Translation Analysis and Translation Automation

ABSTRACT

We argue that the concept of *translation analysis* provides a suitable foundation for a new generation of translation support tools. We show that pre-existing translations can be analyzed into a *structured translation memory* and describe our TransSearch bilingual concordancing system, which allows translators to harness such a memory. We claim that translation analyzers can help detect *translation errors* in draft translations and we present the results of an experiment on the detection of deceptive cognates conducted as part of our TransCheck project. Finally, we claim that translation analysis can facilitate the *speech-to-text* transcription of dictated translations and introduce our new TransTalk project.

1. Introduction

In 1951, Y. Bar-Hillel, the first full-time researcher in MT, wrote the following:

> "For those targets in which high accuracy is a conditio sine qua non, pure MT has to be given up in favor of mixed MT, i.e., a translation process in which a human brain intervenes. There the question arises: Which parts of the process should be given to a human partner?" (Bar-Hillel [1], p. 230)

Forty-two years and three 'generations' of systems later, pure MT is not more widely applicable than it was then(1). More discouraging still, neither is mixed MT. While precise figures are not readily available, it appears safe to assume that the current share of anything that could be called MT, pure or mixed, is well below 1% of the total translation market. One is forced to conclude that the MT community has so far failed to come up with realistic and practical answers to Bar-Hillel's question about the optimal division of labor between man and machine.

Bar-Hillel himself ventured to suggest a man-machine tandem in which the human partner would intervene either before or after the mechanical process, "but preferably not somewhere in the midst of it." That is, the machine would take care of the core part of the translation process. Ever since, 'human-aided MT' has remained the predominant paradigm within the MT community. Machines have persistently been asked to do something they fail to do well: namely, translate. And humans have persistently been asked to do things they would rather not do, like inserting strange codes into source texts, answering odd questions about phrase bracketings or rearranging bizarre jumbles of target language words. In any case, the market response to this kind of man/machine modus vivendi has consistently been less than enthusiastic.

---

(1) Though it has been shown that MT can be remarkably successful in the rather marginal case of some extremely narrow sublanguages like weather bulletins (Isabelle [12]).

It has become obvious that, generally speaking, **machines still cannot successfully assume control over the core part of the translation process**. As far back as 1980, Martin Kay [18] forcefully argued for a reversal of roles in which the machine is sent back to its 'proper place', that of an assistant to the human translator:

> "I want to advocate a view of the problem in which machines are gradually, almost imperceptibly, allowed to take over certain functions in the overall translation process. First they will take over functions not essentially related to translation. Then, little by little, they will approach translation itself. The keynote will be modesty. At each stage, we will do only what we know we can do reliably. Little steps for little feet!" (p. 11)

It is precisely this kind of down-to-earth approach that the Center for Information Technology Innovation (CITI chose to pursue when it launched its translator's workstation project back in 1987. In its current incarnation, the CITI's workstation provides the translator with a windowing environment where he/she has simultaneous access to a number of tools such as split screen word processing, spelling correction, terminology and dictionary lookup, file comparison, word counting, full-text retrieval, etc. (Macklovitch [17]). Admittedly, this has more to do with office automation for translators than with translation automation per se. But following Kay's proposed scenario, we can now take advantage of this computer base and progressively enrich it with translation-oriented tools. From this perspective, the central issue can be formulated as follows: **beyond office automation, but short of machine translation, what else can be done to support translators?**

In the remainder of this paper, we argue that the concept of *translation analysis* constitutes a suitable foundation for the development of a new generation of translation support tools. Section 2 is a general discussion of the notion of translation analysis. Sections 3, 4 and 5 describe our work on three applications: the translation memory, the translation checker and the translator's dictation machine.

2. Translation Analysis

In recent literature (e.g. Isabelle, Dymetman & Macklovitch [14]), translation is often conceptualized as a relation $tr_{L1,L2}(S, T)$ whose extension is a set of pairs $<S,T>$ such that S is a text of language L_1 and T is a text of language L_2. Since the number of texts in each language is infinite, $tr_{L1,L2}$ has to be be defined recursively, with the consequence that the relation will have a compositional character: down to the level of some finite set of primitive elements, S and T will be decomposed respectively into sets of elements $\{s_1, s_2, ..., s_n\}$ and $\{t_1, t_2, ..., t_n\}$, in such a way that for any i, $tr_{L1,L2}(s_i, t_i)$ is also satisfied.

An ordinary MT system embodies some (possibly partial) specification of a translation relation $tr_{L1,L2}$, together with a procedure which, given any value of S, will return one or several values T such that $<S,T>$ belongs to $tr_{L1,L2}$. A reversible MT system (see for example Dymetman [8], Van Noord [21]) can in addition compute, for any value of T, the values S for which $<S,T>$ belongs to $tr_{L1,L2}$.

While MT systems deal with the problem of producing translations, we can also, as noted by Debili [7], view translations from a recognition perspective. We will call a *translation acceptor* any procedure which, given some particular pair $<S,T>$, can decide whether or not $tr_{L1,L2}(S,T)$ holds. Furthermore, we will call a *translation analyzer* any recursive procedure $ta(<S,T>, TAT)$ that assigns to those pairs $<S,T>$ that satisfy $tr_{L1,L2}(S,T)$ a translation analysis tree TAT. A TAT makes explicit the compositional makeup of the translation relation. For example, given some suitable definition of the English-French translation relation, a translation analyzer could produce a TAT such as the one shown in Figure 1.

