# Application of Translation Knowledge Acquired by Hierarchical Phrase Alignment for Pattern-based MT

Kenji Imamura
ATR Spoken Language Translation Research Laboratories
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0288, Japan
kenji.imamura@atr.co.jp

**Abstract**

Hierarchical phrase alignment is a method for extracting equivalent phrases from bilingual sentences, even though they belong to different language families. The method automatically extracts transfer knowledge from about 125K English and Japanese bilingual sentences and then applies it to a pattern-based MT system. The translation quality is then evaluated. The knowledge needs to be cleaned, since the corpus contains various translations and the phrase alignment contains errors. Various cleaning methods are applied in this paper. The results indicate that when the best cleaning method is used, the knowledge acquired by hierarchical phrase alignment is comparable to manually acquired knowledge.

## 1 Introduction

Translation knowledge is necessary for machine translation (MT) systems. Automatic translation knowledge construction is an effective way to reduce costs when applying a system to other task domains.

Statistical translation methods (e.g., Brown et al. 1993) automatically acquire statistical models, which are considered elements of translation knowledge, so little cost is necessary. However, in most cases, these methods are applied to the same language families, such as English and French. In the case of different families, the translation quality is still unclear.

A hierarchical phrase alignment method has been proposed (Imamura 2001). This method hierarchically extracts equivalent phrases from a sentence-aligned bilingual corpus even though they belong to different language families. Kaji et al. (1992), Yamamoto & Matsumoto (2000), and Meyers et al. (2000) have also proposed methods to acquire translation knowledge automatically. They have evaluated the knowledge, but there are few examples in which the translation quality was evaluated when the entire knowledge was applied to translation systems (Menezes & Richardson 2001). This comprehensive level of quality should be measured on an actual translation system to judge whether the acquired knowledge is useful from a practical point of view.

In this paper, translation knowledge is acquired automatically by hierarchical phrase alignment and integrated into a pattern-based MT system, and then the resulting translation quality is evaluated. Through the integration, the problem of ungeneralized patterns contained within the knowledge became clear. Because this problem caused bad translations or increased ambiguities, it became obvious that the knowledge needed to be cleaned. The translation process studied here is from English to Japanese.

## 2 Abstract of Hierarchical Phrase Alignment

### 2.1 Basic Method

Phrase alignment refers to the extraction of equivalent partial word sequences between bilingual sentences. We use the term phrase alignment since these word sequences include not only words but also noun phrases, verb phrases, relative clauses, and so on.

For example, in the case of the following sentence pair,

**English:** *I have just arrived in New York.*
**Japanese:** *NewYork ni tsui ta bakari desu.*

the phrase alignment should extract the following word sequence pairs.

- *in New York* ↔ *NewYork ni*
- *arrived in New York* ↔ *NewYork ni tsui*
- *have just arrived in New York* ↔ *NewYork ni tsui ta bakari desu*

We call these 'equivalent phrases' in this paper and defined this task as extracting phrases that satisfy the following two conditions.

