our findings:

1. *TransSearch* users mostly submit queries concerning two or more contiguous words; the vast majority of these queries are *verbatim* searches (few make use of generalization operators)
2. these sequences are not just arbitrary sequences of words, as they clearly have a "linguistic status"; many of them correspond to sequences of syntactic chunks, possibly omitting some premodifiers (determiners, adjectives, auxilliaries, etc.) while retaining the lexical head;
3. the exact nature of these chunk sequences is extremely varied, but most queries are made up of three chunks or less, and most of these are NP, VP and PP chunks.

initial chunk	frequency	final chunk	frequency
NP-	0.3711	-NP	0.5152
VP-	0.2789	-PP	0.2008
PP-	0.2204	-VP	0.1499
ADVP-	0.0542	-ADJP	0.0451
ADJP-	0.0384	-ADVP	0.0362
SBAR-	0.0171	-SBAR	0.0302
UNK-	0.0064	-PRT	0.0122
PRT-	0.0048	-UNK	0.0039
CONJP-	0.0015	-INTJ	0.0007
INTJ-	0.0008	-CONJP	0.0001
UCP-	0.0007	-UCP	0.0000

Table 5: Distribution of chunks that most frequently match the beginning and the end of *TransSearch* queries

chunk sequence	relative frequency
NP	0.1988
PP-NP	0.1451
VP	0.0629
VP-NP	0.0585
VP-PP	0.0577
NP-PP	0.0414
NP-VP	0.0381
PP-NP-PP	0.0380
NP-PP-NP	0.0194
ADJP	0.0172
VP-PP-NP	0.0169
VP-ADVP	0.0114
VP-ADJP	0.0113
ADVP-VP	0.0105
VP-PRT	0.0101
ADVP	0.0100
VP-NP-PP	0.0099
NP-VP-NP	0.0099
PP-NP-PP-NP	0.0086
ADJP-PP	0.0086
VP-SBAR	0.0085
NP-VP-SBAR	0.0077
ADVP-PP	0.0064
SBAR-NP-VP	0.0063
NP-VP-ADJP	0.0059
PP-VP	0.0053
NP-VP-PP	0.0053
ADJP-PP-NP	0.0049
ADVP-PP-NP	0.0047
VP-ADJP-PP	0.0045
PP-NP-VP	0.0045
NP-VP-NP-PP	0.0045
others	0.1472

Table 6: Distribution of chunk sequences that most frequently match *TransSearch* queries

Actually, what the second of these conclusions likely indicates is that the notion of syntactic chunks does not match very well with *TransSearch* queries, and that syntactic *constituents* or *dependencies* may provide for a better characterization. Interestingly, this would concord with psycholinguistic evidence about the translation unit, collected via experiments based on *think-aloud protocols* and similar procedures.

A Better *TransSearch*

Can we directly transpose the conclusions of the study reported on in the previous section to translation units? Is there a direct link between the way translators mentally segment and process the source-language material they are faced with, and the way they submit queries to a system such as *TransSearch*? It is obviously tempting to think so, but making that jump may be premature at this point.

One obvious problem with using data from the *TransSearch* log-files is that we have little control over the "experiment". As outlined above, we don't know for certain who the users of the system are, we don't know where their queries come from, and we don't know how they use the results.

In this section, we will discuss how an improved *TransSearch* might help us better answer these questions and others, while at the same time providing the end-users with a tool that better suits their needs. The fundamental idea here is to move the *TransSearch* user-interface right into the translator's "battlefield", i.e. the word-processor.

In-line *TransSearch* Queries

While many translators rely on dedicated software (CAT systems) to do their work, it is probably fair to say that the vast majority of translations are still produced using general purpose word-processing software (*Microsoft Word*, not to name it). What we have in mind is a simple add-on functionality to such a word-processor, that would both make the life of translators easier, and provide the research community with more useful, accurate and complete data.

We propose a functionality, which we call *In-line TransSearch* (ITS), that would allow translators to submit queries to *TransSearch* (or any other similar translation database) directly from within a word-processor. In practice, the user would select some segment of text within the edit pane of the word-processor and submit it as a query to *TransSearch* either by selecting a command from a pop-up menu or via some reserved keystroke sequence.

A separate window would then pop up to display the results of the search. In the simplest version of this idea, this could be a Web browser window. (We will discuss later on how we can possibly improve on this way of displaying matches.)