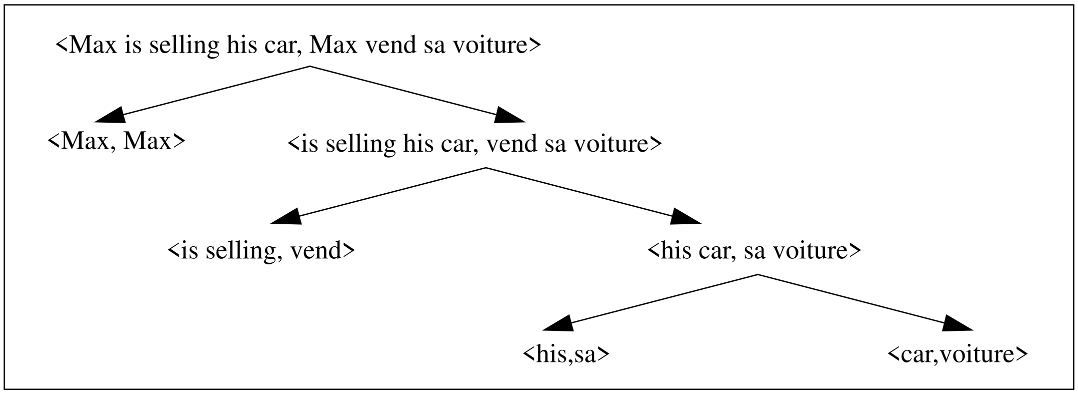

Figure 1: A translation analysis tree (TAT)

Isabelle [13] uses the term *bi-text* to designate structures which, like TAT's, are meant to decompose translations into their constituent correspondences. TAT's are structural descriptors for translation analyses in the same way that parse trees are structural descriptors for grammatical analyses.

In principle, translation analysis and MT are very similar problems: the computation is based on the same abstract relation $tr_{L1,L2}$. The difference is only in the computing modes. Does this mean that in practice translation analyzers and MT systems are subject to exactly the same limitations? In particular, does this mean that useful translation analyzers are feasible if and only if useful MT systems are feasible?

Clearly not. Of course, in those rare cases where high-quality MT is feasible, it should be possible to build a translation analyzer for the output of the MT system. But more importantly, in cases where MT is not possible, we claim that it is still possible to develop analyzers for the translations produced by human translators, and that there will be many uses for these devices. This difference stems from the practical requirements that different tasks (MT versus translation analysis) impose on the level of precision in the formal characterization of $tr_{L1,L2}(S,T)$.

Consider for example the model that underlies the sentence alignment method proposed by Brown & al. [3]. Conceptually, this model generates sequences of pairs of $<S,T>$ in such a way that a) S is a sequence $<s_1, s_2, ... s_n>$ in which each s_i is itself a sequence of 0, 1 or 2 'sentences' and T is a similar sequence $<t_1, t_2, ... t_n>$; b) a 'sentence' is any string of tokens terminated by a punctuation token; c) a token is any string of characters appearing between delimiter characters; d) the length $l(s_i)$ of each s_i (in terms of the number of tokens it contains) is correlated with the length $l(t_i)$ of the corresponding t_i according to a probability distribution $Pr(l(s)|l(t))$; and e) this probability distribution can be estimated from frequencies observed in corpora of translations, like the Hansard corpus of English/French texts.

This model does capture one specific aspect of the translation relation between two languages, namely length correlations between sentences that are mutual translations. In this sense it constitutes a translation model, albeit an extremely weak one.

If we were to apply a model of this kind to the task of translating English texts into French, an English sentence *e* would be translated more or less as random sequence of characters *f*, whose only notable property is to have a length *l(f)* that is typical for a translation of an English sentence of length *l(e)*. Such an ‘MT system’ would appear perfectly useless in practice.

On the other hand, if like Brown & al. we apply their model to the task of translation analysis, we get a system capable of analyzing pre-existing translations into representations in which their compositional makeup is made explicit down to the level of sentences. The result is a TAT of the form shown in Figure 2, in which texts *S* and *T* are decomposed into *n* successive pairs of blocks s_i and t_i of sentences.

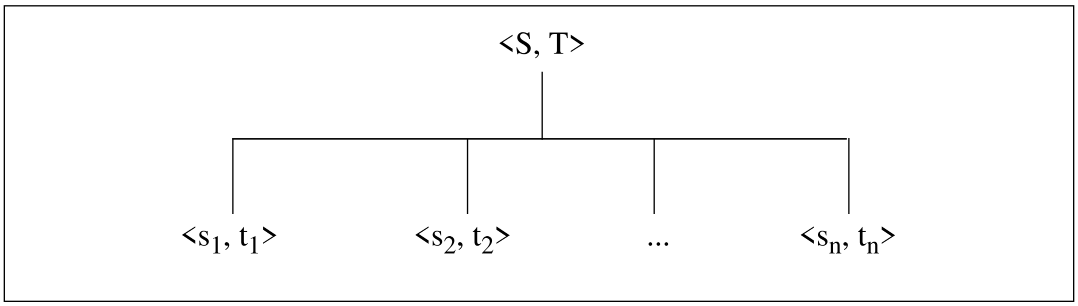

Figure 2: Sentence alignment as a simple case of TAT

Admittedly, the analysis is very crude: no correspondences are established below the sentence level. Still, as we will see shortly, these ‘low-resolution’ bi-texts provide an adequate basis for some very useful translation support tools.

Of course, richer analyses would open up even more possibilities in this respect. And in fact, is not too hard to imagine families of somewhat stronger translation models which, while still insufficient for successful MT, could be used to successfully uncover more structure in pre-existing translations (e.g. phrase or word correspondences).

