wEBMT: Developing and Validating an Example-Based Machine Translation System Using the World Wide Web

Andy Way*
Dublin City University

Nano Gough†
Dublin City University

We have developed an example-based machine translation (EBMT) system that uses the World Wide Web for two different purposes: First, we populate the system's memory with translations gathered from rule-based MT systems located on the Web. The source strings input to these systems were extracted automatically from an extremely small subset of the rule types in the Penn-II Treebank. In subsequent stages, the ⟨source, target⟩ translation pairs obtained are automatically transformed into a series of resources that render the translation process more successful. Despite the fact that the output from on-line MT systems is often faulty, we demonstrate in a number of experiments that when used to seed the memories of an EBMT system, they can in fact prove useful in generating translations of high quality in a robust fashion. In addition, we demonstrate the relative gain of EBMT in comparison to on-line systems. Second, despite the perception that the documents available on the Web are of questionable quality, we demonstrate in contrast that such resources are extremely useful in automatically postediting translation candidates proposed by our system.

1. Introduction

In quite a short space of time, translation memory (TM) systems have become a very useful tool in the translator's armory. TM systems store a set of ⟨source, target⟩ translation pairs in their databases. If a new input string cannot be found exactly in the translation database, a search is conducted for close (or "fuzzy") matches of the input string, and these are retrieved together with their translations for the translator to manipulate into the final, output translation. From this description, it should be clear that TM systems do *not* translate: Indeed, some researchers consider them to be little more than a search-and-replace engine, albeit a rather sophisticated one (Macklovitch and Russell 2000).

We can illustrate this with respect to the TM entries in (1), taken from the Canadian Hansards:

(1) a. While most were critical, some contributions were thoughtful and constructive ⟹ La plupart ont formulé des critiques, mais certains ont fait des observations réfléchies et constructives.
b. Others were plain meanspirited and some contained errors of fact ⟹ D'autres discours comportaient des propos mesquins et même des erreurs de fait.

* School of Computing, Dublin 9, Ireland. E-mail: away@computing.dcu.ie
† School of Computing, Dublin 9, Ireland. E-mail: ngough@computing.dcu.ie

Consider the new source string in (2):

(2) While most were critical, some contributions were plain meanspirited.

Despite the fact that this new input in (2) is extremely close to the source strings in the TM entries in (1), no TM system containing just these translation pairs in its database would be able to translate (2); the best they could do would be to identify one or both of the two source sentences in the TM in (1) as fuzzy matches and display these, together with their French translations. The translator would then manipulate the target strings in the TM into the final translation (3):

(3) La plupart ont formulé des critiques, mais certains ont fait des observations mesquines.

An alternative translation that might be derived from the TM entries in (1) is that in (4):

(4) La plupart ont formulé des critiques, mais certains comportaient des observations mesquines.

At all stages in the translation process, therefore, the translators themselves are the integral figures: They are free to accept or reject any suggested matches, they construct the translations, and they may or may not use any translations proposed by the TM system to formulate the translations in the target document. Finally, they are free to insert the translations produced into the TM itself as they see fit: that is, either (3) or (4) could be inserted into the TM with the source string (2), or some other translation, if that were preferred.

A prerequisite for TM (and example-based machine translation [EBMT]) applications is a parallel corpus aligned at sentential level. Such a corpus may be presented to translators *en bloc*, or translators may help construct it themselves. Here too the translator maintains a large degree of autonomy: Using a tool such as *Trados WinAlign*, for example, he or she may manually overwrite some of the aligner's decisions by linking ⟨source, target⟩ sentence pairs using the graphical interface provided.

Nevertheless, TM systems are currently falling far short of their potential, given the limitation that the smallest accessible translation units are ⟨source, target⟩ strings aligned only at sentential level. Consider the fuzzy matching operation, for instance: Translators are able to set a fuzzy match threshold below which no translation pairs are proposed by the TM system. If this threshold is set too low, then potentially useful translation pairs will be presented along with a lot of noise, thereby risking that this useful translation information will be obscured (high recall, low precision); if it is set too low, then good matches will be presented, but potentially useful matches will not be (low recall, high precision). We noted above that faced with the new input in (2), a TM system might be able to present the translator with the fuzzy matches in (1). However, if a translator were to set the level of fuzzy matching at 80% (a not unreasonable level), then neither of the translation pairs in (1) would be deemed to be a suitably good fuzzy match, as only 7/9 (77%) of the words in (1a) match those in (2) exactly, and only 3/9 (33%) of the words in (1b) match those in (2) exactly. Indeed, setting an appropriate fuzzy match level is such a difficult problem that some translators switch off this option and use the TM only to find exact matches.

If subsentential alignment could be integrated into the TM databases, more useful fragments could be put at the disposal of the translator. If we could fragment the

sententially aligned TM examples in (1) so that subsentential chunks were displayed to the user, then the chance of finding exact matches or good fuzzy matches would increase considerably. This is currently beyond the scope of TM systems.

In contrast, EBMT systems have overcome this constraint by storing subsentential translational correspondences in addition to the sententially aligned pairs from which they are derived. As a consequence, where a TM system can only propose a number of close-scoring matches in its database for the translator to adapt into the final translation, an EBMT system can produce translations itself by automatically combining chunks from different translation examples stored in its memories.

In Section 2, we describe how we automatically obtain a hierarchy of lexical resources that are used sequentially by our EBMT system, *wEBMT*, to translate new input. The primary resource gathered is a "phrasal lexicon," constructed by extracting over 200,000 phrases from the Penn Treebank and having them translated into French by three Web-based machine translation (MT) systems.

Each set of translations is stored separately, and for each set the "marker hypothesis" (Green 1979) is used to segment the phrasal lexicon into a "marker lexicon." The marker hypothesis is a universal psycholinguistic constraint which states that natural languages are "marked" for complex syntactic structure at surface form by a closed set of specific lexemes and morphemes. That is, a basic phrase-level segmentation of an input sentence can be achieved by exploiting a closed list of known marker words to signal the start and end of each segment.

Consider the following example, selected at random from the *Wall Street Journal* section of the Penn-II Treebank:

(5) The Dearborn, Mich., energy company stopped paying a dividend in the third quarter of 1984 because of troubles at its Midland nuclear plant.

Here we see that three noun phrases start with determiners and one with a possessive pronoun. The sets of determiners and possessive pronouns are both very small. Furthermore, there are four prepositional phrases, and the set of prepositions is similarly small. A further assumption that could be made is that all words that end with *-ed* are verbs, such as *stopped* in (5). The marker hypothesis is arguably universal in presuming that concepts and structures like these have similar morphological or structural marking in all languages.

The marker hypothesis has been used for a number of different language-related tasks, including

- language learning (Green 1979; Mori and Moeser 1983; Morgan, Meier, and Newport 1989)
- monolingual grammar induction (Juola 1998)
- grammar optimization (Juola 1994)
- insights into universal grammar (Juola 1998)
- machine translation (Juola 1994, 1997; Veale and Way 1997; Gough, Way, and Hearne 2002)

With respect to translation, a potential problem in using the marker hypothesis is that some languages do not have marker words such as articles, for instance. Green's (1979) work showed that artificial languages, both with and without specific marker words, may be learned more accurately and quickly if such psycholinguistic cues exist. The

research of Mori and Moeser (1983) showed a similar effect due to case marking on pseudowords in such artificial languages, and Morgan, Meier, and Newport (1989) demonstrated that languages that do not permit pronouns as substitutes for phrases also provide evidence in favor of the marker hypothesis. Juola's (1994, 1998) work on grammar optimization and induction shows that context-free grammars can be converted to "marker-normal form." However, marker-normal form grammars cannot capture the sorts of regularities demonstrated for languages that do not have a one-to-one mapping between a terminal symbol and a word. Nevertheless, Juola (1998, page 23) observes that "a slightly more general mapping, where two adjacent terminal symbols can be merged into a single lexical item (for example, a word and its case-marking), can capture this sort of result quite handily." Work using the marker hypothesis for MT adapts this monolingual mapping for pairs of languages: It is reasonably straightforward to map an English determiner-noun sequence onto a Japanese noun–case marker segment, once one has identified the sets of marker tags in the languages to be translated.

Following construction of the marker lexicon, the ⟨source, target⟩ chunks are generalized further using a methodology based on Block (2000) to permit a limited form of insertion in the translation process. As a byproduct of the chosen methodology, we also derive a standard "word-level" translation lexicon. These various resources render the set of original translation pairs far more useful in deriving translations of previously unseen input.

In Section 3, we describe in detail the segmentation process, together with the procedure whereby target chunks are combined to produce candidate translations. In Section 4, we report initially on two experiments in which we test different versions of our EBMT system against test sets of NPs and sentences. We then conduct a set of further experiments which show that using the resources developed from more than one on-line MT system may improve both translation coverage and quality. Furthermore, seeding the system databases with more fragments improves translation quality. In addition, we calculate the net gain of our EBMT system by comparing translation quality against that of the three on-line MT systems. Finally, we comment on the relative strengths and weaknesses of the three on-line MT systems used.

Like most EBMT systems, our approach suffers from the problem of "boundary friction": where chunks from different translation examples are recombined, the quality of the resulting translations may be compromised. Assume that the aligned examples in (6) are located in the system database:

(6) a. You can attach a phone to the connector $\Longrightarrow$ Vous pouvez rélier un téléphone au connecteur.
b. Connect only the keyboard and a mouse $\Longrightarrow$ Connectez uniquement le clavier et une souris.

Let us now confront the EBMT system with the new input string in (7):

(7) You can attach a mouse to the connector.

This could be correctly translated by the EBMT system by isolating the useful translation fragments in (8):

(8) a. You can attach $\Longrightarrow$ Vous pouvez rélier (from (6a))
b. a mouse $\Longrightarrow$ une souris (from (6b))
c. to the connector $\Longrightarrow$ au connecteur (from (6a))

Recombining the French chunks gives us the correct translation in (9):

(9) Vous pouvez rélier une souris au connecteur.

However, a number of mistranslations could also ensue, including those in (10):

(10) a. *Vous pouvez rélier un souris au connecteur.
b. *Vous pouvez rélier un souris au le connecteur.

The mistranslation (10a) could be formed via the set of inferences in (11):

(11) You can attach a $\Longrightarrow$ Vous pouvez rélier un (from (6a))
mouse $\Longrightarrow$ souris (from (6b))
to the connector $\Longrightarrow$ au connecteur (from (6a))

The mistranslation (10b) could be formed via the set of inferences in (12):

(12) You can attach a $\Longrightarrow$ Vous pouvez rélier un (from (6a))
mouse $\Longrightarrow$ souris (from (6b))
to $\Longrightarrow$ au (from (6a))
the $\Longrightarrow$ le (from (6b))
connector $\Longrightarrow$ connecteur (from (6a))

It is clear, therefore, that unless the process by which the original ⟨source, target⟩ sentence pairs are fragmented is well defined and strictly controlled, chunks may be combined from different contexts that result in agreement errors such as those in (10).[1] Depending on the input string, our *wEBMT* system may generate thousands of candidate translations, including many mistranslations like those in (10). A major advantage of MT systems based on probabilities is that output translations can be ranked (and pruned, if required): One would hope that such systems would rank good translations such as that in (9) more highly than poor ones such as those in (10). We demonstrate that in almost all experiments, our EBMT system consistently ranks the "best" translation in the top 10 output translations, and always in the top 1% of the translations generated.

In order to minimize errors of boundary friction, in Section 5 we develop a novel, post hoc procedure via the World Wide Web to validate and, if necessary, correct translations prior to their being output to the user.[2] Finally we conclude and point to areas of future research.

1 Note also that with respect to the translations given in (3) and (4), the translator interacting with the TM has used his or her translation knowledge to avoid a problem of boundary friction: Given the TM entries in (1), the translation of *plain meanspirited* would appear to be *mesquins*. This is correct in this context, as it co-occurs with a masculine plural noun *propos*. In translating (2), however, *observations* is a feminine plural noun, so the adjective *mesquines* is inserted to maintain agreement throughout the NP. If the translation pair ⟨*plain meanspirited, mesquines*⟩ were not found in the system's memories, then only the mistranslation *observations mesquins* could be produced by an EBMT system.