It is quite obvious how ITS would work for simple queries. While it would certainly be possible to develop similar ways of submitting complex queries such as those of Figure 1, it is unlikely that users would want to learn or use such maneuvers any more than they currently do[5]. A possibly simpler and more versatile solution for those rare users who do use "advanced queries" would be to also provide a pop-up window as an alternative way to submit queries. This *ITS popup* would behave somewhat like the *Find* popup of most word-processors, and allow all the required flexibility to submit arbitrarily complex queries, when the simple "select-and-submit" mechanism is not appropriate.

Enriching Queries with Context

It is not difficult to see how the ITS would make the translator's life easier; he or she would no longer have to leave the word processor to submit queries. But how would it help us researchers? Simple: when the user submits a query by selecting words in the text, the software behind the ITS could send not only the selected words, but also their surrounding context. This context could either be the sentence or paragraph that contains these words, or a fixed number of words or characters around them. This additional information, stored in the log-files alongside the query, would help answer the question "*where do user queries come from?*".

In turn, such contextual information could be used by *TransSearch* to provide more precise results to the queries. For example, context might be used to re-rank *TransSearch* matches according to their similarity with the query's context[6]. Note in passing how this sort of mechanism would lead to a behavior much closer to that of existing translation memory systems: with this kind of re-ranking, a close or exactly matching sentence would likely be displayed first.

Context would also make it possible for *TransSearch* to resolve grammatical ambiguities in queries. The problem is most obvious with queries consisting of categorically ambiguous words like "will" and "lead". Knowing whether the user is looking for a noun or a verb can be critical in proposing useful matches. But it also occurs quite frequently when the inflection-matching operator is used. For instance, a user looking for translations for the noun-phrase "*surplus estimate*" might be tempted to use this operator on the last word *estimate*, so as to also match plural forms. However, because *TransSearch* does not know whether the user is looking for the noun or the verb, it will look for both noun and verb inflections *estimates*, *estimated* and *estimating*, and mix both types of matches.

If the query's surrounding context is available, then we can rely on automatic part-of-speech tagging to determine what the user is really looking for, then restrict expansions to the relevant forms, and re-rank or filter matches based on their parts-of-speech.

Query Translation Spotting

As proposed above, results of an ITS query could simply be presented in a separate popup window, essentially as *TransSearch* does in its current state. However, recuperating translations from this kind of display is rather inconvenient, for two reasons: first, *TransSearch* does not currently have the capacity to locate the exact translation of the query, and therefore it is left to the user to scan the target-language examples to find the equivalent(s) they are looking for; and second, because the display window and the text editor are "unconnected", users must either cut-and-paste the translation they intend to re-use, or directly type them themselves in the word processor.

While word-alignment, or more precisely in this case *translation spotting* (Véronis & Langlais 2000), is a notoriously difficult problem, a number of workable solutions have been proposed that would make it possible to overcome the first obstacle above, if only in a limited number of situations (for example, see Mihalcea & Pedersen 2003 for recent work on word alignment).

Word-level alignments between the query and its target-language translations in matched examples would make it possible to highlight these correspondences in the display. It also opens the door to efficient recuperation mechanisms, similar to those used in spell-checking functionalities: the user could insert one of the proposed translations into his text simply by clicking on it in the display of results.

Word alignments would also make it possible to group together similar translations, thus making the display more compact and easier to read. Each distinct translation of the query would be represented by a "prototypical" example; additional examples would then be available on demand. As an alternative to ranking results based on contextual similarity as proposed above, the user could chose to view the most frequent translation first.

Interestingly, the kind of functionality proposed here could very well affect the way users submit their queries. We have seen earlier how *TransSearch* users tend to shorten their queries, either in an attempt to make them more general or to just save time, keystrokes, etc. Knowing that a longer (and therefore less general) query could be rewarded by a complete translation, ready for re-use, might lead users into re-thinking their strategies.

Here again, such improved user functionalities could benefit the research community as well, by answering the question: "How are *TransSearch* results used?" A log-file that would systematically record which translation was picked as the most appropriate for a query in a given context could be a very useful tool to study human translation strategies. Interestingly, it could also be a very valuable resource for current approaches to machine translation that rest on bilingual phrases extracted from word-aligned corpora (Marcu & Wong 2002; Tillmann & Xia 2003; Och & Ney 2004). A major advantage here is that the phrasal correspondences recorded in the log-file have been explicitly validated by human translators.

Conclusions

We have presented the *TransSearch* system, and have attempted to show how the raw data found in its log-files could be converted into a wealth of information and knowledge for the benefit of researchers in the fields of translation studies

[5] As we have seen in the previous Section, "simple" queries make up close to 95% of all queries submitted to *TransSearch*.