With respect to their general architecture, models used for translation analysis can be very close to those used for MT. The most obvious possibility is perhaps the tripartite model illustrated in Figure 3. Just as in the well-known transfer model of MT, there are two language-specific components

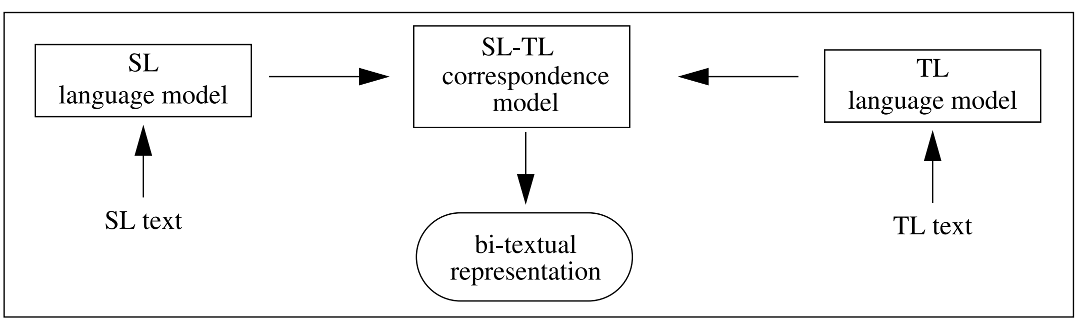

Figure 3: A tripartite model for translation analysis

(the language models), and one pair-specific, ‘contrastive’ component (the correspondence model). Both monolingual components operate in the analysis mode and the language-specific representations that they produce are fed into the correspondence model, which connects them into a single bi-textual representation in which translation correspondences are made explicit. This model remains a natural one regardless of whether its components are implemented by means of rule-based or corpus-based techniques. In fact, even the simple length-based alignment method mentioned above is best conceptualized as an instance of it.

In the development of general-purpose translation analyzers, there is some evidence to suggest that probabilistic models will turn out to be extremely useful. While rule-based methods work well for the development of 'deep' models in narrow domains, probabilistic methods appear especially well-suited to the development of shallow models potentially capable of providing reasonably good partial analyses of non-restricted translations.

In any case, our basic claim here is only that translation analysis, even based on weak translation models, provides the right foundation for a new generation of translation support tools. We now turn to an examination of some of these tools.

3. Translation Memory

3.1 Existing Translations as a Resource

The trend towards corpus-based approaches in MT stems in part from a realization that the existing body of translations is an immensely rich resource whose potential has so far been neglected. In fact, it is clear that **existing translations contain more solutions to more translation problems than any other available resource**.

But translators will only be able to tap the riches buried in their past production once they are provided with tools capable of managing it as translation data rather than as word-processing data. This is precisely what a translation analyzer sets out to do: upgrade word-processing data into bi-textual structures that make translation correspondences explicit.

Once pre-existing translations are organized in that way, corresponding source and target language segment are systematically linked together. In particular, any segment containing an instance of some translation problem is linked with a segment containing a ready-made solution for that problem. If we provide translators with the means to create, store and search such bi-textual structures, their past production becomes a highly effective translation memory.

3.2 TransBase

In order to render accessible the results of translation analyses of large quantities of text, we have devised a simple model for a structured translation memory, which we call TransBase. It shares the basic characteristics of full-text retrieval systems: it can manage arbitrary amounts of text, it can be enlarged incrementally and it allows rapid access to the textual contents of the database. Its essential difference with these systems is its ability to also store bi-textual representations.

A TransBase database is constructed using a translation analyzer similar to the one depicted in Figure 2. Each document in a pair of mutual translations is submitted to a language-specific analysis which breaks it down into its structural elements (paragraphs, sentences, etc.) and determines its lexical content. This information is stored in two distinct language-specific components of the database, and indexed so as to allow rapid access to any part of the text. A “correspondence analyzer” based on the techniques described in Simard, Foster & Isabelle [20] then uses these language-specific analyses to construct a sentence-level “translation map”, which is also stored into the database. The structure and construction scheme of the database are illustrated in Figure 4.

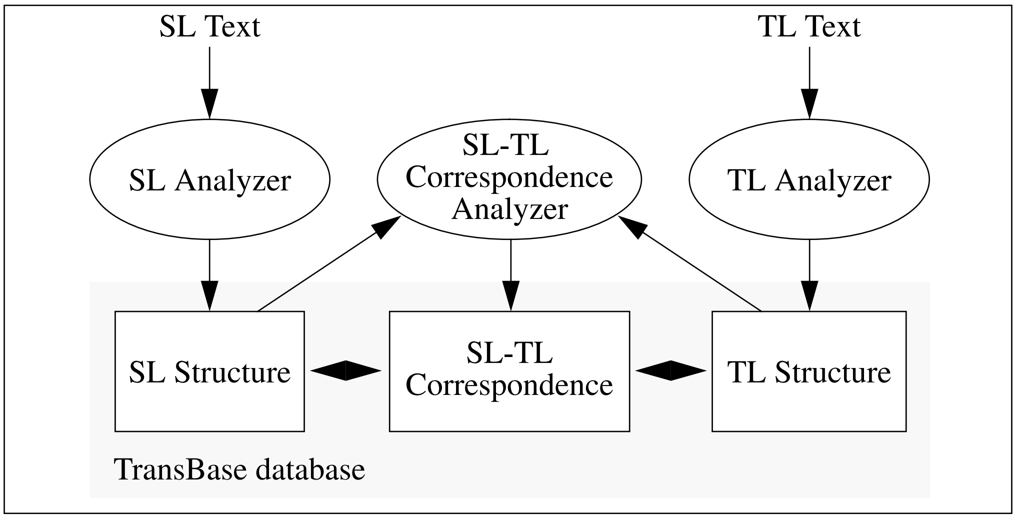

Figure 4: General organization of a TransBase database

The texts of the source and target languages are handled symmetrically in the database. However, since the directionality of the translation may be important to the user, TransBase can record which language is the source.