2 One of the areas of boundary friction that we use our post hoc validation procedure to correct is that of subject-verb agreement. Note that with examples such as (18), this is not usually (such) a problem for marker-based approaches to MT as we face here, as verbs are contained within (part of) the same chunk as their subject NPs. However, given that we translate phrases rather than sentences, it is a considerable problem for our approach, yet one that we overcome satisfactorily. In further work, if we were to store the translations of the VPs with their dummy subject NPs in a sentential lexicon and derive all marker lexicons from this database, the problem of subject-verb agreement would be largely overcome.

2. Deriving Translation Resources from Web-Based MT Systems

All EBMT systems, from the initial proposal by Nagao (1984) to the recent collection of Carl and Way (2003), are premised on the availability of subsentential alignments derived from the input bitext. There is a wealth of literature on trying to establish subsentential translations from a bilingual corpus.[3] Kay and Röscheisen (1993) attempt to extract a bilingual dictionary using a hybrid method of sentence and word alignment on the assumption that the ⟨source, target⟩ words have a similar distribution. Fung and McKeown (1997) attempt to translate technical terms using word relation matrices, although the resource from which such relations are derived is a pair of nonparallel corpora. Somers (1998) replicates the work of Fung and McKeown with different language pairs using the simpler metric of Levenshtein distance. Boutsis and Piperidis (1998) use a tagged parallel corpus to extract translationally equivalent English-Greek clauses on the basis of word occurrence and co-occurrence probabilities. The respective lengths of the putative alignments in terms of characters is also an important factor. Ahrenberg, Andersson, and Merkel (2002) observe that for less widely spoken languages, the relative lack of linguistic tools and resources has forced developers of word alignment tools for such languages to use shallow processing and basic statistical approaches to word linking. Accordingly, they generate lexical correspondences by means of co-occurrence measures and string similarity metrics.

More specifically, the notion of the phrasal lexicon (used first by Becker 1975) has been used successfully in a number of areas:

- Learnability (Zernik and Dyer 1987)
- Text generation (Hovy 1988; Milosavljevic, Tulloch, and Dale 1996)
- Speech generation (Rayner and Carter 1997)
- Localization (Schäler 1996)

More recently, Simard and Langlais (2001) have proposed the exploitation of TMs at a subsentential level, while Carl, Way, and Schäler (2002) and Schäler, Way, and Carl (2003, pages 108–109) describe how phrasal lexicons might come to occupy a central place in a future hybrid integrated translation environment. This, they suggest, may result in a paradigm shift from TM to EBMT via the phrasal lexicon: Translators are on the whole wary of MT technology, but once subsentential alignment is enabled, translators will become aware of the benefits to be gained from ⟨source, target⟩ phrasal segments, and from there they suggest that "it is a reasonably short step to enabling an automated solution via the recombination element of EBMT systems such as those described in [Carl and Way 2003]."

In this section, we describe how the memory of our EBMT system is seeded with a set of translations obtained from Web-based MT systems. From this initial resource, we subsequently derive a number of different databases that together allow many new input sentences to be translated that it would not be possible to translate in other systems. First, the phrasal lexicon is segmented using the marker hypothesis to produce a marker lexicon. This is then generalized, following a methodology based on Block (2000), to generate the "generalized marker lexicon." Finally, as a result of the

3 We refer the interested reader to the excellent and comprehensive bibliography on parallel text processing available at http://www.up.univ-mrs.fr/~veronis/biblios/ptp.htm.

methodology chosen, we automatically derive a fourth resource, namely, a "word-level lexicon."

2.1 The Phrasal Lexicon

Our phrasal lexicon was built by selecting a set of 218,697 English noun phrases and verb phrases from the Penn Treebank. We identified all rule types occurring 1,000 or more times and eliminated those that were not relevant (e.g., rules dealing only with numbers). Where rules contained just a single nonterminal on their right-hand side, only those rules whose left-hand side was VP were retained in order to ensure that we could handle intransitive verbs. In total, 59 rule types out of a total of over 29,000 (i.e., just 0.002% of the rules in Penn-II) were used in creating the various lexical resources. For each of these 59 rule types, the tokens corresponding to the rule right-hand sides were extracted. These extracted English phrases were then translated using three on-line MT systems:

- SDL International's *Enterprise Translation Server*[4] (system A)
- *Reverso* by Softissimo[5] (system B)
- *Logomedia*[6] (system C)

Translating the NPs via these MT systems was reasonably straightforward. We report in Section 4 on the quality of the French NPs produced, and in Section 5 we discuss experiments designed to discover whether our EBMT system could improve on any mistranslations obtained. Translating the VPs involved a little more thought: In the main, on-line MT systems such as these work far better when they translate sentences. In order to obtain finite verb forms rather than the default infinitival forms, we provided dummy subjects. Initially these were third-person plural pronouns, which caused similar verb forms to be created. This obviously biases the EBMT system more in favor of third-person plural sentences. Nevertheless, using the WWW-based post hoc evaluation methodology proposed in Section 5, we were still able to obtain reasonable translations for non-third-person-plural sentences too. In a subsequent experiment, we seed the databases of *wEBMT* with third-person singular verb forms by providing third-person singular pronouns as the dummy subjects, and in a final experiment we combine all third-person fragments (both singular and plural) into the system's memories and compare results on the same test set.

The on-line MT systems were selected purely because they enable batch translation of large quantities of text. In our experience, the most efficient way to translate large amounts of data via on-line MT systems is to send each document as an HTML page with the phrases to be translated encoded as an ordered list. We automatically tagged the English phrases with HTML codes and input them into each translation system using the Unix `wget` function, which takes a URL as input and writes the corresponding HTML document to a file. If the URL takes the form of a query, then the document retrieved is the result of the query, namely, the translated Web page. Once this is obtained, it is a simple process to retrieve the French translations and associate them with their English source equivalents.

4 http://www.freetranslation.com
5 http://trans.voila.fr
6 http://www.logomedia.net

2.2 The Marker Lexicons

Given that the marker hypothesis is arguably universal, it is clear that benefits may accrue by using it to facilitate subsentential alignment of ⟨source, target⟩ chunks. Juola (1994, 1997) conducts some small experiments using his *METLA* system to show the viability of this approach for English ⟶ French and English ⟶ Urdu. For the English ⟶ French language pair, Juola gives results of 61% correct translation when the system is tested on the training corpus, and 36% accuracy when it is evaluated with test data. For English ⟶ Urdu, Juola (1997, page 213) notes that "the system learned the original training corpus . . . perfectly and could reproduce it without errors"; that is, it scored 100% accuracy when tested against the training corpus. On novel test sentences, he gives results of 72% correct translation. In their *Gaijin* system, Veale and Way (1997) give a result of 63% accurate translations obtained for English ⟶ German on a test set of 791 sentences from *CorelDRAW* manuals.

As in *METLA* and *Gaijin*, we exploit lists of known marker words for each language to indicate the start and end of segments. For English, our source language, we use the sets of marker words in (13):

(13)

<DET>	{the, a, an, those, these, . . . }
<PREP>	{in, on, out, with, from, to, under, . . . }
<QUANT>	{all, some, few, many, . . . }
<CONJ>	{and, or, . . . }
<POSS>	{my, your, our,. . . }
<PRON>	{I, you, he, she, it,. . . }

A similar set (14) was produced for French, the target language in our *wEBMT* system:

(14)

<DET>	{le, la, l', les, ce, ces, ceux, cet, . . . }
<PREP>	{dans, sur, avec, de, à, sous, . . . }
<QUANT>	{tous, tout, toutes, certain, quelques, beaucoup, . . . }
<CONJ>	{et, ou, . . . }
<POSS>	{mon, ma, mes, ton, ta, tes, notre, nos, . . . }
<PRON>	{je, j', tu, il, elle, . . . }

In a preprocessing stage, the aligned ⟨source, target⟩ pairs in the phrasal lexicon are traversed word by word, and whenever any such marker word is encountered, a new chunk is begun, with the first word labeled with its marker category (<DET>, <PREP>, etc.). The example in (15) illustrates the results of running the marker hypothesis over the source phrase *all uses of asbestos*:

(15)

<QUANT>	all uses
<PREP>	of asbestos

In addition, we impose a further constraint that each chunk must also contain at least one non–marker word, so that the phrase *out in the cold* will be viewed as one segment (labeled with <PREP>), rather than split into still smaller chunks.

For each $\langle English, French_X \rangle$ pair, where X is one of the sets of translations derived from the three separate MT on-line systems (see above), we derive separate marker lexicons for each of the 218,697 source phrases and target translations. This gives

us a total of 656,091 ⟨source, target⟩ translation pairs (including many repetitions, of course). Given that English and French have essentially the same word order, these marker lexicons are predicated on the naïve yet effective assumption that marker-headed chunks in the source *S* map sequentially to their target equivalents *T*; that is, $\text{chunk}_{S1} \longrightarrow \text{chunk}_{T1}$, $\text{chunk}_{S2} \longrightarrow \text{chunk}_{T2}, \ldots \text{chunk}_{Sn} \longrightarrow \text{chunk}_{Tn}$, subject to their marker categories matching, where possible. Using the previous example of *all uses of asbestos*, this gives us the marker chunks in (16):

(16)
<QUANT>	all uses	:	tous usages
<PREP>	of asbestos	:	d' asbeste

Sometimes the number of marker chunks in the two languages differs, with respect to both the marker categories and the number of chunks obtained. Consider the example in (17):

(17) The man looks at the woman $\Longrightarrow$ L'homme regarde la femme.

Once the marker hypothesis is applied to (17), it would be marked up as in (18):

(18) <DET> The man looks <PREP> at <DET> the woman $\Longrightarrow$
<DET> L' homme regarde <DET> la femme.

That is, the English verb subcategorizes for a PP complement which in this case contains two marker words, whereas the French verb *regarder* is a straightforward transitive verb. It may appear, therefore, that there are three chunks in the English string and only two on the French side, but this is not the case: The restriction that each segment must contain at least one non–marker word ensures that we have just two marker chunks for the English string in (18). However, it remains the case that the chunks are tagged differently; we obtain the marker chunks in (19):

(19) *English*:
<DET> The man looks
<PREP> at <DET> the woman

French:
<DET> L' homme regarde
<DET> la femme

Our alignment method would therefore align the first English chunk with the first French chunk, as their marker categories match. Note, of course, that this contains a translation error: *regarde* translates not as *looks* but rather as *looks at*. Errors such as this will adversely affect translation quality, but as we report in Section 4, good-quality translations are obtainable on the whole. The second pair in (19), however, cannot be mapped straightforwardly onto one another, as the marker categories differ. Nevertheless, our algorithm would align "<DET> the woman" with "<DET> la femme," as their marker categories match. This ensures that as many potentially useful translation fragments are generated as possible.

This naïve alignment procedure works well between (broadly) similar languages such as English and French, but there are cases even between quite closely related languages in which the procedure breaks down. In order to increase translation quality

still further, the mapping function needs to be improved to account for examples such as (20):

(20) The man likes the woman ⟹ La femme plaît à l'homme.

The *like* ⟹ *plaire* case is an argument-switching (or relation-changing) example, in that the subject in English becomes the indirect object in French, and the English object translates as the French subject. If we were to apply the marker hypothesis to (20), we would derive (21):

(21) <DET> The man likes <DET> the woman ⟹
<DET> La femme plaît <PREP> à <DET> l' homme.