**Condition 1 (Semantic constraint):**
Words in the phrase pair correspond with no deficiency and no excess.

**Condition 2 (Syntactic constraint):**
The phrases are of the same syntactic category.

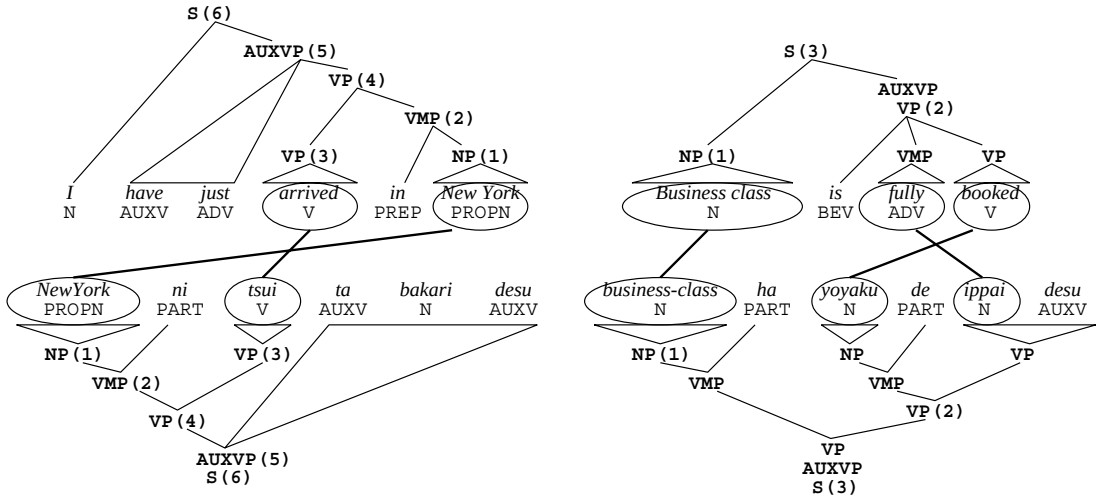
In order to extract phrases that satisfy two conditions, corresponding words (called 'word links,' represented as $WL(word_e, word_j)$) are first extracted by word alignment. Next, the sentence pair is parsed respectively, and phrases and their syntactic categories are acquired. Finally, the phrases, which include some word links, exclude other links, and are of the same syntactic categories, are regarded as equivalent.

For example, in the case of Figure 1(a), `NP(1)` and `VMP(2)` are regarded as equivalent because they only include $WL(New\ York, NewYork)$, and are of the same syntactic category. In the case of $WL(arrived, tsui)$, `VP(3)` is regarded as equivalent, and in the case of both word links, `VP(4)`, `AUXVP(5)`, and `S(6)` are regarded as equivalent. Consequently, six equivalent phrases are extracted hierarchically.

Even though word links are available, the part-of-speech (POS) of the words is sometimes different in different languages, as shown in the second example in Figure 1(b). In this case, the phrases that contain only $WL(fully, ippai)$ or only $WL(booked, yoyaku)$ are not regarded as equivalent because of the syntactic constraint, and `VP(2)` nodes are extracted first. Thus, few unnatural short phrases are extracted as equivalent.

### 2.2 Increasing Robustness

The problem in the above method is that the result of the phrase alignment directly depends on the parsing result. We solved this problem by using the following features and techniques, and partial correspondences were extracted even though parsing failed. In the experiment of Imamura (2001), about twice as many equivalent phrases were extracted compared with the basic method and almost no deterioration was observed.




(a) Example of Simple Translation (b) Example of Different POS Translation

Figure 1: Examples of Hierarchical Phrase Alignment
(Upper and lower trees indicate English and Japanese, respectively;
lines between languages indicate word links.)

**Disambiguation Using Structural Similarity:** As Kaji et al. (1992) and Matsumoto et al. (1993) showed, some parsing ambiguities can be eliminated when the two languages are made to correspond. This disambiguation utilizes structural similarity. For example, a PP attachment in English is ambiguous as to whether it modifies a noun or a verb, but this is nearly always definite in Japanese. Hence, when the attachment is assumed to modify the same word, the ambiguity is eliminated. Accordingly, the structures between the two languages become similar.

We employ a '**phrase score**' to measure structural similarity. This measure is calculated by counting the phrases that satisfy the above two conditions, and the parsing candidate that has the maximum score is selected.

**Combination of Partial Trees:** Partial parsing is an effective way to avoid a lack of grammar or to parse ungrammatical sentences. It is used to combine partial candidates in the parser. Therefore, a criterion as to whether the part is valid or not is necessary for the combining process. We utilize the phrase score as the criterion, and a partial tree sequence that maximizes the sum of the phrase scores is searched for. The forward DP backward A* search algorithm (Nagata 1994) is employed to speed up the combination.

## 2.3 Placement in Translation Knowledge Acquisition

The phrase alignment result by this method maintains correspondent parse trees and hierarchical information, so it is especially suitable for MT systems using syntactic transfer methods.

Moreover, a characteristic of this method is the introduction of a syntactic constraint

(Condition 2). [1] There are two effects of the syntactic constraint. One is that few unnatural short phrases are extracted, as described above. The other is that it is easy to construct translation patterns because the phrases can be grammatically replaced.