3.3 TransSearch

There are many possible ways to exploit such a translation memory. The first one that comes to mind, and probably the most universally useful, is to provide translators with tools to search a TransBase database on the basis of its textual content. It has already been suggested that a tool capable of producing *bilingual concordances* would be useful to bilingual lexicographers (see for example Church & Gale [6]). It is rather obvious that bilingual concordancing would also be useful to translators. For example, upon encountering some occurrence of an expression like *to be out to lunch* or *to add insult to injury* in his English source text, a translator might be hesitant as to an appropriate French equivalent. He/she might also find that conventional bilingual dictionaries do not provide satisfactory answers. Bilingual concordancing would enable him/her to retrieve examples of these expressions together with their translations in a database of the TransBase kind. This could be useful not only for idiomatic expressions, but also for specialized terminology or domain-specific formulae (*To whom it may concern...*, *Attendu que...*). See Macklovitch [16] for a more detailed discussion of this issue.

TransSearch is just such a tool: it allows one to extract occurrences of specific ‘expressions’ from the database, and to visualize them within their bilingual context. Because the software is primarily aimed at translators, who are likely to use it as just another reference source, it is designed to be used interactively and to provide answers in real-time. This is just what the inclusion of word-form indexes within the TransBase model is meant to allow for.

Because most translators are not computer experts, much attention has been devoted to the user-friendliness of the TransSearch interface. Using an intuitive, graphically-oriented query language, it is easy for a user to submit complex queries to the database. Every such query defines a logical expression on sequences of word-forms: when the query is submitted, the system produces all the couples that satisfy this expression in the alignment component of the database. In addition, the inclusion of dictionaries and morphological descriptions of both French and English allows TransSearch to automatically match any inflectional variant of query items.

The result of a query is normally presented in a two-column format, where mutual translations appear side-by-side. The user can either examine one match at a time within the document from which it was drawn, or collect all matches with a small portion of their immediate context, the way concordances are usually presented.

Figure 5 gives an idea of a typical session with TransSearch. In this example, the user has queried the system for occurrences of the English expression *take X to court* which are <u>not</u> translated in French as *poursuivre X* or *intenter un (or des) procès à X*, and the database searched consists of the 1986 Canadian Hansard translations. The translators to whom we have shown the system invariably concluded that bilingual concordancing would be very useful to them.

4. Translation checking

4.1 Translation Analysis and Error Detection

In recent years, we have witnessed the appearance on the market of text critiquing tools meant to help writers improve their texts by spotting potential problems in spelling, grammar and even style. While these tools can in principle help translators correct writing errors in the target language text, there is no way they can help them correct *translation errors* in the strict sense of the term, that is, incorrect correspondences between the source and target texts. For example, they cannot help with cases of *mistranslation* in which both texts are individually correct and meaningful, but do not happen to mean the same thing. Such errors can only be detected by a device that simultaneously examines the source and target texts. In other words, a device that comprises a translation analyzer.

Given a translation analyzer capable of reconstructing some subset *Cset* of the correspondences that are observable in the result of some translation operation, and given some set of constraints *C* on admissible correspondences, a translation checker is a device that helps the translator ensure that *Cset* indeed satisfies *C*. This requires a translation analyzer based on a 'robust' translation model, a model capable of observing actual correspondences that may be deviant with respect to the norm defined by *C*.

The general problem of translation quality is a notoriously complex and vexing issue. It is certainly not our intention to propose any global metric or method for evaluating translations. Our aim is more modest. We only want to identify some particularly simple properties that most translators will want their translations to possess and devise some tools that will help them verify these properties.