That is, without recourse to a lexicon or information about the relative distribution of words and their translations, we would derive the marker chunks in (22):

(22) a. <DET> The man likes ⟹ <DET> La femme plaît
b. <DET> the woman ⟹ <DET> l' homme

Of course, both alignments are wrong. However, our alignment method correctly aligns ⟨source, target⟩ segments in approximately 80% of cases. We calculate this as an approximation by testing all translations of marker chunks to see whether these French chunks appear *anywhere* on the Web: If so, we assume that the translations obtained by the online MT systems are correct. For 39,895 such translations, 75.2% of those produced by system A appear on the Web, with 81.7% of those generated by system B and 81.5% of those produced by system C also appearing on the Web. Note that this gives us only an approximation of the correctness of our alignments, as we are testing whether the French translations are "good French" rather than whether the alignments in which they appear are actually correct.

Correcting misalignments such as those in (22) is a topic for further research. Adding a bilingual lexicon (our word-level lexicon, for example) and incorporating the constraints contained therein into the marker-based alignment process would prevent chunks such as those in (22) from being generated, and we conjecture that translation quality would improve accordingly.

Given marker chunks such as those in (16), we are able to extract automatically a further bilingual dictionary, the word-level lexicon. We take advantage of the assumption that where a chunk contains just one non–marker word in both source and target, these words are translations of each other. Where a marker-headed pair contains just two words, as in (16), for instance, we can extract the word-level translations in (23):

(23)

<QUANT>	all	:	tous
<PREP>	of	:	d'
<LEX>	uses	:	usages
<LEX>	asbestos	:	asbeste

That is, using the marker hypothesis method of segmentation, smaller aligned segments can be extracted from the phrasal lexicon without recourse to any detailed parsing techniques or complex co-ocurrence measures.

Juola (1994, 1997) assumes that words ending in *-ed* are verbs. However, given that verbs are not a closed class, in our approach we do not mark chunks beginning with a verb with any marker category. Instead, we take advantage of the fact that the initial phrasal chunks correspond to rule right-hand sides. That is, for a rule in the Penn Treebank VP ⟶ VBG, NP, PP, we are certain (if the annotators have done their job correctly) that the first word in each of the strings corresponding to this right-hand side is a VBG, that is, a present participle. Given this information, in such cases we tag such words with the <LEX> tag. Taking *expanding the board to 14 members ⟶ augmente le conseil à 14 membres* as an example, we extract the chunks in (24):

(24)

<DET>	the board	:	le conseil
<DET>	the	:	le
<PREP>	to <QUANT> 14 members	:	à 14 membres
<QUANT>	14 members	:	14 membres
<LEX>	expanding	:	augmente
<LEX>	board	:	conseil
<PREP>	to	:	à
<LEX>	members	:	membres

We ignore here the trivially true lexical chunk "<QUANT> 14 : 14."

In a final processing stage, we generalize over the marker lexicon following a process found in Block (2000). In Block's approach, word alignments are assigned probabilities by means of a statistical word alignment tool. In a subsequent stage, chunk pairs are extracted, which are then generalized to produce a set of translation templates for each ⟨source, target⟩ segment.

Block distinguishes chunks from "patterns," as we do: His chunks are similar to our marker chunks, and his patterns are similar to our generalized marker chunks. Once chunks are derived from ⟨source, target⟩ alignments, patterns are computed from the derived chunks by means of the following algorithm: "for each pair of chunk pairs $\langle\langle C_{S1}, C_{T1}\rangle, \langle C_{S2}, C_{T2}\rangle\rangle$, if C_{S1} is a substring in C_{S2} and C_{T1} is a substring in C_{T2}, then $\langle P_S, P_T\rangle$ is a pattern pair where P_S equals C_{S2} with C_{S1} replaced by a variable V and P_T equals C_{T2} with C_{T1} replaced by V" (Block 2000, pages 414–415). Block then gives an example that shows how patterns are derived. Assume the chunk pairs in (25):

(25)
⟨ [das], [which] ⟩
⟨ [ist], [is] ⟩
⟨ [was], [what] ⟩
⟨ [Sie], [you] ⟩
⟨ [wollten], [wanted] ⟩
⟨ [das ist], [which is] ⟩
⟨ [das ist was], [which is what] ⟩
⟨ [das ist was Sie], [which is what you] ⟩
⟨ [das ist was Sie wollten], [which is what you wanted] ⟩

Using the algorithm described above, the patterns in (26) are derived from the chunks in (25):

(26) ⟨ [*V* ist], [*V* is] ⟩
⟨ [das *V*], [which *V*] ⟩
⟨ [das *V* was], [which *V* what] ⟩
⋮
⟨ [*V* ist was Sie], [*V* is what you] ⟩
⋮
⟨ [das ist was *V* wollten], [which is what *V* wanted] ⟩
⋮

Of course, many other researchers also try to extract generalized templates. Kaji, Kida, and Morimoto (1992) identify translationally equivalent phrasal segments and replace such equivalents with variables to generate a set of translation patterns. Watanabe (1993) combines lexical and dependency mappings to form his generalizations. Other similar approaches include those of Cicekli and Güvenir (1996), McTait and Trujillo (1999), Carl (1999), and Brown (2000), inter alia.

In our system, in some cases the smallest chunk obtainable via the marker-based segmentation process may be something like (27):

(27) <DET> the good man : le bon homme

In such cases, if our system were confronted with *a good man*, it would not be able to translate such a phrase, assuming this to be missing from the marker lexicon. Accordingly, we convert examples such as (27) into their generalized equivalents, as in (28):

(28) <DET> good man : bon homme

That is, where Block (2000) substitutes variables for various words in his templates, we replace certain lexical items with their marker tag. Given that examples such as "<DET> a : un" are likely to exist in the word-level lexicon, they may be inserted at the point indicated by the marker tag to form the correct translation *un bon homme*. We thus cluster on marker words to improve the coverage of our system (see Section 5 for results that show exactly how clustering on marker words helps); others (notably Brown [2000, 2003]) use clustering techniques to determine equivalence classes of individual words that can occur in the same context, and in so doing derive translation templates from individual translation examples.

2.3 Summary

In sum, we automatically create four knowledge sources:

- the original ⟨source,target⟩ phrasal translation pairs
- the marker lexicon (cf. (16))
- the generalized marker lexicon (cf. (28))
- the word-level lexicon (cf. (24))

When matching the input to the corpus, we search for chunks in the order given here, that is, from specific examples (those containing more context) to generic (those containing less context). We give in (29) an example of how a particular sentence from our test set is translated via these different knowledge sources:

(29) *Input*:
A major concern for the parent company is what advertisers are paying per page.

Chunks found in marker lexicon:
for the parent company : pour la société mère
what advertisers are paying per page : quels annonceurs paient per page

Chunk found in generalized marker lexicon:
<DET> major concern : inquiétude majeure

Words found in word-level lexicon:
<DET> a : une
<LEX> is : est

Given the fragments shown in (29), a translation can now be derived. First, the word pair "<DET> a : une" is inserted into the generalized template "<DET> major concern: inquiétude majeure" to begin the translation process; the next chunk, "for the parent company : pour la société mère," is retrieved from the marker lexicon; the missing word pair "<LEX> is : est" is retrieved from the word-level lexicon; and finally, the marker chunk "what advertisers are paying per page : quels annonceurs paient per page" is appended to produce the translation in (30):

(30) Une inquiétude majeure pour la société mère est quels annonceurs paient per page.

Of course, this "translation" is not without problems: There is a poor (in this instance) translation of *what* as *quels*, and a nontranslation of *per*. There is little our system can do about errors such as these made by the on-line MT systems. Nevertheless, (29) illustrates how the various knowledge sources play a part in determining the final translation in our system.

Note that none of these aligned resources would be possible in a TM system. The problem of segmentation is not an inconsiderable one in all EBMT systems, but we (and others) have found that using the marker hypothesis can greatly facilitate such a process. We shall show in subsequent sections that because such knowledge sources are derived automatically from the original translations obtained via Web-based MT systems, the translations obtained in our EBMT process are largely of high quality, are ranked highly in the set of output translation candidates, and may be generated in almost all cases—all this despite the fact that the original translations obtained via the Web contain many errors, and that the source phrases to be translated were selected from a mere fraction of the rule types in the Penn-II Treebank.

3. Retrieving Chunks and Producing Translations

In Section 4, we report on a number of experiments using the resources obtained in the previous section to translate two test sets of data, one a set of NPs and the other

a set of sentences. Although we are primarily interested in translating sentences, we translate NPs for two reasons: (1) to assure ourselves that we are in fact translating nominal chunks correctly, and (2) to see whether our methodology can actually correct any NPs mistranslated by the three on-line MT systems. In this section, we describe the processes involved in retrieving appropriate chunks and forming translations for NPs only (these being fewer in number than for sentences, of course).

3.1 Segmentation of the Input

In many cases, a 100% match for a given NP cannot be found in the phrasal lexicon. In order to try and process the NP in a compositional manner, it is segmented into smaller chunks, and the system then attempts to locate these chunks individually and to retrieve their relevant translation(s) from the various lexicons described above. We use an *n*-gram-based segmentation method. Initially, we located all possible bigrams, trigrams and so on within the input string and then searched for these within the relevant knowledge sources.

However, many of these *n*-grams cannot be found by our system, given that new chunks are placed in the marker lexicon when a marker word is found in a sentence. Taking the NP *the total at risk a year* as an example, chunks such as *the total at risk a* or *at risk a* cannot be located, as new chunks would be formed at each marker word (assuming the adjacent word is a non–marker word), so the best that could be expected here might be to find the chunks in (31):

(31) <DET> the total, <PREP> at risk, <DET> a year

The respective translations of these chunks would then be recombined to form the target string. In a recent addition to our work, we have eliminated certain *n*-grams (such as those that end in a marker word, for instance) from the search process, as these would never be found given our chosen method of segmentation.

3.2 Retrieving Translation Chunks

We use translations retrieved from the three different on-line MT systems specified above (see Section 2.1). These translations are further broken down using the marker hypothesis to provide us with an additional three knowledge sources A′, B′, and C′, a marker lexicon, generalized marker lexicon and word-level lexicon derived from chunks produced by each system. These knowledge sources can be combined in several different ways. We have produced translations using

- information from a single source: A/A′, B/B′, or C/C′, that is, a phrasal lexicon and set of marker lexicons derived from translations produced by each on-line system
- information from pairs of sources: A/A′ and B/B′, A/A′ and C/C′ or B/B′ and C/C′, that is, phrasal and marker lexicons derived from translations produced by two different on-line systems
- information from all available knowledge sources: A/A′ and B/B′ and C/C′, that is, phrasal and marker lexicons derived from translations produced by all three on-line systems

The objective here is to see how much translation coverage and quality are improved by using chunks derived from multiple sources. Assuming that the English strings are

not translated in exactly the same manner by the three on-line MT systems means that more knowledge sources could be combined in attempting to translate the new input contained in the test sets of noun phrases and sentences. Results from experiments conducted using multiple knowledge sources are given in Section 4.2.

3.3 Calculation of Weights

Each time a source language (SL) chunk is submitted for translation, the appropriate target language (TL) chunks are retrieved and returned with a weight attached. We use a maximum of six knowledge sources:

- Stage 1: Three sets of translations (A, B, and C) are retrieved using each of the three on-line MT systems.
- Stage 2: Three sets of translations (A′, B′, and C′) acquired by breaking down the translations retrieved in Stage 1 using the marker hypothesis to form the marker lexicon, the generalized marker lexicon, and the word-level lexicon.

Within each knowledge source, each translation is weighted according to the formula in (32):

$$\text{(32)} \qquad \text{weight} = \frac{\textit{number of occurrences of the proposed translation}}{\textit{total number of translations produced for SL phrase}}$$

For the SL phrase *the house*, assuming that *la maison* is found eight times and *le domicile* is found twice, then $P(\text{la maison} \mid \text{the house}) = 8/10$ and $P(\text{le domicile} \mid \text{the house}) = 2/10$. Note that since each SL phrase will only have one proposed translation within each of the knowledge sources acquired at Stage 1, these translations will always have a weight of 1.