In other words, suppose that an equivalent phrase replaces another one that is extracted from another sentence. If the source phrase and the target phrase are in the same syntactic category, the resulting synthesized sentence is appropriate. On the other hand, if they are in different categories, the source or target sentence becomes grammatically inappropriate. The syntactic constraint suppresses such replacement. This is a particular advantage for translation between different language families, since this phenomenon appears more frequently in such case than in translation between languages of the same language family.

# 3 Transfer Driven Machine Translation (TDMT)

The Transfer Driven Machine Translation system, or TDMT (Furuse & Iida 1994; Sumita et al. 1999), used here is an MT system based on the syntactic transfer method. The following sections describe the abstract focusing of the transfer module.

## 3.1 Transfer Patterns

Transfer patterns represent the correspondence between source language expressions and target language expressions. They are the most important kinds of knowledge in TDMT. Examples are shown in Figure 2 that include the preposition '*at.*' In this pattern, source language information is constructed by a source pattern and its syntactic category. The source pattern is a sequence of instantiate-able variables and constituent boundaries (functional words or part-of-speech bigram markers). The instantiation of each variable is restricted by a syntactic category using daughter patterns. Namely, source language information is equivalent to Context Free Grammar such that the right side of each rewrite rule absolutely contains at least one terminal symbol.

Target patterns are constructed with variables and constituent boundaries, but they do not have POS bigram markers. In addtion, each pattern has examples, which are instances of variables. The examples are headwords acquired from training sentences. For instance, the first rule of Figure 2 means that the English phrase "*present at (the) conference*" was translated into the Japanese phrase "*kaigi* "conference" *de happyo-suru* "present"."

## 3.2 Translation Process

At the time of translation, the source sentence is parsed using source patterns. Then, the target structure, which is mapped by target patterns, is generated (Figure 3). However, as shown in Figure 2, one transfer pattern has multiple target patterns. In order to select an appropriate target pattern, semantic distances (node distances on the thesaurus; refer to Sumita & Iida 1991) are calculated between the examples and the daughter headwords of the input sentence, and the target pattern that has the nearest example is selected. Therefore, each pattern also has head information.

[1] The methods of Yamamoto & Matsumoto (2000) and Meyers et al. (2000) do not use syntactic categories. Alternatively, dependency structures are utilized. Chunks and relationships may be substituted

| Syn. Cat. | Source Pattern | | Target Pattern | Example |
|---|---|---|---|---|
| VP | $X_{VP}$ *at* $Y_{NP}$ | ⇒ | Y' *de* X' | ((*present*, *conference*) ...) |
| | | | Y' *ni* X' | ((*stay*, *hotel*), (*arrive*, *p.m*) ...) |
| | | | Y' *wo* X' | ((*look*, *it*) ...) |
| NP | $X_{NP}$ *at* $Y_{NP}$ | ⇒ | Y' *no* X' | ((*man*, *front desk*) ...) |

Figure 2: Examples of Transfer Patterns in which the Constituent Boundary is '*at*'