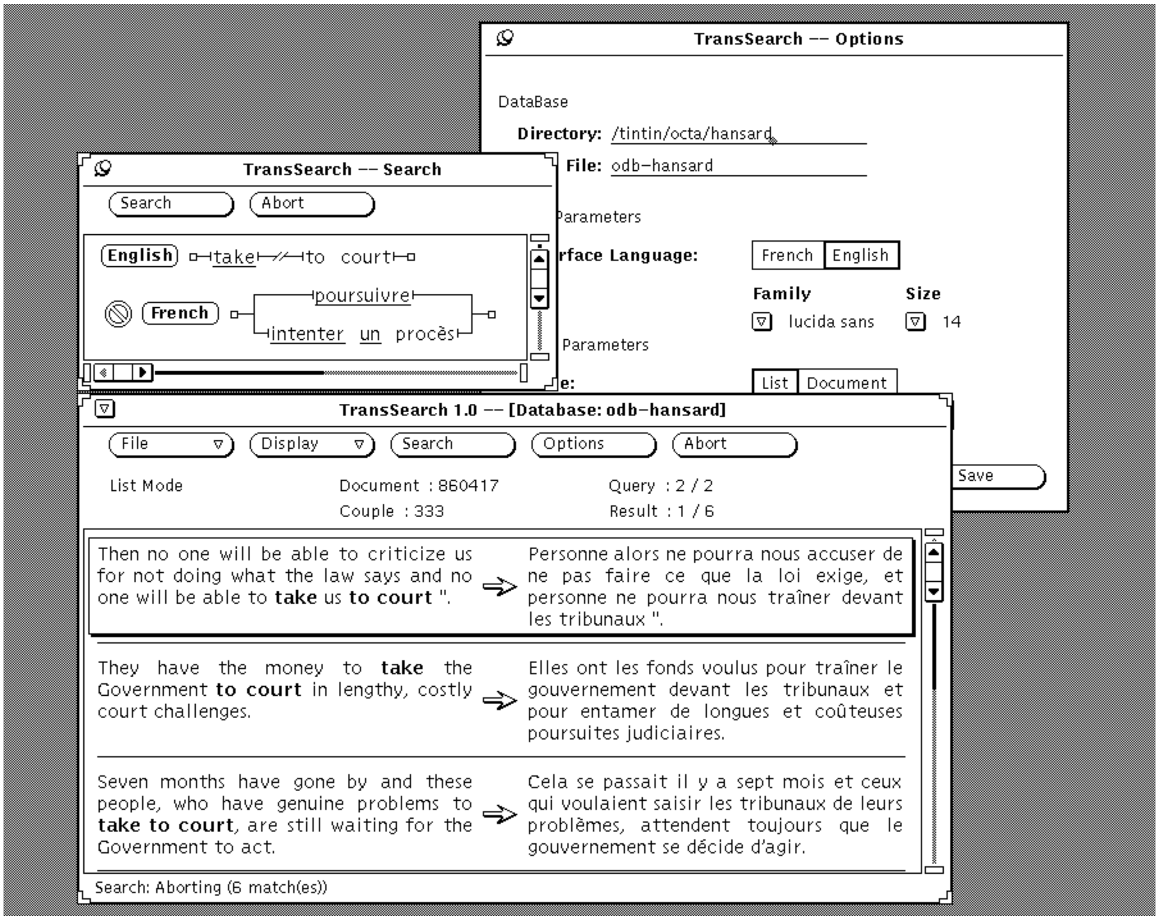

Figure 5: A session with TransSearch

One rather obvious candidate is the property of *exhaustivity*. Normally, all parts of the source text should have a corresponding element in the target text. But translators sometimes make *omission errors*, forgetting for example to translate a sentence, a paragraph, or even a complete page. In such cases an adequate translation analyzer should realize that a source language segment is being mapped onto an empty target language segment. The checking device could then warn the translator, pointing out a possible problem in his draft translation.

Another candidate property is *terminological coherence*. In technical translations, one and the same target language term should be used to translate all occurrences of any particular source language term. A process of translation analysis capable of bringing out term correspondences between a draft translation and its source would presumably make it possible to help translators enforce terminological coherence.

A third constraint that translations are expected to obey is the *absence of source language interference*. Some cases of interference result in constructs that are ill-formed with respect to the target language. Their detection is possible without any need to look at the source text. For example, if the English word *address* is translated as *addresse* (with two d's) in French, an ordinary French spell checker should be able to flag the problem. But there are also cases in which interference results not in ill-formedness but rather in mistranslation. *Deceptive cognates*, for example, tend to generate this kind of interference.

Word w_e of language L_e and word w_f of language L_f are *cognates* when their forms are similar due to shared etymology. For example, the English word *'government'* and the French word *'gouvernement'* are cognates. Most often, these words are not only cross-linguistic homonyms but they are synonyms as well. However, in some cases the synonymy does not hold. For example, the following pairs of English/French cognates have completely disjoint meanings: *<actual, actuel>*, *<library, librairie>*, *<physician, physicien>*. Such cognates are said to be 'deceptive' because of the misleading semantic expectation induced by their morphological similarity. The sentence *Max se rendit à la librairie* is perfectly well-formed in French, but used as a translation for *Max went to the library*, it would constitute a blatant case of mistranslation. To the extent that a translation analyzer is capable of observing in a draft translation an actual correspondence between cognates known to be deceptive, this correspondence can be flagged as a possible error for the translator to verify.

There are probably several other types of translation errors that translation analysis could help detect. Research in this area is just starting. In order to get a better idea of the practical potential of this approach we conducted an experiment on the detection of deceptive cognates in actual translations.

4.2 An Experiment on the Detection of Deceptive Cognates

Deceptive cognates (DC's) can be subclassified as to whether they are *complete* or *partial*. Complete DC's, like the examples given above, have the property that their meanings are completely disjoint, and as a consequence can never be used as mutual translations. Partial DC's, on the other hand, have partially overlapping meanings, and are mutually translatable in some subset of their possible uses. For example, the French verb *examiner* is sometimes equivalent ('≡') and sometimes non-equivalent ('≠') to the English verb *to examine*:

The doctor examined his patient ≡ Le médecin examina son patient
The professor examined his students ≠ Le professeur examina ses étudiants.

Concentrating for the moment on the easier problem of complete DC's, we conducted an experiment aimed at: 1) assessing the amplitude of the problem in actual translations; and 2) evaluating the effectiveness of some straightforward detection methods.

We assembled a simple translation analyzer, TA1, that instantiates the model of Figure 1 as follows: language models for French and English are reduced to processes of 'tokenization' and morphological analysis (based on a dictionary and a set of inflection rules). The output of these language models is a simple morphological representation of the input text: each token is represented as the set of citation forms of the lexical entries of which it is potentially an instance. The correspondence model used in TA1 is simply the sentence alignment program of Simard, Foster & Isabelle [20]. Its output representation is a sequence $<<e_1,f_1>, <e_2,f_2>, ...<e_n,f_n>>$ where each e_i is a sequence of zero, one or two morphologically represented sentences of the English text, each f_j is a sequence of zero, one or two morphologically represented sentences of the French text, and each $<e_i,f_i>$ is a translation correspondence.

We extracted from van Roey & al. [22] a list of 145 word pairs which were classified as DC's of the 'complete' variety: *<accomodate, accomoder>*, *<actually, actuellement>*, etc.(2) We then implemented a straightforward checker that would search the output of TA1 and for each word pair $<w_e, w_f>$ would return the set of sentence pairs $<e_i, f_i>$ such that $w_e \in e_i$ (i.e. e_i contains the word w_e) and $w_f \in f_i$. Obviously, this condition can be met by pairs of sentences in which w_e and w_f appear without being used as mutual translations.