If we wish to consider only those translations produced using a single MT system (e.g., A and A′), we add the weights of translations found in both knowledge sources and divide the weights of all proposed translations by two. For the SL phrase *the house*, assuming $P(\text{la maison} \mid \text{the house}) = 5/10$ in knowledge source A and $P(\text{la maison} \mid \text{the house}) = 8/10$ in A′, then $P(\text{la maison} \mid \text{the house}) = 13/20$ over both knowledge sources. Similarly, if we wish to consider translations produced by all three MT systems, then we add the weights of common translations and divide the weights of all proposed translations by six.

When translated phrases have been retrieved for each chunk of the input string, they must then be combined to produce an output string. In order to calculate a ranking for each TL sentence produced, we multiply the weights of each chunk used in its construction. Note that this ensures that greater importance is attributed to longer chunks, as is usual in most EBMT systems (cf. Sato and Nagao 1990; Veale and Way 1997; Carl 1999).[7]

As an example, consider the translation into French of *the house collapsed*. Assume the conditional probabilities in (33):

7 Note that approaches that prefer the greatest context to be taken into account are not limited to EBMT. Research in the area of data-oriented parsing (cf. Bod, Scha, and Sima'an, 2003) also shows that unless the corpus is inherently biased, derivations constructed using the smallest number of subtrees have a higher probability than those built with a larger number of smaller subtrees.

(33) a. P(la maison | the house) = 8/10
b. P(le domicile | the house) = 2/10
c. P(s'écroula | collapsed) = 1/7
d. P(s'effondra | collapsed) = 6/7

Given the weights in (33), the four translations in (34) can be produced, each with an associated probability:

(34) a. P(la maison s'écroula | the house collapsed) = $\frac{8}{10} \cdot \frac{1}{7} = \frac{8}{70}$
b. P(le domicile s'écroula | the house collapsed) = $\frac{2}{10} \cdot \frac{1}{7} = \frac{2}{70}$
c. P(la maison s'effondra | the house collapsed) = $\frac{8}{10} \cdot \frac{6}{7} = \frac{48}{70}$
d. P(le domicile s'effondra | the house collapsed) = $\frac{2}{10} \cdot \frac{6}{7} = \frac{12}{70}$

Where different derivations result in the same TL string, their weights are summed and the duplicate strings are removed.

The examples in (33) and (34) are reasonably straightforward if we assume, as here, that the chunks in (35) exist in the system databases shown:

(35) *Marker lexicon*:
<DET> the house : la maison
<DET> the house : le domicile

Word-level lexicon:
<LEX> collapsed : s'écroula
<LEX> collapsed : s'effondra

If the input string were instead *a house collapsed*, and the NP *a house* were absent from the marker lexicon, then a translation could be formed via the chunks in (36):

(36) *Generalized marker lexicon*:
<DET> house : maison
<DET> house : domicile

Word-level lexicon:
<LEX> collapsed : s'écroula
<LEX> collapsed : s'effondra
<DET> a : un
<DET> a : une

Given the aligned segments in (36), the correct translations (37) would be built:

(37) a. Une maison s'écroula.
b. Une maison s'effondra.
c. Un domicile s'écroula.
d. Un domicile s'effondra.

However, in addition, the mistranslations in (38) would be constructed:

(38) a. *Un maison s'écroula.

b. *Un maison s'effondra.
c. *Une domicile s'écroula.
d. *Une domicile s'effondra.

These mistranslations are all caused by boundary friction.

Each of the translations in (37) and (38) would be output with an associated weight and ranked by the system. We would like to incorporate into our model a procedure whereby translation chunks extracted from the phrasal and marker lexicons are more highly regarded than those constructed by inserting words from the word-level lexicon into generalized marker chunks. That is, we want to allocate a larger portion of the probability space to the phrasal and marker lexicons than to the generalized or word-level lexicons. We have yet to import such a constraint into our model, but we plan to do so in the near future using the weighted majority algorithm (Littlestone and Warmuth 1992).

4. Experiments and System Evaluation

We report here on a number of experiments using test sets of 200 sentences and 500 noun phrases. Some typical examples from the two test sets are given in (39):

(39) *Noun phrases*:
- the heavy use of management fees last year
- an increase through issues of new shares and convertible bonds
- a space-based defense shield for official acts by the congressman

Sentences:
- The bright red one interferes with the genes that are responsible for collecting pollen.
- A more recent novel permitted the new basket product.
- The area with the museums and the charities is under something of a cloud.
- Reducing the supply of goods as commissions to middlemen permitted a chaotic sex life.

The test sets were created automatically from words contained in at least one of the systems' knowledge bases, with the proviso that the strings corresponding to the 59 rule types we extracted from the Penn Treebank reflected the frequency bias of these rule types, as far as possible. That is, we wanted to ensure that strings corresponding to a rule type that was (approximately) twice as frequent as some other rule type occurred (approximately) twice as often in the test sets. Finally, we ensured that strings corresponding to all 59 rule types were present in the sentence test set.

The experiments were designed to evaluate the coverage and translation quality of different versions of our EBMT system. We contrast the results obtained when the memories of our system are seeded with source strings and their translations derived from

- each of the three individual on-line MT systems (A, B, and C)
- each pair of on-line MT systems (AB, AC, and BC)
- all three on-line MT systems (ABC).

We also compare and contrast the results obtained when the memories of our system are seeded with source strings and their translations derived using third-person singular, third-person plural, and both third-person singular and third-person plural dummy subjects.

Both sets of experiments are designed to test whether coverage and translation quality improve when more ⟨source, target⟩ fragments are taken into account. With respect to quality, the translations output (for sentences, using chunks derived with third-person plural dummy subjects) were scored according to the following scale by two native speakers of French with excellent English:

- 1: Contains major syntactic errors and is unintelligible
- 2: Contains minor syntactic errors and is intelligible
- 3: Contains no syntactic errors and is intelligible

This scale is used to measure the impact on translation quality, both for sentences and NPs, of using multiple knowledge sources. Although we are primarily interested in the translation of sentences, we use the NP test set to see whether we are in fact translating nominal chunks correctly, and also to investigate whether our methodology can actually correct any NPs mistranslated by the three on-line MT systems. We discuss this further in Section 5.

We also measure the ability of our system to rank the "best" translation (as determined by our human experts) highly in the set of output translations. Statistical MT systems such as *wEBMT* may derive many different translations for a particular input, each of which is output with a confidence weighting. We are keen to ensure that if our system is able to produce high-quality translations, these are ranked as highly as possible: We do not consider it feasible for a human to have to sift through many hundreds or thousands of translations in order to determine the "correct" one.

In a further experiment, we translate the test set of sentences via the three on-line MT systems and test to see whether our system *wEBMT* can improve on these translations. In so doing we calculate the "net gain" of performing example-based MT compared to using on-line MT systems.

Finally, we offer some thoughts on the relative merits of the three on-line MT systems used in our research. Although this was not the primary focus of our research, it turned out that we were able, as a direct consequence of our methodology, to evaluate the on-line MT systems chosen.

4.1 Experiments Using Single Knowledge Sources

Here we report on experiments in which the two test sets are tackled by our system when its memory is seeded with translations obtained by the individual on-line MT systems specified in Section 2.1. A parameter that is altered in the experiment on the sentential test set is the nature of the dummy subjects used to gather the initial translation fragments: third-person singular, third-person plural, and both third-person singular and third-person plural.

4.1.1 Sentences. The test set comprised 200 sentences, with an average sentence length of 8.5 words (minimum 3 words, maximum 18). The input strings were segmented by applying the *n*-gram segmentation approach outlined in Section 3.

Experiment 1: Third-Person Plural Subjects

As far as coverage is concerned, our system *wEBMT* translated 184 (92%) of the sentences using chunks derived from Systems A and C, and using chunks from system B, our system managed to translate 180 sentences (90%). The same 16 sentences were not translated by any of the systems owing to their failure to locate one or more words in the sentence within the word-level lexicon. Recall that despite the fact that all words in the test set were seen by the system in the training phase, only those content (i.e., non-marker) words that occur in bigram marker chunks are inserted into the word-level lexicon (cf. (23)). In cases such as these, in which one or more words cannot be translated by our system, partial translations such as those in (40) are output:

(40) A little girl misplaced a full page $\Longrightarrow$ Une petite fille *misplaced* une pleine page.

That is, although *misplaced* was present in the system's database, it was not present in the correct context. That is, it appeared in the phrasal lexicon, as shown in (41):

(41) were misplaced $\Longrightarrow$ ont été égarés

The form required in (40) is a simple past-tense verb, but *misplaced* appears in (41) only as a passive participle. The word *were* in (41) is not a marker word, so this fragment cannot be broken down any further by our segmentation method.[8] In such cases we output the partial translation with source equivalents for any untranslated words, as shown in (40).

Ninety-six (48%) of the sentences were translated by combining fragments contained in the original phrasal lexicon or the marker lexicon, 56 (28%) of the translations were obtained by locating single content (i.e., non-marker) words in the word-level lexicon and inserting these into the translation at the appropriate position, and 32 (16%) were produced by inserting marker words into generalized templates.

Table 1 shows the results obtained from our human evaluators' ratings of the translations produced by our system when it was populated with fragments derived from one of the individual on-line MT systems. Evaluators rated more than one-third of translations as intelligible and without syntactic errors (score 3), with over 85% of translations deemed intelligible (scores 2 and 3) for all systems. Unintelligble translations (score 1) ranged from 14% for chunks derived from *SDL* to just 4.4% for translations formed from knowledge sources created from *Logomedia*. These initial results provide some evidence in favor of the hypothesis that *Logomedia* might be the best system. Although such evaluation is not a primary focus of our work at the outset, our methodology provides as a spin-off an evaluation of the three MT systems. We discuss this further in Section 4.5.

When the system cannot produce a translation for a particular input, the main reason is an absent word in the word-level lexicon. Adding more lexical entries would improve translation coverage and would also affect translation quality (possibly adversely, in some cases). We plan to measure the impact of a larger lexicon in future work. Low-quality translations are almost invariably caused by inappropriate verb forms in the word-level dictionary: For the experiments carried out in which all verbs

8 Of course, even if it could be, we would be able to derive only the mistranslation *Une petite fille égarés une pleine page*, assuming there to be no other relevant fragments in the system's databases.

Table 1
Translation quality for sentences: Chunks derived from individual on-line MT systems, third-person plural dummy subjects.

System	Score 1	Score 2	Score 3
A	14.2%	51.2%	34.6%
B	8.9%	54.7%	36.4%
C	4.4%	59.1%	36.5%

were third-person plural, any NP with a third-person singular subject in the test set would be accompanied by a third-person plural verb in the translation. A similar effect is seen where the databases of *wEBMT* were seeded with third-person singular verbs, of course. However, we should expect an improvement in translation quality when both sets of verb forms are included in the memories of the system (see Experiment 2).

Table 2 shows where the "best" translation, as defined by a human expert, was ranked among the of translations output by our system. In over 65% of cases, the system itself had ranked the "best" translation first, and the "best" translation was never located outside the top five ranked translations. This is remarkable given that over 2,000 translations are output for certain source sentences.

Experiment 2: Seeding the Databases with More Examples

The results for the previous experiment were obtained when the databases were seeded with third-person plural dummy subjects. We ran two variations on this experiment: (1) we tested the system by seeding its memories with third-person singular dummy subjects, and (2) We tested the system by seeding its memories with both third-person singular and third-person plural dummy subjects.

Figure 1 shows that translation quality improves when the system databases are seeded with more translation pairs. We can see that the system does slightly better when it uses third-person plural chunks compared to when it uses their singular counterparts. When third-person singular dummy subjects are inserted in order to derive the initial translation fragments inserted into our system's memories, the number of translations rated 3 for quality deteriorates by about 5% for systems B and C and by about 3% for system A. Given a larger number of third-person plural NP subjects in our test set, this was to be expected.

However, a considerable improvement in quality can be seen when fragments from

Table 2
Ranking of "best" translation for sentences: Chunks derived from individual on-line MT systems, third-person plural dummy subjects.