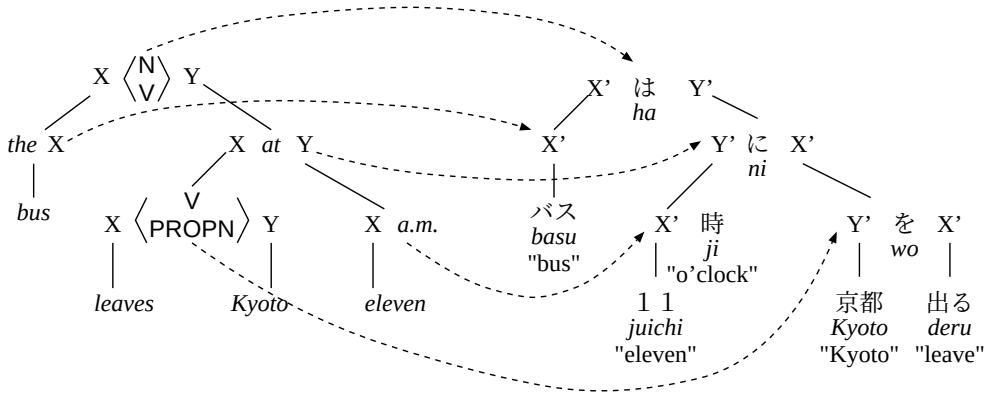



Figure 3: Example of TDMT Transfer Process

For example, in the case of the input sentence "*The bus leaves Kyoto at eleven a.m.*," the source pattern (X *at* Y) is used. Then, the headword of the variable X is '*leave*,' and Y is '*a.m.*' According to the semantic distance calculation, the example (*arrive*, *p.m.*) is the nearest. Therefore, the target pattern (Y' *ni* X') is selected. The semantic distance calculation is also utilized for parsing disambiguation.

### 3.3 Content Word Selection

Functional words are translated by the above process. In the case of content words, TDMT generates a default translation at leaf variables by referring to a translation dictionary. However, a single word is often translated into different words in different contexts. For example, in the case of the English phrase "*leave Kyoto*," '*leave*' should be translated into '*deru* "go out".' On the other hand, in the case of "*leave my wallet on the table*," '*leave*' should be translated into '*okisaru* "put and go".'

Content word selection is achieved in two ways. One is by using local dictionaries, which are translation dictionaries created for each target pattern. When an instantiated variable of a source pattern equals an example, the system refers to the local dictionary and generates the translated word (Yamada et al. 1998). Another way is by embedding content words that can generate different translations into the source pattern in advance.

for categories. However, this approach is not declarative.

# 4 Application of Phrase Alignment Results for TDMT

In this section, we describe how to generate TDMT transfer patterns from the results of phrase alignment and the problems of this method.

## 4.1 Transfer Pattern Generation

The transfer patterns described in Section 3 are constructed by source patterns that include their syntactic category, target patterns, examples, head information, and local dictionaries. They are generated as follows from the phrase alignment results.

- Source patterns and target patterns are generated from the parsing tree by eliminating non-corresponding nodes and regarding daughter corresponding nodes as variables. POS bigram markers are generated from leaf word sequences and embedded into source patterns.
- Head information is acquired from grammar, and examples are identified by tracing the parsing tree to the head branch.
- Local dictionaries are created by word links and by extracting leaf equivalent phrases in which the source phrase contains only a word.

In addition, because the inputs of phrase alignment are aligned sentences, sentence correspondences are added to the phrase alignment results as equivalent.

## 4.2 Pattern Cleaning

Even after transfer patterns are generated, they may contain many ungeneralized patterns. The reasons for this are as follows:

**(1) Reasons for Bad Translation**

- Ellipses or additional words are contained in patterns due to context-dependent equivalent phrases. For instance, the determiner '*the*' is not generally translated when English is translated into Japanese. However, when a human translator cannot semantically identify the following noun, a determinant modifier such as '*watashi-no* "my"' or '*sono* "its"' is supplied. These patterns depend on the context, so if they are used in the wrong context, the translation will be wrong.
- Incorrect phrase alignment. This causes the wrong transfer patterns not only in themselves but also in parent patterns that have variables instantiated by the result.

  For example, in Figure 1(b), suppose that an incorrect pair of the English phrase "*book*" and the Japanese phrase "*yoyaku de ippai desu*" are extracted as equivalent. This resulet will deliver the incorrect transfer patterns (*book*) ⇒ (*yoyaku de ippai desu*) and (X *is fully* Y) ⇒ (X' *ha* Y').

The experiment described in Imamura (2001) shows that 6% of the phrases were context dependent and 8% were incorrect even if word alignment was carried out by hand.