We then tested this rather simplistic device on one year of Hansard translations. Hand-checking the results, we found out that many genuine cases of translation errors were retrieved, as in the following:

> The peace movement in Canada is composed of **physicians**, members of the church, [...]
> -> Le mouvement canadien pour la paix compte dans ses rangs des **physiciens**, des ecclésiatiques, [...] (Hansard, 1987/09/29)

> There are parts of this bill which concern **librarians** and the artistic community.
> -> Quelque part dans ce projet de loi, il est question des **libraires** et des artistes.
> (Hansard, 1987/11/30)

But as Table 1 shows, the results were also very noisy.

	No. of cases	**Percent**
Hits (real errors)	57	7.4
Noise	718	92.6
Total	775	100

Table 1: Results of DC retrieval in TA1's output

The noise was generated by three different sources. First, there are cases where the 'deceptivity' of $<w_e, w_f>$ is relative to their part of speech (POS). For example the French noun *local* and the English noun *local* are complete DC's but their homograph adjectives are not. Since POS information was not taken into account, irrelevant cases were retrieved. Second, some of the noise was engendered by untranslated quotations. For example, *agenda* (English) and *agenda* (French) are complete DC's. Since the forms are perfectly identical, our checker was unable to distinguish among the two, and would consequently retrieve cases where *agenda* appears on both sides simply as a consequence of the fact that one of the texts contains it in the form of an untranslated quotation from the other language. Third, there were cases where w_e and w_f did appear in sentences that were mutual translations, but in such a way that these words were not themselves used as mutual translations. Our correspondence model (that is, sentence alignment) was simply too

(2) We do not yet know what proportion of the actual problem is covered by these 145 pairs, but we strongly suspect it is only the tip of the iceberg.

coarse to filter out these cases.(3) The breakdown between these noise sources was as shown in Table 2.

Noise category	No. of cases	Percent
Wrong POS	703	97.9
Quotation	6	.8
Not mutual trans.	9	1.3
Total	718	100

Table 2: Noise categorization for DC retrieval in TA1's output

Given these figures, POS tagging was obviously called for. The translation analyzer was therefore replaced with a new one, TA2, that differed from TA1 only in that its two language models were augmented with the POS tagger of Foster [10]. The search process was modified so as to take into account POS information associated with our 145 pairs of DC's. This scheme produced much better results, as shown in Tables 3 and 4.

	No. of cases	Percent
Hits (real errors)	56	76.7
Noise	17	23.3
Total	73	100

Table 3: Results of DC retrieval in TA2's output

Noise category	No. of cases	Percent
Wrong POS	6	35.3
Quotation	2	11.8
Not mutual trans.	9	52.9
Total	17	100

Table 4: Noise categorization for DC retrieval in TA2's output

POS tagging dramatically reduced the noise, with no more than a marginal effect on the recall (one case is lost). This spectacular effect is in a large measure attributable to the resolution of problems associated with a small number of frequent words (like the case of *local* mentioned

(3) Note however that none of the noise could be attributed to incorrect sentence alignments, which our algorithm gets right about 98% of the time.

above). Part of the remaining noise is due to tagging errors, but the largest proportion is now attributable to the coarseness of our correspondence model.

Better models would no doubt improve DC detection. However, the performance level of the computationally cheap method tested here may well prove sufficient for real-life applications.

5. Translation Dictation: TransTalk

A recurring theme of this paper has been that weak models of translation, if used realistically, can provide useful tools for the human translator, without imposing artificial constraints on his activity. One invaluable addition to the translator's workstation would be an automatic dictation module: many professional translators prefer to dictate their translations rather than doing the typing themselves (Gurstein & Monette [11]).

At the present time, speech-recognition technology is severely limited when confronted with large-vocabularies, and is therefore inapplicable to the task of most translators. An intriguing possibility, however, is that of *teaming* the speech-recognition module with a (weak) translation model. The MT model would then be used to make probabilistic predictions of the possible target language verbalizations freely produced by the translator, so as to dynamically reduce the "effective probable vocabulary" considered by the speech-recognition module on each dictation unit (sentence or paragraph) to such an extent that complete recognition of these units can be attempted.

For example, it is clear that the probabilistic composition of the vocabulary considered by a speech recognizer attempting to decode the spoken French sentence: *Ces impôts cachés doivent être acquittés par les pauvres aussi bien que par les riches* should be markedly different depending on whether its English source *The poor as well as the rich have to pay these extra hidden taxes* is available or not. A French translation of this English sentence is for instance much more likely to contain the word *impôts* than is a French sentence taken at random. It seems reasonable to hope that a weak translation model could make this composition available to the speech recognizer.

This idea was independently advanced by Dymetman, Foster & Isabelle [9] and by Brown & al. [4]. We have launched a joint project with the speech-recognition group at CRIM (Centre de Recherche Informatique de Montréal), the *TransTalk* project, aimed at proving the feasibility of the approach, using English as the source language and French as the dictation language. Initially we intend to restrict dictation to an isolated-word mode, then to progress to a connected-speech mode. The TransSearch and TransCheck projects discussed above involved the development of translation analyzers comprising French and English language models and a French-English correspondence model (sentence alignment) that were trained on the Hansard corpus. The Hansard domain is thus a natural choice for the TransTalk project, since existing modules will then provide fundamental resources for TransTalk. Actually, one can view TransTalk as incorporating a translation analyzer much like those described above, except that it has the capability of dealing with target language that is a spoken rather than written.