System	Ranked 1	Ranked 2–5
A	71.6%	28.4%
B	65.3%	34.7%
C	70.3%	29.7%

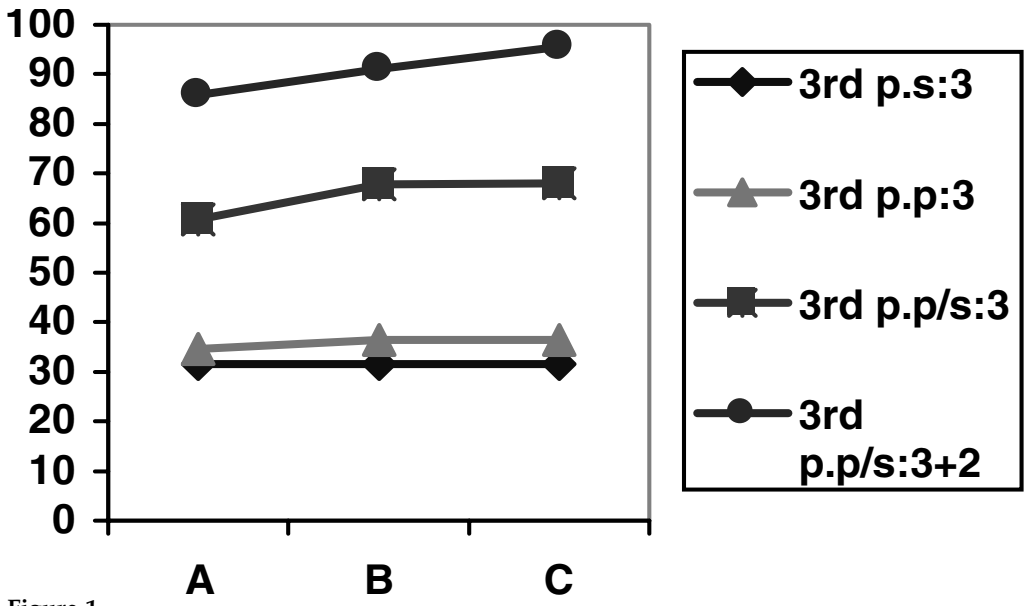

Figure 1
Translation quality improves when system databases are seeded with more translation pairs: Measuring % translation quality using fragments derived from single on-line MT systems.

Table 3
Ranking of "best" translation for sentences: Chunks derived from individual on-line MT systems, third-person singular and third-person plural dummy subjects.

System	Ranked 1	Ranked 2–5	Ranked 6–10	Ranked 10–20
A	65.2%	30.5%	0.0%	4.3%
B	60.8%	34.9%	0.0%	4.3%
C	64.1%	31.6%	0.0%	4.3%

both singular and plural forms are inserted into the system's memories. Translations produced from chunks derived from system A are rated 3 in 66.1% of cases, and this rises to 67.9% for system B and 68% for system C. Regarding intelligibility (scores 2 and 3 for quality), system A scores 85.8%, system B scores 91.1%, and system C scores 95.6%. We consider these to be very reasonable results.

Table 3 summarizes the relative ranking of the "best" translation when more translation pairs are used to seed the system's memories. Now that both third-person singular and third-person plural dummy subjects are provided, we see that the number of "best" translations ranked first deteriorates, by about 6% for systems A and C and by 4.5% for system B. In Table 2, we saw that the "best" translation was in no case ranked lower than fifth, but now that many more translations are output per test set sentence, we sometimes have to search as low as 20th in order to find the "best" translation.

4.1.2 Experiment 3: Noun Phrases. The test set comprised 500 noun phrases, with an average NP length of 6.14 words (minimum 3 words, maximum 12). The noun phrases in the test set also need to be fragmented using our *n*-gram segmentation method, as

it is highly probable that they do not exist *en bloc* in the phrasal lexicon and therefore need to be analyzed using smaller fragments in the system's databases.

We give results for coverage and translation quality in Table 4. These results are for NPs translated via chunks derived from the three individual on-line MT systems. As with the sentence test set, fragments derived from systems A and C achieve the broadest coverage, producing translations for 474 out of the 500 NPs; those obtained from system B enable 463 of the 500 NPs to be translated.

As for quality, *wEBMT* clearly performs best when using translation fragments derived from system C: 47.3% of these translations were awarded a quality score of 3, more than 10% better than for chunks derived from system B. For system C, a total of 452 (96%) of the generated translations were deemed intelligible (scores 2 and 3), that is, 31 (6.6%) more translations than with system B.

On average, about 54% of translations are formed by combining chunks from the phrasal lexicon with those from the marker lexicon, about 9% are produced by inserting marker words into the generalized templates, and about 37% are generated by inserting single non–marker words from the word-level lexicon at the appropriate locations in phrasal chunks. The major reason that translations fail to be produced in 6% of cases is the absence of a relevant generalized template. For example, the unseen input *her negative TV ads* is generalized to (42):

(42) <POSS> negative TV ads

However, the nearest relevant generalized template found in the system's memory is (43):

(43) <DET> negative TV ads

That is, the template in (43) allows the insertion of any determiner, but no other marker word. Deriving translation fragments from more examples would lead to an improvement in coverage. Alternatively, for marker words that appear in the same relative position, such as determiners and possessive pronouns, we could "back off" to a more general marker tag to allow mutual substitution of such words in a subsequent operation to enable translation of examples like these. This remains an area of investigation in future work.

The results in Table 4 further substantiate our findings on the sentence test set, namely, that system C may be the best of the three on-line MT systems used to populate the memories of our EBMT system. We comment further on this in Section 4.5. In addition, these figures provide strong evidence that our system can indeed translate most noun phrases with which it is confronted and with more than reasonable quality.

Table 4
Translation coverage and quality for NPs: Chunks derived from individual on-line MT systems.

System	Coverage	Quality		
		Score 1	Score 2	Score 3
A	94.8%	13.7%	52.5%	33.8%
B	92.6%	10.6%	52.3%	37.1%
C	94.8%	4.0%	48.7%	47.3%

Table 5
Ranking of "best" translation for NPs: Chunks derived from individual on-line MT systems.

System	Ranked 1	Ranked 2	Ranked 3–5	Ranked 6–10
A	64.6%	9.1%	23.6%	2.7%
B	57.7%	15.6%	24.8%	1.9%
C	60.0%	7.6%	29.3%	3.1%

Table 6
Number of translations produced for the NP *a plan for reducing debt over 20 years.*

System(s)	Number of Translations
A	14
B	10
C	5
AB	108
AC	72
BC	42
ABC	224

The results obtained regarding the ranking of the "best" translation appear in Table 5. Our system ranks the "best" translation first over 57% of the time, and in over 96% of cases, it ranks it in the top five, and at worst in the top ten.

4.2 Experiments Using Multiple Knowledge Sources

If the three on-line MT systems translate the phrases extracted from the Penn-II Treebank in different ways, then combining systems to obtain results for AB, AC, BC, and ABC always involves an increase in the number of translations produced, both for sentences and noun phrases. That is, if an input string receives a translation via chunks derived from the individual on-line systems, when chunks are combined from different systems, more translations will be output for that input string.

As an example, the number of translations produced by each system for the NP *a plan for reducing debt over 20 years* is shown in Table 6. Whereas the greatest number of translations for this NP produced from chunks from any individual on-line system is 14, when translation fragments from all three systems are merged (ABC), 224 translations are produced. Combining systems in this way means that all possible combinations of chunks from the systems are produced: That is, the number of translations generated via AB is much larger than those derived from either A or B, as now chunks from A and B may be combined to produce new translations that could not be generated from the individual knowledge sources. As a further example, consider the translation of the NP *the total at risk a year*. When fragments from systems A and B are combined, the "best" translation is comprised of the chunk combination AAB, that is, the three-chunk combination in (44), with the first two chunks obtained from system A, and the last from system B:

(44) $[_A$*the total*], $[_A$*at risk*], $[_B$*a year*]

That is, the translation of this NP improves when the performance of system AB is

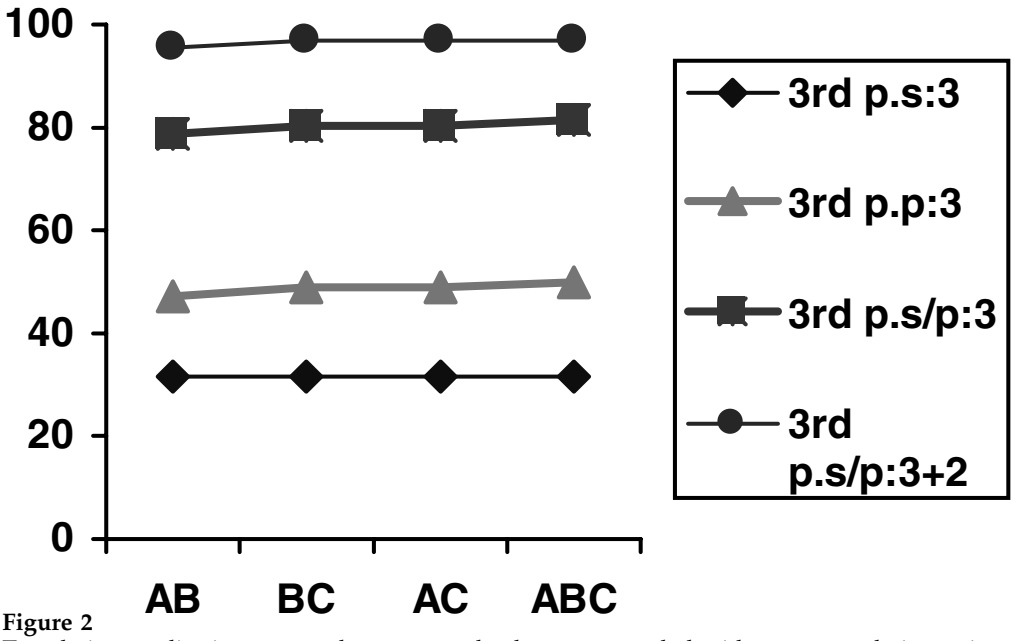

Figure 2
Translation quality improves when system databases are seeded with more translation pairs and when more knowledge sources are used: Measuring % translation quality using fragments derived from combinations of on-line MT systems.

evaluated: Of course, if we consider (say) three-chunk combinations from either system A or B, the only possibilities are AAA or BBB, respectively.

However, the number of translations produced by the system is less significant than their quality. The ranking process outlined in Section 3 classifies the translations produced with regard to their position as the "best" translation. In the sections below, we also discuss the issue of quality and show that it improves when more translation fragments are taken into account. Furthermore, we show below that despite generating more translations per input string, *wEBMT* still ranks the "best" translation in the top 1% of all output translation candidates.

4.2.1 Experiment 4: Translating Sentences by Combining Fragments from Different Systems. We saw in Experiment 1 that 16 strings in the test set were left untranslated by systems A, B, and C individually. When knowledge sources are combined, these 16 strings remain untranslated. However, as Figure 2 shows, the translation quality improves significantly. The best individual system performance was 36.5% scoring 3. This rises to a best performance of 48.9% among pairs of systems combined and improves still further to 50% when chunks from all three knowledge sources are combined.

Table 7 provides results regarding the relative location of the "best" translation for sentences. For all system combinations, the "best" translation is to be found among the top 10 ranked translations in all permutations of combinations of chunks, with at least 54% ranked first. Despite a corresponding rise in the number of translations produced per input sentence when all three knowledge sources are combined (ABC), in over 97% of cases, the "best" translation continues to be found in the top five output candidates.

Table 7
Ranking of "best" translation for sentences: Chunks derived from combinations of on-line MT systems, third-person plural dummy subjects.