Table 1: Statistics of the Corpus

| | English | Japanese |
|---|---|---|
| Sentence# | 125,579 | |
| Total Word# | 721,848 | 774,711 |
| Vocabulary# | 9,945 | 14,494 |
| Equivalent Phrase# | 404,664 | |
| (including Sentence Correspondence) | (463,869) | |
| Different Pattern# | 56,851 | 53,317 |

**(2) Reasons for Good Translation but Ungeneralized Patterns**

- The corpus contains a variety of translations even for a single source sentence. For example, in the corpus used for the experiments of Section 5, the English sentence "*How can I get there?*" is translated into thirty Japanese sentences. These translations cause various patterns. However, these translations can be unified, so most patterns will be unnecessary.
- When the phrase alignment result partially lacks correspondence, patterns in which the variables are instantiated in advance are generated.

  For example, if the correspondence `VP(2)` is missing in Figure 1(b), the transfer pattern (X *is fully booked*) ⇒ (X *ha yoyaku de ippai desu*) will be generated from `S(3)`. These patterns are correct but clearly ungeneralized.

Meyers et al. (2000) referred to this problem as an explosive number of rules and decreasing translation speed. They tried to solve it by selecting rules based on the frequency during translation. TDMT performs pattern selection by semantic distance calculation, so the translation speed decreases only slightly even if there are many patterns. On the other hand, because TDMT does not employ frequency, low-frequency patterns of type (1) cause bad translation. Therefore, they should be cleaned in advance.

The experiment in the next section compares the translation quality among different cleaning methods.

# 5 Evaluation

English to Japanese translation is evaluated in this paper.

## 5.1 Experimental Settings

**Corpus for Pattern Generation** We built a collection of Japanese sentences and their English translations based on expressions that are usually found in phrasebooks for foreign tourists. We used about 125K sentences in the corpus, and the basic statistics are shown in Table 1. The different pattern numbers in Table 1 indicate the numbers of different source (English) and target (Japanese) patterns, respectively.

**Evaluation Measure** Each experiment used the same test set, which was composed of 508 sentences randomly selected in advance from the corpus and excluded from the training set. The evaluation was carried out by one Japanese native speaker. He/She evaluated the EJ translation into the following four ranks (Sumita et al. 1999) from the viewpoint of a user. In this paper, we call (A+B+C) the 'translation rate.'

**(A)** Perfect: no problem in either information or grammar.
**(B)** Fair: easy-to-understand with some unimportant information missing or flawed grammar.
**(C)** Acceptable: broken but understandable with effort.
**(D)** Nonsense: important information has been translated incorrectly.

**Cleaning Methods** We employed the following pattern cleaning methods.

- **No cleaning:**
  All patterns were integrated into the TDMT.
- **Cutoff by frequency:**
  The frequency was counted for each source and target pattern pair, and transfer patterns were generated only from high-frequency pairs in the same manner as in Menezes & Richardson's (2001) experiment. In this experiment, the pairs that appeared more than two times were used.
- **$\chi^2$ test:**
  Considering that source and target patterns occur independently, the $\chi^2$ test was performed. In this process, only high-frequency patterns were tested in order to rely on the $\chi^2$ value. That is, the co-occurrence frequency was over 40, or the co-occurrence frequency was over 20 and the independent occurrence was 5 more than the co-occurrence frequency. In addition, the threshold was set at the 95% reliable point ($\chi^2 \geq 3.841$).
- **Manual cleaning:**
  Based on the $\chi^2$ patterns, manual adjustment of the test set was made by only eliminating or adding patterns. Additional patterns were obtained from the unused "No cleaning" patterns. The purpose of this experiment was to measure the translation quality when a theoretically perfect cleaning method is applied.

## 5.2 Result of Experiments

The pattern numbers for each cleaning method are shown in Table 2, and the translation quality is shown in Figure 4. The transfer pattern pair in Table 2 means a pair of source and target patterns. A TDMT that is integrated with fully hand-made transfer patterns is also shown for reference. The hand-made patterns were created from a different corpus (dialogue corpus; refer to Furuse et al. 1994), so it cannot be compared directly, but it contains a sufficient number of appropriate patterns.

First, among the fully automatic pattern generation methods (No cleaning, Cutoff by frequency, and $\chi^2$ test), the best method was Cutoff by frequency, which achieved a 72% translation rate.