TransTalk is based on a probabilistic model p of translation dictation relating an English written textual unit e, its French written translation f (to simplify matters, we assume here that textual units are sentences), and the acoustic counterpart *s* of *f*. Both *e* and *s* are known to the system, and TransTalk's job is to provide an estimate $\hat{f}$ of the actual *f* intended by the translator:

$$
\begin{tikzcd}
e \arrow[r, "\text{Trans}"] \arrow[dr] & f \arrow[r, "\text{Talk}"] & s \arrow[dl] \\
& \boxed{\text{TransTalk}} \arrow[d] & \\
& \hat{f} &
\end{tikzcd}
$$

One is thus led to define $\hat{f}$ as:

$$\hat{f} = \mathrm{argmax}_f \; p(f \mid e, s)$$

that is, $\hat{f}$ is the most probable French sentence according to the model *p*, given both the source English sentence and the acoustic realisation of the French sentence.

By Bayes's formula, this equation can be rewritten as:

$$\begin{aligned} \hat{f} &= \mathrm{argmax}_f \; p(s \mid e, f) \; p(f \mid e) \\ &= \mathrm{argmax}_f \; p(s \mid f) \; p(f \mid e) \end{aligned}$$

where the last equality is a consequence of the mild assumption that once *f* is known, further knowledge of *e* cannot add anything to the determination of *s*.

This equation is strongly reminiscent of the "fundamental formula" of statistical speech-recognition (Bahl & al [2]):

$$\hat{f} = \mathrm{argmax}_f \; p(s \mid f) \; p(f)$$

where the distributions *p(s | f)* and *p(f)* are known as the acoustic model and language model respectively. In the situation considered here, the pure language model *p(f)* has been replaced by a "conditional language model" *p(f | e)*, where knowledge of *e* "sharpens" the statistical structure of the language model, in particular by making it "concentrate" its attention on a limited lexical subset of the whole language. A quantitative measure of this "sharpening" can be given in terms of *perplexity*, an information-theoretic quantity which measures the average uncertainty a given language model entertains about the next word to appear in a natural text, having seen the preceding words: the less the perplexity, the more predictive the model (Jelinek [15]). Brown et al. [4] report the results of an experiment with the Hansards, using one of their simpler translation models (from French to English, in their case), which show the per-word perplexity of their pure (English) language model to average 63.3, while the perplexity of their conditional language model drops to an average of 17.2. These results are highly encouraging for the dictation task, for they mean that the acoustic module should be able to discriminate, given one English spoken word, in average between 17.2 equiprobable candidates proposed by the conditional language model, as opposed to 63.3 equiprobable candidates proposed by the pure language model.

Several approaches are possible to the modelling of $p(f \mid e)$. *A first approach*, proposed by the IBM team, is to use Bayes' formula and to write, by analogy to the standard formulation of the speech-recognition problem:

$$p(f \mid e) \propto p(e \mid f)\ p(f)$$

where (in their terminology) $p(e|f)$ is the "translation model", which plays a role similar to the acoustic model in speech recognition. One is thus led to a symmetrical formula for the whole translation dictation model where $p(f)$ is the language model, $p(s|f)$ the acoustic model, and $p(e|f)$ the translation model. This method has two advantages: (1) it relies on a unique language model for French, and (2) the work at IBM on statistical MT seems to indicate that even rough approximations to $p(e|f)$, *when teamed with a good language model for French*, result in acceptable approximations to the conditional language model $p(f|e)$. It is as if there were a "division of work" between $p(f)$, responsible for the well-formedness of French sentences, and $p(e|f)$, responsible for pairing between English and French sentences (hence the somewhat misleading terminology "translation model") without much regard for either the internal structure of French or the internal structure of English[4](see [9] for details). The method has, however, one important shortcoming in terms of processing: it requires an extensive search among the sentences f in order to maximize $p(e|f)\ p(f)$ (not counting the $p(s|f)$ factor, which only makes matters worse). This is known to present serious practical difficulties in terms of non-optimal search results as well as in terms of processing time, this last factor being obviously of central importance in a dictation application.

A second approach to the modelling of $p(f \mid e)$ is to consider a priori a certain parametrized family $p_\lambda(f)$ of language models for French, to describe a mapping $e \to \lambda(e)$, and then to define the conditional language model through:

$$p(f \mid e) = p_{\lambda(e)}(f)$$

Although it presents the inconvenience of dispensing with a unique reference language model for French, this approach can be efficiently implemented if the family $p_\lambda(f)$ is well-chosen. One possibility that we are currently investigating is to adapt a language model proposed in [10]. This model is a kind of "tri-POS" hidden Markov model, depending on two families of parameters. The first family $a_{i,j,k}$ gives the probability of generating a word having part-of-speech POS_k, given that words with parts-of-speech POS_i and POS_j have been previously generated. The second family $b_{i,w}$ gives the probability that a given part-of-speech POS_i is associated with word *w*. That is, conceptually at least, the model first generates part-of-speech strings, using a context window of the two previously generated parts-of-speech, then "decorates" each part-of-speech with a given word form, depending only on this part-of-speech. The $a_{i,j,k}$ parameters represent an approximation to the "grammatical" structure of French, while the $b_{i,w}$ parameters represent an approximation to its "lexical" structure.