System	Ranked 1	Ranked 2–5	Ranked 6–10
AB	67.6%	31.1%	1.3%
AC	54.0%	46.0%	0.0%
BC	63.6%	35.1%	1.3%
ABC	62.2%	35.1%	2.7%

4.2.2 Experiment 5: Combining Fragments from Different Systems and Seeding the Databases with More Examples. Figure 2 demonstrates that considerable improvements in translation quality are achieved when the memory of *wEBMT* is seeded with both third-person singular and third-person plural fragments. For the pairwise combinations, 78.7% of the translations derived from AB are rated 3 for quality, compared to 80.4% of those derived from AC and BC. The results for ABC improve again, to 81.5%. Regarding intelligibility (scores 2 and 3 for quality), we can see from Figure 2 that near perfect results are obtained: AB scores 95.6%, and all other combinations score 96.7%.

Table 8 shows the ranking of the "best" translation when multiple knowledge sources are employed and both third-person singular and third-person plural dummy subjects are used to populate the system's memories. The number of instances in which the "best" translation is ranked first by *wEBMT* deteriorates: by 24% for AB, by 15% for AC, by 20% for BC, and by 27% for ABC. For all system combinations, Table 7 shows that the "best" translation was ranked no lower than 10th; for the system combinations in Table 8, sometimes the "correct" translation is ranked as low as 36th. As expected, the worst ranking results are for system combination ABC, in which all system chunks are combined for both third-person singular and third-person plural dummy subjects. However, even here the "best" translation is ranked in the top five in over 63% of cases, and 72.6% of the time it is located among the top 10 ranked translations. For this system configuration, the lowest we have to look to find the "best" translation is 36th. For that particular sentence (i.e., the one for which the "best" translation is ranked 36), over 4,000 possible translations are generated, so even here the "best" translation remains in the top 1% of translation candidates.

Table 8
Ranking of "best" translation for sentences: Chunks derived from combinations of on-line MT systems, third-person singular and third-person plural dummy subjects.

System	Ranked 1	Ranked 2–5	Ranked 6–10	Ranked 10–20	Ranked 20–40
AB	43.4%	32.6%	2.3%	13.0%	8.7%
AC	39.1%	34.8%	5.4%	12.0%	8.7%
BC	43.4%	31.7%	2.7%	13.5%	8.7%
ABC	35.2%	28.0%	9.4%	18.3%	9.1%

Table 9
NPs: Coverage and quality improve when fragments from different sources are included.

Knowledge Source Combination	Coverage Percentage	Quality Score 3
A	94.8	33.8
B	92.6	37.1
C	94.8	47.3
AB	95.4	54.1
BC	95.6	64.0
AC	94.8	72.0
ABC	96.0	77.8

4.2.3 Experiment 6: Translating Noun Phrases by Combining Fragments from Different Systems.

As we did with sentences, we seeded our EBMT system with fragments derived from the three different on-line MT systems and confronted it with the NP test set. Table 9 clearly shows that as more knowledge sources are added, translation quality improves considerably. The worst-performing individual system scores 3 for quality in just over a third of cases, but when all system chunks are combined, this rises to 77.8%. Note also that, unlike with sentences, we see an increase in coverage when more knowledge sources are used, from a low of 92.6% for system B to a high of 96% when all chunks are combined. Many more improvements are seen when our post hoc validation and correction methodology, described in Section 5, is used, but the merging of fragments derived from different on-line systems also leads to an improvement in translation quality. Consider the examples in (45):

(45) *Input*: an old story in common
System B: une vieille histoire dans commun
System Combination BC: une vieille histoire en commun

That is, the optional PP *in common* was mistranslated by system B as *dans commun*, but when knowledge from system C is added to that of system B, the improved translation *en commun* is generated.

We saw in Section 4.1.2 that when translating the NP test set, the "best" translation, as adjudged by our human evaluators, was to be found no lower than tenth of all translations output by our system. When knowledge sources are combined, it is

Table 10
Ranking of "best" translation for NPs: Chunks derived from more that one on-line MT system.

System	Ranked 1	Ranked 2	Ranked 3–5	Ranked 6–10
AB	42.2%	13.8%	41.3%	2.7%
AC	62.1%	14.1%	21.3%	2.5%
BC	66.4%	11.4%	19.8%	2.4%
ABC	62.0%	17.5%	13.5%	7.0%

important to measure whether the "best" translation is still highly ranked. The results of such an assessment are summarized in Table 10. When the "best" translation is ranked first by the system, we see that the optimal combination of knowledge sources is the pair BC, with 66.4%. This is an interesting result given that in Table 5, only 57.7% of the NPs translated by system B were ranked first, with 60% of those produced by system C ranked in first place. That is, we see a 6.4% improvement (31 NPs) when fragments from systems B and C are combined. All other combinations cause the number of "best" translations ranked first to deteriorate, as may be expected. When the "best" translation is ranked either first or second by the system, the best combination of fragments is that from system ABC, with 79.5% (376 NPs).

Importantly, for all combinations, the "best" translation remains among the top 10 ranked translations. This is encouraging, as any translator using our system needs only to examine a small subset of the translations produced to find the "best" one. Indeed, given the various results shown, we are confident that we could prune the number of translations generated for presentation to the translator for selection of the "best" one: For the most part, this is the top ten translations for both NPs and sentences, but in the worst case, we need present no more than the top 1% of the candidate translations.

4.3 Summary

The results presented in this section show that seeding the *wEBMT* system's databases with more fragments improves both coverage and translation quality. We do this additional seeding in two ways: (1) by combining fragments derived from different on-line MT systems, and (2) by obtaining translations using both third-person singular and third-person plural dummy subjects. The best combination of these parameters is to use chunks derived from *Logomedia* with third-person plural dummy subjects provided. Nevertheless, despite the fact that *Logomedia* appears to be the best on-line MT system, adding chunks from the other two on-line MT systems improves coverage and translation quality. In sum, therefore, the best results are obtained when chunks from all three on-line systems (combination ABC) are used and the *wEBMT* system's databases are seeded with translations from these systems for both third-person singular and third-person plural versions of sentences.

The disadvantage of using more knowledge sources, of course, is that many more candidate translations are generated, which sometimes causes the "best" translation to appear lower in the ranked order of output translations. Nevertheless, the "best" translation is almost always to be found in the top 10 translations produced by *wEBMT* and always in the top 1% of the candidate translations.

4.4 Relative Gain of EBMT

In order to try to calculate the relative gain of EBMT, we translated all 200 strings in the sentence test set via the three on-line MT systems used elsewhere in our experiments. Of course, the main advantage of using such Web-based systems is that they are extremely robust: no matter what they are confronted with, they will always produce some translation. With respect to coverage, therefore, the on-line systems currently win out over *wEBMT*: the size of the lexicons available to the on-line systems means that they will generate translations with all source words translated more often than our system will. Nevertheless, the fact that our system outputs partial translations in situations in which it encounters a word that it cannot translate demonstrates a certain level of robustness in our system.

Where quality is concerned, however, *wEBMT* can improve on the translations produced by the three on-line MT systems. We provided our human evaluators with the source sentences from the test set, together with the translation generated by our

wEBMT system using combinations of chunks derived from all three on-line systems (ABC), and the translation obtained directly from the on-line systems themselves. The pairs of translations (*wEBMT* and on-line MT system) were presented in a random order. The evaluators were simply asked to state, for all three on-line MT systems, which of the pair of translations they preferred: that from the on-line MT system or that from *wEBMT*.

Translations produced by *wEBMT* using chunks derived from all three systems were preferred to those from system A in 30/184 cases (16.3%) (we ignored the 16 cases for which our system could produce only partial translations); for system B, our system's translations were preferred in 8/180 cases (4.4%); and for system C, our system was judged as producing better translations in 6/184 cases (3.3%). Some examples in which our system offered improvements over the translations provided by the on-line MT systems are shown in (46):

(46) *Input*: Her short term interest rates link the issues.
System A: Son lien à court terme de taux d'interêt les questions.
wEBMT ABC: Ses taux d'interêt à court terme lient les questions.

Input: The researchers air the shows.
System B: L'air de chercheurs les expositions.
wEBMT ABC: Les chercheurs aérent les expositions.

Input: A group hire lawyers to provide information about clients.
System C: Un avocats de la location du groupe fournir de l'informations au sujet de clients.
wEBMT ABC: Un groupe embauche des avocats à fournir de l'informations au sujet de clients.

Regarding the first two translation pairs in (46), *wEBMT* has provided a finite verb where the on-line MT systems have none. In addition, the translation of the subject NPs is much improved. As for the final translation pair in (46), whereas *Logomedia* managed to retrieve the correct translation of *provide*, this was not the main verb in the English input string. *wEBMT*, on the other hand, does translate *hire* correctly as a verb rather than a noun.

We consider three ways in which the net gain of EBMT may be calculated. First, we assume it is equal to the number of translations produced by *wEBMT* that are preferred by the human evaluator, minus those derived by the on-line MT systems that are preferred, divided by the total number of translations. In fact, where both *wEBMT* and the on-line systems produce a translation, those derived via *wEBMT* are *always* preferred. Those translations produced by the on-line systems that are preferred are those in which *wEBMT* was unable to generate a complete translation. This is quite a harsh measure: As can be seen from the translations in (46), although the words in the translations produced by the on-line systems are all French (but recall (29)–(30), in which this was not the case), the translations themselves are poor. In some cases, despite the fact that the translations derived via *wEBMT* may contain an untranslated English word, the accompanying partial translations may in fact be deemed superior to the "complete" translations derived via the on-line systems.

Nevertheless, assuming that the on-line systems win out in these situations, the net gain compared to system A is 14/200 (7%), whereas for systems B and C, we see in effect a net loss: −12/200 (−6%) for system B, and −10/200 (−5%) for system C. If we can obtain complete translations in those cases in which we currently encounter

an untranslatable word, we are confident that we can convert these net losses into net gains. With respect to system A, we can assume that our net gain would increase further.

However, we provide two other interpretations of the net gain of EBMT, calculated using the formula in (47):

(47) *Net Gain = Coverage Percentage + K*(*Translation Quality*)

The term Translation Quality in (47) refers to the number of translations preferred by the human evaluator, excluding cases in which one system failed to produce a translation, which is already factored into the equation under the term *Coverage Percentage*. Where *K*=1, we view coverage and translation quality as equally important. If we consider quality to be more important, we can increase *K*. We provide results for *K*=1 and *K*=2 in (48):

(48) Where *K*=1:
$Net\ Gain_{MT} = 100$

$Net\ Gain_{EBMT} = 92 + 30 = 122$ (compared to system A)
$Net\ Gain_{EBMT} = 92 + 8 = 100$ (compared to system B)
$Net\ Gain_{EBMT} = 92 + 6 = 98$ (compared to system C)

Where *K*=2:
$Net\ Gain_{MT} = 100$

$Net\ Gain_{EBMT} = 92 + 60 = 152$ (compared to system A)
$Net\ Gain_{EBMT} = 92 + 16 = 108$ (compared to system B)
$Net\ Gain_{EBMT} = 92 + 12 = 104$ (compared to system C)

That is, where *K*=1, *wEBMT* outperforms *SDL* by a factor of 22, whereas there is no gain with respect to *Reverso*, and a slight loss compared to *Logomedia*. However, *wEBMT* outperforms all three on-line systems when translation quality is viewed as twice as important as coverage. This is a reasonable view, we feel, and our system shows a net gain against all three on-line systems in this context. Although the coverage obtained with the on-line systems is better, the improved translation quality obtained with *wEBMT* ensures a net gain.

We expect to obtain more insightful results regarding the relative gain of EBMT over on-line MT systems when automatic evaluation metrics (such as IBM's Bleu, or dynamic programming, or sentence- and word-error rates) have been obtained. This is a priority in future work.

4.5 Evaluating Individual On-line MT Systems

The previous sections detail the results obtained when translation fragments derived from the three individual on-line MT systems are used, together with various combinations of knowledge sources. We provided results both for coverage and translation quality. As we noted above, we were able, as a consequence of our chosen methodology, to evaluate the on-line MT systems used.