Table 2: Pattern Numbers for Each Cleaning Method

| Cleaning Method | Number of Source and Target Pattern Pairs | Transfer Pattern# |
|---|---|---|
| No cleaning | 92,005 | 56,910 |
| Cutoff by freq. | 10,011 | 5,478 |
| $\chi^2$ test | 922 | 504 |
| Manual cleaning | 1,172 | 635 |
| (Hand-made patterns) | 4,878 | 2,235 |

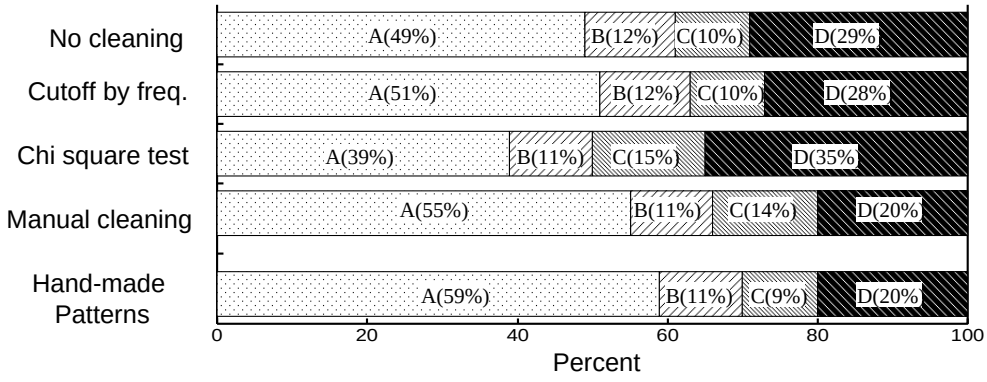



Figure 4: Translation Quality for Each Cleaning Method

In comparison with No cleaning, the pattern numbers decreased to about 1/9 in the case of Cutoff by frequency. However, the translation rate slightly increased. This means that almost all low-frequency patterns are redundant or inappropriate, and the Cutoff by frequency method performs moderately well and is simple.

The $\chi^2$ test patterns are reliable from the viewpoint of statistics, but the translation quality was lower. This is because the pattern number was insufficient, and the translations were divided into segments [2]. However, the translation quality did not deteriorate to 1/10 compared with Cutoff by frequency even though the pattern number decreased to about 1/10. This is because only general patterns remained. Therefore, if there are comprehensive and reliable patterns from the viewpoint of statistics, good translations can be achieved.

Finally, the translation quality of the Manual cleaning method was almost the same as that of the Hand-made patterns. The patterns generated from phrase alignment results contained comprehensive patterns in the same way as the Hand-made patterns. Therefore, if there is an effective cleaning method, the quality will increase in the same way as in hand-made TDMT.

[2]TDMT has a partial translation function if there are no patterns for parsing.

## 6 Discussion

**Corpus Size for Stastical Pattern Cleaning** The $\chi^2$ test is one of the methods that acquire word translations from a bilingual corpus (Gale & Church 1991). Since transfer patterns are regarded as word correspondences, the hypothesis test can be applied and good transfer patterns will be acquired. However, a sufficient number of patterns were not acquired (i.e., the coverage was low) in this experiment because of the small corpus.

Melamed (2000) shows an experiment using the Hansard Corpus (English and French). He used 300K bilingual sentences, and extracted translation words with a precision of 87% and coverage of 90%. There were about 41,000 different words for English and 36,000 for French.

Suppose that source and target patterns are regarded in the same way as translation words. 57,000 source patterns and 53,000 target patterns are generated in this experiment. About $\frac{57000*53000}{41000*36000} \simeq 2.0$ times resolution is necessary in comparison with Melamed (2000)'s experiment, and the sentence number becomes $300K * 2.0 = 600K$. Consequently, it is estimated that anywhere from a half million to one million bilingual sentences are necessary for statistical pattern cleaning.

**Longer Sentences** The corpus used here contains many short sentences. In the case of long sentences such as newswires, the accuracy of phrase alignment will decreased. However, it can be somewhat maintained if the techniques we described in Section 2.2 are applied. In fact, we could expect the problem that the transfer pattern number will increase because longer sentences contain more complex expressions. Even though TDMT translates them with short units, a larger corpus will be necessary to maintain coverage of the knowledge.