We propose to experiment with a scheme where these parameters vary dynamically depending on the observed source sentence *e*. One interesting possibility is to keep the "grammatical parame-

(4) In fact, it is easy to see that, for the purpose of English-to-French translation, it is equivalent to use $p(e|f)$ or $p(e|f)\,/\,p(e)$ as the "translation model". This last quantity is the exponential of the mutual information between e and f, a quantity symmetrical in *e* and *f*, and bearing no memory of the internal statistical structure of either *e* or *f*, but only of their statistical relationship.

ters” fixed at their global French language values (neglecting the influence of the grammatical make-up of the English sentence on its translation), while modifying the “lexical” parameters depending on the lexical make-up of the English sentence. The first family of parameters can be estimated reliably on a sufficiently large French corpus, while the second family of parameters, depending on *e*, can be estimated if certain simplifying assumptions akin to the Translation Model 1 of Brown & al. [5] are made. Basically, each $b_{i,\,w_f}(e)$ is considered to be the average of the contributions $p\,(w_f\,|\,w_e\,,POS_i)$ made by each English word w_e in *e* to the probability of realising part-of-speech POS_i as the French word w_f. In order to estimate the parameters $p\,(w_f\,|\,w_e\,,POS_i)$, it is necessary to have a pre-aligned training corpus of English-French bitexts (see section 3). It is then possible to start with initial guesses for the $p\,(w_f\,|\,w_e\,,POS_i)$ parameters, and use standard reestimation techniques (see [5]) on this training corpus to maximize the predictive power of these parameters, while holding grammatical parameters fixed.

The main advantage of this approach is that, for each source sentence *e*, the conditional language model in effect reduces to a simple Hidden Markov Model $p_{\lambda(e)}(f)$, and the translation dictation problem then takes the form familiar in speech-recognition:

$$\hat{f} = \operatorname{argmax}_f \; p_{\lambda(e)}(f)\; p\,(s\,|\,f)$$

for which powerful search techniques are available (Bahl & al [2]).

6. Conclusions

A new generation of translation support tools is just around the corner. Thanks to the development of translation analysis techniques, translator's workstations will soon be able to offer much more to their users than mere office automation functions. Translators will soon be in a position to tap the vast potential lying dormant in their past production. They will soon be able to receive assistance in checking their translations for errors. And speech input stands a good chance of becoming a reality for them long before it does for monolinguals.

We would not be surprised to see the list of applications based on the concept of translation analysis expand rapidly. We wish classical MT well, but the real action is likely to be with translator's aids for quite a few years to come!

REFERENCES

[1] Bar-Hillel Y., *The State of Machine Translation in 1951*, in **American Documentation**, vol. 2, 1951, pp. 229-237.

[2] Bahl L., Jelinek F., Mercer R. *A maximum likelihood approach to continuous speech recognition*, **IEEE Transactions on Pattern Analysis and Machine Intelligence**, PAMI-5(2):179--191, March 1983.

[3] Brown P., Lai J., Mercer R., *Aligning Sentences in Parallel Corpora*, **Proceedings of the 29th Meeting of the ACL**, 1991.

[4] Brown P., Chen S., Della Pietra S., Della Pietra V., Kehler S., Mercer R. **Automatic speech**

recognition in machine aided translation, 1992. (to appear).

[5] Brown P., Della Pietra S, Della Pietra V, Mercer R. **The Mathematics of Machine Translation: Parameter Estimation**, (to appear).

[6] Church K., Gale W., *Concordances for Parallel Texts*, in **Proceedings of the 7th Annual Conference the UW Centre for the NOED and Text Research**, Oxford, 1991.

[7] Debili F., Sammouda E., *Appariement des phrases de textes bilingues Français-Anglais et Français-Arabes*, **Proceedings of COLING-92**, Nantes, 1992.

[8] Dymetman M., **Transformations de grammaires logiques et réversibilité en traduction automatique**, thèse d'État, Université de Grenoble 1, France, 1992.

[9] Dymetman M., Foster G., Isabelle P., **Towards an Automatic Dictation System for Translators (Transtalk)**, Tech. report, CITI, Laval, Quebec, Canada, 1992.

[10] Foster G., **Statistical Lexical Disambiguation.** Master's thesis, McGill University, School of Computer Science, 1991.

[11] Gurstein M. and Monette M. **Functional Specifications for a Translator's Workstation**. Technical Report 12SD.36902-5-0003, Socioscope Inc, Ottawa, Canada, October 1988. Report submitted to the Canadian Workplace Automation Research Center.

[12] Isabelle P., *Machine Translation at the TAUM Group*, in Margaret King (ed.), **Machine Translation Today: The State of the Art**, Edinburgh University Press, 1987.

[13] Isabelle P., *Bi-Textual Aids for Translators*, **Proceedings of the Eight Annual Conference of the UW Centre for the New OED and Text Research**, University of Waterloo, Waterloo, Canada, 1992.

[14] Isabelle P., Dymetman M., Macklovitch E., *CRITTER: a Translation System for Agricultural Market Reports*, **Proceedings of COLING-88**, Budapest, 1988.

[15] Jelinek F. *Self-Organized Modeling for Speech Recognition*, in Alex Waibel and Kai-Fu Lee, editors, **Readings in Speech Recognition**, pages 450--506. Morgan Kaufmann, San Mateo, CA, 1990.

[16] Macklovitch E., *Corpus-Based Tools for Translators*, P**roceedings of the 33rd Annual Conference of the American Translators Association**, San Diego, 1992.

[17] Macklovitch E., **A Third Version of the CWARC's Workstation for Translators**, Tech. report, CITI, Laval, Quebec, Canada, 1993.

[18] Kay M., **The Proper Place of Men and Machines in Translation**, CSL-80-11, Xerox PARC, 1980.

[19] Sato S., Nagao M., *Toward Memory-Based Translation*, **Proceedings of COLING-90**, 247-252, 1990.

[20] Simard M., Foster G., Isabelle P. *Using Cognates to Align Sentences in Parallel Corpora*, **Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation**, Montreal, 1992.

[21] Van Noord G., **Reversibility in Natural Language Processing**, CIP-Gegevens Konincklijke Bibliotheek, The Hague, 1993.

[22] Van Roey J., Granger S., Swallow H., **Dictionnaire des faux-amis français-anglais**, Paris, Duculot, 1988.