Where sentences are concerned, we saw that coverage was approximately the same for each individual system, and that combinations of multiple knowledge sources did

not improve coverage. For NPs, however, coverage improved when more fragments were considered: Whereas 474 NPs could be translated by both systems A (*SDL*) and C (*Logomedia*), this number rose to 480 when all three knowledge sources were combined.

With respect to translation quality, whereas *Logomedia* and *Reverso* could hardly be distinguished when it came to numbers of translations of sentences adjudged as intelligible and syntactically correct, if we consider those translations considered unintelligible by our human evaluators, about twice as many translations produced by chunks derived from *Reverso* were unintelligible compared to those produced by *Logomedia*. This would indicate that *Logomedia* may be better.

When fragments from combinations of systems are considered, we note that for sentences, no improvement in coverage results, but quite significant improvements in quality are seen. System C slightly outperformed system B in the individual face-off, and we see that combinations that utilize chunks from system C outperform those that do not: AC and BC both score 3 for quality in 80.4% of cases, compared to AB's 78.7%, and ABC improves still further, to 81.5%. When the databases of *wEBMT* are seeded with more chunks (using both singular and plural dummy subjects), system C continues to outperform the other two systems.

When NPs are considered, system C considerably outperforms the other two systems, obtaining a score of 3 for quality in 10.2% more cases than its nearest challenger, system B. When chunks from different systems are combined, again we see that combinations with chunks derived from system C outperform those that omit them: BC and AC improve over AB by 10% and 18%, respectively, and ABC shows a further increase.

Finally, in Section 4.4, we discussed the relative gain of using *wEBMT* over the three on-line MT systems. Further evidence that *Logomedia* may be the best of the three systems is provided by the fact that the relative gain compared to *Logomedia* was much lower than with the other two systems.

It is worth considering why some combinations seem to work better than others. For NPs, in cases where system A fails to produce the "best" translation, this is often due to incorrect word order. For example, in the NP *cellular mobile lines for the workmen*, system A produces the translation *cellulaire mobile lignes pour les ouvriers*.[9] It is also the case that when system A fails to retrieve a translation for a particular chunk, it simply proposes the English for that chunk as its translation. This is useful to some extent, in that a default translation is produced (cf. the similar approach that we have taken with respect to (40), for instance). However, in all of these cases this utility is lessened considerably given that either system B or system C produces a better translation. System B has the added advantage that it sometimes provides an alternative translation in brackets. If no translation is available, the English is output. System C often produces a correct translation of a verb where the other systems are lacking. It is probably because of this aspect of translation that system C's translations are preferred over those of the other two systems.

5. Validation and Correction of Translations via the Web

A translation can be formed in our system only when the recombination of chunks causes the input string to be matched exactly. Therefore, if all chunks cannot be retrieved, then no complete translation can be produced (cf. (40) and resultant discus-

9 Such a translation would be a candidate for post hoc validation via the Web (cf. Section 5), but the correct translation *lignes cellulaires mobiles pour les ouvriers* is produced in any case by system C, rendering this unnecessary.

sion). We have shown that when a translation cannot be produced by combining phrasal chunks, translations can be formed by the insertion of single marker words into generalized templates. This can be compared to the idea of "hooks" (Somers, McLean, and Jones 1994), where some context in which fragments have occurred is maintained in the translation templates. The hooks indicate which words and POS tags can occur in the immediate left and right context of a fragment, together with a weight that reflects how often this context is found in a corpus. The "best" translation, therefore, is simply that which is output by the system with the highest score.

Consider the translation of the NP *the personal computers*. There are three possible ways in which this may be segmented using the marker hypothesis, namely, the chunks in (49):

(49) *Phrasal lexicon*: the personal computers
Marker lexicon: <DET> the personal computers
Generalized lexicon: <DET> personal computers

In our system, the only chunk retrieved is the generalized chunk in (49). The system stores a list of marker words and their translations in the word-level lexicon. A weight derived via the method in (32) is attached to each translation. The system searches for marker words within the string and retrieves their translations.[10] In this case, the marker word in the string is *the* and its translation can be one of *le, la, l',* or *les,* depending on the context. The system simply attaches the translation with the highest weight to the existing chunk *ordinateurs personnels* to produce the mistranslation in (50):

(50) *la ordinateurs personnels

The problem of boundary friction is clearly visible here: We have inserted a feminine singular determiner into a chunk that was generalized from a masculine plural NP.

However, rather than output this wrong translation directly, we use a post hoc validation and (if required) correction process based on Grefenstette (1999). Grefenstette shows that the Web can be used as a filter on translation quality simply by searching for competing translation candidates and selecting the one that is found most often. Rather than search for competing candidates, we select the "best" translation and have its morphological variants searched for on-line. In the example above, namely, *the personal computers*, we search for *les ordinateurs personnels* versus the wrong alternatives *le/la/l'ordinateurs personnels*. Interestingly, using Lycos, and setting the search language to French, the correct form *les ordinateurs personnels* is uniquely preferred over the other alternatives, as it is found 2,454 times, whereas the others are not found at all. In this case, this translation overrides the highest-ranked translation (50) and is output as the final translation. In fact, in checking the translations obtained for NPs using system combination ABC, we noted that 251 NPs out of the test set of 500 could be improved. Of these 251, 207 (82.5%) were improved post hoc via the Web, with no improvement for the remaining 43 cases. We consider this to be quite a significant result.

In addition to determiner-noun agreement, we use this methodology to check for agreement between the head noun in the subject NP with the head verb in the main

10 Although this is not relevant for the example discussed here, if non–marker words remain untranslated yet exist in the word-level marker lexicon, these too would be inserted at this stage.

Table 11
Validating translations using *AltaVista*

n-Gram Searched For	Number of Web Occurrences
empire sont	353
sont au	91,197
empire est	1,809
est au	217,820

clause VP. We extracted a list of all verbs in the Penn-II Treebank and obtained translations for all verb forms using the three on-line MT systems by inserting appropriate third-person dummy subjects. We use the list of translated verbs to attempt to find the main verb and identify the head noun as the rightmost non–marker word or the rightmost word before any other marker word in a nominal chunk. Having extracted the noun and the verb from the mistranslation in this way, we then search for this bigram on the Web and correct the verb if its morphological variant (third-person singular or third-person plural form) is found more often than in the translation obtained by our system.

To exemplify this procedure with a sentence from the test set, *his empire is beyond the reach of the president*, system A produces the translations in (51):

(51) a. *son empire sont au delà de la portée du président (ranked first, with probability 0.614)
b. son empire est au delà de la portée du président (ranked fifth, with probability 0.028)

This shows that a correct translation such as (51b) may be ranked lower than an incorrect variant (51a) with considerably less probability. The higher ranking is accounted for because the pair ⟨*is beyond the reach of the president, sont au delà de la portée du président*⟩ is contained in the phrasal lexicon, whereas the pair ⟨*is beyond the reach of the president, est au delà de la portée du président*⟩ does not appear.

Prior to outputting translations such as (51a), we search for the relevant *n*-grams via the Web. For this example, using AltaVista, we obtained the results in Table 11. With the counts for all other bigrams for the two translation candidates in (51) being exactly the same, in order to evaluate which of these proposed target strings is the "better" translation, one can simply add the occurrences found on the Web for all four different bigrams and report their relative probabilities. This gives us the probabilities in (52):

(52) a. #empire sont + #sont au = 91,550/311,179 = 0.294
b. #empire est + #est au = 219,629/311,179 = 0.706

That is, the string *empire est au* is about 2.4 times more likely than the string *empire sont au*. However, the count for the second bigram in each example in (52) can of course be discounted, as the juxtaposition of *sont* or *est* with *au* bears no relevance to the correctness or otherwise of the translations in (51). The amended probabilities are, therefore, those in (53):

(53) a. #empire sont = 353/2,162 = 0.163
b. #empire est = 1,809/2,162 = 0.837

Table 12
Using the Web to improve noun-verb agreement

Improvement	No Improvement	N-V Confusion	Not on Web
System A: *Enterprise Translation Server*			
58.6%	3.4%	17.3%	20.7%
System B: *Reverso*			
62%	3.4%	17.3%	20.7%
System C: *Logomedia*			
76%	3.4%	17.2%	3.4%

These figures accurately reflect the likelihood of the translations in (51). Given that *P*(empire est) is about five times higher than *P*(empire sont), translation (51a) is rejected in favor of (51b).

Having given examples of how post hoc validation works within both NPs and sentences, we summarize in Table 12 the results obtained when we tested the 58 sentences whose translations contained subject-verb agreement errors owing to boundary friction. Improvements were seen for translations derived from each of the on-line MT systems: from a minimum of 58.6% (34 translations) for system A to a maximum of 76% (44 translations) for system C. For each system, no improvement was found for two translations, and for 10 translations, our methodology could not tell definitively whether the word to be corrected was a noun or a verb, so no change was made. Finally, in a small number of cases (between 2 and 12 translations), the target string was not found on the Web, so again, no change was made.

In order to be as accurate and relevant as possible, statistical language (and translation) models should be derived from corpora that are as large as possible, representative, and of high quality. Although the Web is large, this post hoc validation process shows that despite the inherent noise contained on the Web because of the heterogeneous nature of the documents contained therein, it remains a resource that is of great use in evaluating translation candidates.

6. Conclusions and Further Work

We have presented an EBMT system based on the marker hypothesis that uses post hoc validation and correction via the Web.[11] Over 218,000 NPs and VPs were extracted automatically from the Penn-II Treebank using just 59 of its 29,000 rule types. These phrases were then translated automatically by three on-line MT systems. These translations gave rise to a number of automatically constructed linguistic resources: (1) the original ⟨source,target⟩ phrasal translation pairs, (2) the marker lexicon, (3) the gen-

11 Thanks are due to one of the anonymous reviewers for pointing out that our *wEBMT* system, seeded with input from multiple translation systems, with a postvalidation process via the Web (amounting to an *n*-gram target language model), in effect forms a multiengine MT system as described by Frederking and Nirenburg (1994), Frederking et al. (1994), and Hogan and Frederking (1998).

eralized lexicon, and (4) the word-level lexicon. When the system is confronted with new input, these knowledge sources are searched in turn for matching chunks, and the target language chunks are combined to create translation candidates.

We presented a number of experiments that showed how the system fared when confronted with NPs and sentences. For the test set of 500 NPs, we obtained translations in 96% of cases, with 77.8% of the 500 NPs being translated correctly. For sentences, we obtained translations in 92% of cases, with a completely correct translation obtained 81.5% of the time. Translation quality improved both when chunks from different on-line systems were used and when the system's memories were seeded with both third-person singular and third-person plural forms. For both NPs and sentences, we obtained intelligible translations in over 96% of cases. In most cases, the "best" translation was ranked in the top 10 translations output by the system and was always ranked in the top 1% of translation candidates. This facilitates the task of any translator interacting with our system who needs to search for the "best" translation among the alternatives provided.

We calculated the net gain of using *wEBMT* compared to the three on-line MT systems. In some cases, an improvement of 50% was seen when EBMT was used. As a consequence of the methodology chosen, we were able to perform a detailed evaluation of the strengths and weaknesses of the three Web-based systems used in our research, with *Logomedia* clearly outranking the other systems used. Nevertheless, adding chunks from the other two on-line MT systems improves both coverage and translation quality. In sum, therefore, the best results are obtained when chunks from all three on-line systems are used, and the system's databases are seeded with translations from these systems for both third-person singular and third-person plural versions of sentences.

In addition, prior to the system's outputting the best-ranked translation candidate, morphological variants of certain components in the translation are searched for via the Web in order to confirm it as the final output translation or to propose a corrected alternative. Currently we validate our translations only with regard to subject head noun–head verb agreement and determiner-noun agreement, but we plan to extend this validation to cover more cases of boundary friction. We demonstrated that considerable improvements can be made to the translations derived by the system by submitting them to the Web for validation and correction.

A number of issues for further work present themselves. The decision to take only those Penn-II rules occurring 1,000 or more times was completely arbitrary, and it might be useful to include some strings corresponding to the less frequently occurring structures in our database. Similarly, it would be a good idea to extend our word-level lexicon by including more entries using rules in which the right-hand side contains a single non-terminal.