[6] Currently, *TransSearch* displays matches in anti-chronological order, i.e. those entries most recently added to the database are displayed first.

and machine translation. We have also proposed a number of improvements to the *TransSearch* system itself which would both provide the research community with richer data and possibly increase the usability of the application from the point of view of its end-users.

Of course, whatever knowledge we can extract from these log-files certainly cannot be seen as "voluntary contribution" from the community of *TransSearch* users. To what extent translators would behave differently if they knew that they were being "observed" is not clear. But clearly, from the point of view of translation studies, this aspect is a positive characteristic of the data.

Yet it does raise an ethical issue. The *TransSearch* system was not designed with this kind of data collection in mind, and as outlined earlier, the log-files were initially intended mostly for administrative purposes. As a result, the users were never consulted as to whether they agreed to participate in this kind of study. In practice, as administrators of the service, we did have access to confidential information about the users, but it was a point of honour for us never to make use of this information for this study or any other. At the very least, if we were to make the existing data publicly available to the scientific community, we would first have to ensure that the anonymity of the users is preserved.

Our proposal for an in-line *TransSearch* querying facility also raises an additional issue. As we propose to recuperate and store in the log-files not only the queries, but also their context, in the form of the surrounding sentences or paragraphs, issues of confidentiality and intellectual property arise. First, it seems quite obvious that many users will be reluctant to use the system, knowing that bits and pieces of the material they are working on will be traveling freely over the Internet. To ensure confidentiality, it will be necessary to encrypt communications between the user and *TransSearch*, and to install the necessary saveguards around the system's log-files. Furthermore, questions of intellectual property are somewhat complicated when it comes to translation, because translators do not usually own the texts they are working on. Therefore, they are not technically in a position to authorize its dissemination.

These are complex and sensitive issues, and it is not yet obvious how to solve them. Nevertheless, we do believe that ethical solutions can be found to these questions, and that we must actively seek them. A wealth of valuable knowledge is lying dormant inside these log-files and others, waiting to be exploited. We have only just begun to scratch the surface of their enormous potential.

References

Abney, S. 1991. Parsing by Chunks. In Berwick, R., ed., *Principle-Based Parsing: Computation and Psycholinguistics*. Dordrecht, The Netherlands: Kluwer Academic Publishers. 257–278.

Abney, S. 1992. Prosodic Structure, Performance Structure and Phrase Structure. In *Proceedings, Speech and Natural Language Workshop*, 425–428. San Mateo, USA: Morgan Kaufmann Publishers.

Bernardini, S. 2001. Think-aloud Protocols in Translation Research: Achievements, Limits, Future Prospects. *Target* 13(2):241–263.

Bies, A.; Ferguson, M.; Katz, K.; and MacIntyre, R. 1995. Bracketing Guidelines for Treebank II Style Penn Treebank Project. Technical report, University of Pennsylvania.

Isabelle, P.; Dymetman, M.; Foster, G.; Jutras, J.-M.; Macklovitch, E.; Perrault, F.; Ren, X.; and Simard, M. 1993. Translation Analysis and Translation Automation. In *Proceedings of the 5th Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*.

Macklovitch, E.; Simard, M.; and Langlais, P. 2000. TransSearch: A Free Translation Memory on the World Wide Web. In *Proceedings of the Second International Conference on Language Resources & Evaluation (LREC)*.

Marcu, D., and Wong, W. 2002. A Phrase-Based, Joint Probability Model for Statistical Machine Translation. In *Proceedings of the 7th Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Mihalcea, R., and Pedersen, T. 2003. An evaluation exercise for word alignment. In *Proceedings of the HLT/NAACL Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*.

Och, F. J., and Ney, H. 2004. The Alignment Template Approach to Statistical Machine Translation. *Computational Linguistics* 30(4):417–449.

Osborne, M. 2000. Shallow Parsing as Part-of-Speech Tagging. In Cardie, C.; Daelemans, W.; Nédellec, C.; and Sang, E. T. K., eds., *Proceedings of the Fourth Conference on Computational Natural Language Learning*.

Simard, M. 2003. *Mémoires de traduction sous-phrastiques*. Ph.D. Dissertation, Université de Montréal.

Tillmann, C., and Xia, F. 2003. A Phrase-Based Unigram Model for Statistical Machine Translation. In *Human Language Technologies (HLT) Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 106–108.

Véronis, J., and Langlais, P. 2000. Evaluation of Parallel Text Alignment Systems – The ARCADE Project. In Véronis, J., ed., *Parallel Text Processing*, Text, Speech and Language Technology. Dordrecht, The Netherlands: Kluwer Academic Publishers.