## 7 Conclusions

Using hierarchical phrase alignment, translation knowledge was acquired from a bilingual corpus of different language families. The acquired knowledge was applied to a translation system, TDMT, and its translation quality was evaluated. When the transfer patterns were cleaned automatically, the translation rate was about 72%.

Phrase alignment results contain high coverage patterns. If the patterns are combined correctly, it is possible to obtain good translations that are similar to hand-made patterns.

Since the corpus contains context-dependent translations and the phrase alignment results have errors, the transfer patterns need to be cleaned. Although we used a large corpus of 125K sentences, in which over fifty thousand transfer patterns appeared, the patterns could not be cleaned to the level that made them as useful and reliable as hand-made patterns.

Future research topics will include enriching our corpus and investigating cleaning methods that offer reliability and high coverage despite sparse data.

## References


Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra & Robert L. Mercer: 1993, 'The mathematics of machine translation: Parameter estimation', *Computational Linguistics*, **19**(2): 263–311.

Furuse, Osamu & Hitoshi Iida: 1994, 'Constituent boundary parsing for example-based machine translation', in *Proceedings of COLING-94*, pp. 105–111.

Furuse, Osamu, Y. Sobashima, Toshiyuki Takezawa & N. Uratani: 1994, 'Bilingual corpus for speech translation', in *Proceedings of AAAI'94 Workshop 'Integration of Natural Language and Speech Processing'*, pp. 84–91.

Gale, William A. & Kenneth W. Church: 1991, 'Identifying word correspondences in parallel texts', in *Proceedings of 4th DARPA Workshop on Speech and Natural Language, Asilomar, CA*, pp. 152–157.

Imamura, Kenji: 2001, 'Hierarchical phrase alignment harmonized with parsing', in *Proceedings of 6th Natural Language Processing Pacific Rim Symposium (NLPRS 2001)*, pp. 377–384.

Kaji, Hiroyuki, Yuuko Kida & Yasutsugu Morimoto: 1992, 'Learning translation templates from bilingual text', in *Proceedings of COLING-92*, pp. 672–678.

Matsumoto, Yuji, Hiroyuki Ishimoto & Takehito Utsuro: 1993, 'Structural matching of parallel texts', in *the 31st Annual Meeting of the ACL*, pp. 23–30.

Melamed, I. Dan: 2000, 'Models of translational equivalence among words', *Computational Linguistics*, **26**(2): 221–249.

Menezes, Arul & Stephen D. Richardson: 2001, 'A best first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora', in *Proceedings of the 'Workshop on Example-Based Machine Translation' in MT Summit VIII*, pp. 35–42.

Meyers, Adam, Michiko Kosaka & Ralph Grishman: 2000, 'Chart-based translation rule application in machine translation', in *Proceedings of COLING-2000*, pp. 537–543.

Nagata, Masaaki: 1994, 'A stochastic Japanese morphological analyzer using a forward-DP backward-A* N-best search algorithm', in *Proceedings of COLING-94*, pp. 201–207.

Sumita, Eiichiro & Hitoshi Iida: 1991, 'Experiments and prospects of example-based machine translation', in *Proceedings of the 29th ACL*, pp. 185–192.

Sumita, Eiichiro, Setsuo Yamada, Kazuhide Yamamoto, Michael Paul, Hideki Kashioka, Kai Ishikawa & Satoshi Shirai: 1999, 'Solutions to problems inherent in spoken-language translation: The ATR-MATRIX approach', in *Machine Translation Summit VII*, pp. 229–235.

Yamada, Setsuo, Kazuhide Yamamoto & Hitoshi Iida: 1998, 'Word selection on cooperative integrated machine translation (in Japanese)', in *Proceedings of The 4th Annual Meeting of The Association for Natural Language Processing*, pp. 508–511.

Yamamoto, Kaoru & Yuji Matsumoto: 2000, 'Acquisition of phrase-level bilingual correspondence using dependency structure', in *Proceedings of COLING-2000*, pp. 933–939.