Furthermore, the quality of the output was not taken into consideration when selecting the on-line MT systems from which all our system resources are derived, so that any results obtained may be further improved by selecting a "better" MT system that permits batch processing.

We could expect a significant improvement in the results obtained if we were to import the original sentences and their translations into a sentential database. Although we insert dummy subject pronouns to derive appropriate finite verb forms, we do not maintain these translation pairs as a resource for subsequent consultation and retrieval. Although the chance of finding an exact match at sentential level is very low, it will increase as more sentence pairs are added to the database, especially if we restrict the domain of applicability of (a version of) our system to a particular sublanguage area. However, the major improvement that can be expected is in the segmentation process:

Given that verbs are not a closed class, any verb will be contained within (part of) a chunk pertaining to its subject NP. That is, although subject-verb agreement poses a considerable problem to our system given the choice of original input material, this particular instance of boundary friction will disappear if we segment our translation pairs at the sentential rather than at the phrasal level.

In addition, we want to evaluate our system further with respect to larger data sets. Manual evaluation is costly, both in terms of time and effort required. Accordingly, in future work we plan to use automatic evaluation methodologies such as sentence error rate or word error rate. These are very harsh metrics: Consider the example in (54), extracted from the Canadian Hansards:

(54) Again this was voted down by the Liberal majority $\Longrightarrow$
Malheureusement, encore une fois, la majorité libérale l'a rejeté.

Automatic evaluation measures presuppose the existence of an "oracle" (i.e., "correct") translation produced by a human, such as here. Translations derived by the MT system to be evaluated are then compared against the human translation. In the example in (54), the human has inserted *malheureusement* although there is no sign of *unfortunately* in the English source. If the perfect translation *Encore une fois, la majorité libérale l'a rejeté* were produced by an MT system, therefore, it would be penalized, as the human translation is always considered to be the "correct" translation. We have obtained a number of translation memories from two major computer companies, as well as a large amount of monolingual data from the same domain, with which we plan to test our system using automatic evaluation metrics in future work. This will also enable us to test our EBMT methodology against other language pairs, which may present the segmentation method employed with new challenges to overcome.

Finally, we plan to prioritize the lexical resources produced so that more weighting would be given to translations derived from the phrasal and marker lexicons as opposed to those derived via word insertion from the word level lexicon and the generalized templates.

In sum, we have demonstrated that using a "linguistics-lite" approach based on the marker hypothesis, with a large number of phrases extracted automatically from a very small number of the rules in the Penn Treebank, many new reusable linguistic resources can be derived automatically that can be utilized in an EBMT system capable of translating new input with quite reasonable rates of success. We have demonstrated that a net gain may be achieved by using EBMT over on-line MT systems. We have also shown that the Web can be used to validate and correct candidate translations prior to their being output.

Acknowledgments

The authors wish to thank Mary Hearne for helpful input in the initial stages of this project. In addition, the insightful comments provided by four anonymous reviewers helped improve this article considerably. All remaining errors are our own.

References

Ahrenberg, Lars, Mikael Andersson, and Magnus Merkel. 2002. A system for incremental and interactive word linking. In *Proceedings of the Third International Conference on Language Resources and Evaluation* (LREC), pages 485–490, Las Palmas, Canary Islands, Spain.

Becker, Joseph. 1975. The phrasal lexicon. In *Proceedings of the International Workshop on Theoretical Issues in Natural Language Processing*, pages 70–73, Cambridge, MA.

Block, Hans-Ulrich. 2000. Example-based incremental synchronous interpretation. In Wolfgang Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translation*, Springer Verlag, Berlin/Heidelberg/New

York, pages 411–417.
Bod, Rens, Remko Scha, and Khalil Sima'an, editors. 2003. *Data-Oriented Parsing*. CSLI Publications, Stanford, CA.
Boutsis, Sotiris, and Stelios Piperidis. 1998. Aligning clauses in parallel texts. In *Proceedings of the Third Conference on Empirical Methods in Natural Language Processing*, pages 17–26, Granada, Spain.
Brown, Ralf. 2000. Automated generalization of translation examples. In *Eighteenth International Conference on Computational Linguistics: COLING 2000 in Europe*, pages 125–131, Saarbrücken, Germany.
Brown, Ralf. 2003. Clustered transfer-rule induction for example-based translation. In Michael Carl and Andy Way, editors, *Recent Advances in Example-Based Machine Translation*. Kluwer Academic, Dordrecht, the Netherlands, pages 287–305.
Carl, Michael. 1999. Inducing translation templates for example-based machine translation. In *Machine Translation Summit VII*, pages 250–258, Singapore.
Carl, Michael, and Andy Way, editors. 2003. *Recent Advances in Example-Based Machine Translation*. Kluwer Academic, Dordrecht, the Netherlands.
Carl, Michael, Andy Way, and Reinhard Schäler. 2002. Toward a hybrid integrated translation environment. In Stephen Richardson, editor, *Machine Translation: From Research to Real Users: Fifth Conference of the Association for Machine Translation in the Americas (AMTA-2002)*. Lecture Notes in Artificial Intelligence 2499. Springer Verlag, Berlin/Heidelberg, pages 11–20.
Cicekli, Ilyas, and Altay Güvenir. 1996. Learning translation rules from a bilingual corpus. In *Proceedings of the Second International Conference on New Methods in Language Processing*, pages 90–97, Ankara, Turkey.
Frederking, Robert, and Sergei Nirenburg. 1994. Three heads are better than one. In *Proceedings of the Fourth Conference on Applied Natural Language Processing (ANLP-94)*, pages 95–100, Stuttgart, Germany.
Frederking, Robert, Sergei Nirenburg, David Farwell, Steven Helmreich, Eduard Hovy, Kevin Knight, Stephen Beale, Constantin Domashnev, Donna Attardo, Dean Grannes, and Ralf Brown. 1994. Integrating translations from multiple sources with the Pangloss Mark III machine translation system. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, pages 73–80, Columbia, MD.
Fung, Pascale, and Kathleen McKeown. 1997. Finding terminology translations from non-parallel corpora. In *Proceedings of the Fifth Annual Workshop on Very Large Corpora*, pages 192–202, Hong Kong.
Gough, Nano, Andy Way, and Mary Hearne. 2002. Example-based machine translation via the Web. In Stephen Richardson, editor, *Machine Translation: From Research to Real Users: Fifth Conference of the Association for Machine Translation in the Americas (AMTA-2002)* Lecture Notes in Artificial Intelligence 2499. Springer Verlag, Berlin/Heidelberg, pages 74–83.
Green, Thomas. 1979. The necessity of syntax markers: Two experiments with artificial languages. *Journal of Verbal Learning and Behavior*, 18:481–496.
Grefenstette, Gregory. 1999. The World Wide Web as a resource for example-based machine translation tasks. In *Proceedings of the ASLIB Conference on Translating and the Computer,* volume 21, London.
Hogan, Christopher, and Robert E. Frederking. 1998. An evaluation of the multi-engine MT architecture. In *Machine Translation and the Information Soup: Proceedings of the Third Conference of the Association for Machine Translation in the Americas (AMTA '98)*, Lecture Notes in Artificial Intelligence 1529. Springer Verlag, Berlin/Heidelberg, pages 113–123.
Hovy, Edward. 1988. Generating language with a phrasal lexicon. In David McDonald and Leonard Bolc, editors, *Natural Language Generation Systems*. Springer Verlag, New York, pages 353–384.
Juola, Patrick. 1994. A psycholinguistic approach to corpus-based machine translation. In *CSNLP 1994: Third International Conference on the Cognitive Science of Natural Language Processing*, Dublin.
Juola, Patrick. 1997. Corpus-based acquisition of transfer functions using psycholinguistic principles. In Daniel Jones and Harold Somers, editors, *New Methods in Language Processing*. UCL Press, London, pages 207–218.
Juola, Patrick. 1998. On psycholinguistic grammars. *Grammars*, 1(1):15–31.
Kaji, Hiroyuki, Takuya Kida, and Yuji Morimoto. 1992. Learning translation templates from bilingual text. In *Proceedings of the 15th [sic] International Conference on Computational Linguistics* (COLING), pages 672–678, Nantes, France.
Kay, Martin, and Martin Röscheisen. 1993. Text-translation alignment. *Computational*

Linguistics, 19(1):121–142.
Littlestone, Nick, and Manfred Warmuth. 1992. The weighted majority algorithm. Technical Report UCSC-CRL-91-28, University of California, Santa Cruz.
Macklovitch, Elliott. 2000. Two types of translation memory. In *Proceedings of the ASLIB Conference on Translating and the Computer,* volume 22, London.
Macklovitch, Elliott, and Graham Russell. 2000. What's been forgotten in translation memory. In *Envisioning Machine Translation in the Information Future: Proceedings of Fourth Conference of the Association for Machine Translation in the Americas (AMTA-2000)*, pages 137–146, Cuernavaca, Mexico.
McTait, Kevin, and Arturo Trujillo. 1999. A language-neutral sparse-data algorithm for extracting translation patterns. In *Proceedings of the Eighth International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 98–108, Chester, England.
Milosavljevic, Maria, Adrian Tulloch, and Robert Dale. 1996. Text generation in a dynamic hypertext environment. In *Proceedings of the 19th Australasian Computer Science Conference*, pages 417–426, Melbourne, Australia.
Morgan, James, Richard Meier, and Elissa Newport. 1989. Facilitating the acquisition of syntax with cross-sentential cues to phrase structure. *Journal of Memory and Language,* 28:360–374.
Mori, Kazuo, and Shannon Moeser. 1983. The role of syntax markers and semantic referents in learning an artificial language. *Journal of Verbal Learning and Verbal Behavior,* 22:701–718.
Nagao, Makoto. 1984. A framework of a mechanical translation between Japanese and English by analogy principle. In Alick Elithorn and Ranan Banerji, editors, *Artificial and Human Intelligence*. North-Holland, Amsterdam, pages 173–180.
Rayner, Manny, and David Carter. 1997. Hybrid language processing in the spoken language translator. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 107–110, Munich.
Sato, Satoshi, and Makoto Nagao. 1990. Toward memory-based translation. In *COLING-90: Papers Presented to the 13th International Conference on Computational Linguistics*, pages 247–252, Helsinki.
Schäler, Reinhard. 1996. Machine translation, translation memories and the phrasal lexicon: The localisation perspective. In *Proceedings of TKE-96: EAMT Workshop on Machine Translation*, pages 21–33, Vienna.
Schäler, Reinhard, Andy Way, and Michael Carl. 2003. Example-based machine translation in a controlled environment. In Michael Carl and Andy Way, editors, *Recent Advances in Example-Based Machine Translation*, Kluwer Academic, Dordrecht, the Netherlands, pages 83–114.
Simard, Michel, and Philippe Langlais. 2001. Subsentential exploitation of translation memories. In *Machine Translation Summit VIII*, pages 335–339, Santiago de Compostela, Spain.
Somers, Harold. 1998. Further experiments in bilingual text alignment. *International Journal of Corpus Linguistics,* 3:115–150.
Somers, Harold, Ian McLean, and Daniel Jones. 1994. Experiments in multilingual example-based generation. In *CSNLP 1994: Third International Conference on the Cognitive Science of Natural Language Processing*, Dublin.
Veale, Tony, and Andy Way. 1997. *Gaijin*: A bootstrapping, template-driven approach to example-based machine translation. In *International Conference, Recent Advances in Natural Language Processing*, pages 239–244, Tzigov Chark, Bulgaria.
Watanabe, Hideo. 1993. A method for extracting translation patterns from translation examples. In *Proceedings of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI '93): MT in the Next Generation*, pages 292–301, Kyoto, Japan.
Zernik, Uri, and Michael Dyer. 1987. The self-extending phrasal lexicon. *Computational Linguistics*, 13(3–4):308–327.