Corpus-based Generation of Numeral Classifier using Phrase Alignment

Michael PAUL[*†] and **Eiichiro SUMITA**[*] and **Seiichi YAMAMOTO**[*†]
[*]ATR Spoken Language Translation Research Laboratories, [†]Kobe University
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0288, Japan
{Michael.Paul, Eiichiro.Sumita, Seiichi.Yamamoto}@atr.co.jp

Abstract

A severe problem for NLP applications dealing with multilingual language resources is the acquisition of knowledge that is obligatory in one language but not explicitly expressed in another language. In this paper, we focus on *numeral classifiers*, which are required in languages like Japanese but are usually not explicitly used in languages like English, which don't have such a classifier system.

We propose a uniform method to assign the numeral classifiers of languages that have a numeral classifier system to the numerals of non-classifier languages. The omitted classifier information is extracted from a bilingual corpus based on phrasal correspondences in the contexts of the respective sentences.

1 Introduction

Various NLP applications, such as machine translation, are confronted with the need to acquire knowledge that is obligatory in one language but not explicitly expressed in another language. However, based on the assumption that the word sequences of bilingual sentences have the same semantic information, knowledge of specific features of one language that do not exist in the other language can be extracted from phrasal correspondences.

In this paper, we focus on *numeral classifiers* (cf. Section 2), which are required in languages like Japanese to categorize the objects that speakers count or quantify but are usually not explicitly used in languages like English, which don't have such a classifier system. In the case of a machine translation application, knowledge of omitted classifier information has to be recovered when translating into languages with numeral classifier systems.

Our corpus-based method extracts the omitted classifier information from corresponding phrases of bilingual samples similar to the input. First, we retrieve bilingual sentence pairs from the training corpus by performing a *dynamic programming* match between the input sentence and training sentences of the same language. In the second step, a *phrase alignment* method is applied to the selected bilingual sentence pairs in order to identify corresponding phrases (cf. Section 4). The aligned phrasal knowledge of numeral classifiers is then reused to generate numeral classifiers corresponding to numeral expressions of the non-classifier language as described in Section 5.

2 Numeral Classifier

Numeral classifiers are components of a grammatical system that reflects how speakers categorize objects that they count or quantify. They form a group of morphemes that usually occur adjacent to quantity expressions that include numbers. Around 37 East and Southeast Asian languages have numeral classifier systems (Adams and Conklin, 1973).

Japanese, for example, has more than 150 different numeral classifiers, but only 30-80 are found in daily use (Downing, 1996). Japanese classifiers (*josūshi*) are a subclass of nouns and can postfix to numerals and quantificational nouns, such as *sū* (some) or *nani* (what), to form a noun phrase. They cannot form grammatical noun phrases on their own.

However, in languages without such a classifier system, like English, numerals can directly modify countable nouns or can even be used anaphorically, i.e. without any quantified noun. Uncountable nouns have to be reclassified as countable ones or embedded in a partitive construction (Quirk et al., 1985).

Although similar patterns for languages with classifier systems have been reported in (Croft, 1994), classifiers, their relationship to nouns,

and their realization (i.e. the syntactic position of numeral classifier phrases within the sentence), are language-dependent. The selection of numeral classifiers is determined by the inherent semantic properties of the noun whose quantity is being specified in the context of the respective utterance.

Concerning the realization of quantifier constructions, (Asahioka et al., 1990) distinguishes the types given in Table 1 for Japanese[1].

Table 1: Japanese Quantifier Constructions

Type	Form
prenominal	XC "*no*" N m
appositive	N XC m
floating	N m XC
partitive	N "*no*" XC m
attributive	XC N m
anaphoric	XC m
predicative	N "*wa*" XC "*da*"

Therefore, the generation of numeral classifiers requires not only the selection of an appropriate classifier but also its type for its realization within the context of the sentence.

3 Related Research

Previous work in numeral classifiers focused mainly on the syntactic, semantic and functional aspects of numeral classifier systems within a particular language (Downing, 1996). Less research has been conducted on the generation of numeral classifier phrases.

(Sornlertlamvanich et al., 1994) proposed an algorithm for generating numeral classifiers in Thai that uses default classifiers associated with each noun defined in a lexicon. However, no evaluation results have been published.

(Bond and Paik, 2000) reuses semantic classes from an ontology for the generation of Japanese numeral classifiers. Their algorithm associated classifiers with semantic classes and used inheritance to dynamically select the classifier of a noun according to its most typical semantic class. They evaluated their algorithm using 90 noun phrases modified by sortal classifiers and achieved an accuracy of 81%. The upper boundary of their method was given as 88%.

(Bond et al., 1996) used bilingual information for the translation of Japanese numeral classifiers into English. Their analysis of prenominal classifiers is based on the characteristics of, and the differences between, Japanese and English, i.e. restrictions on countability and number of the embedded English noun phrase in a partitive construction. Although their rule-based approach was implemented in a Japanese-to-English machine translation system, no evaluation results of its accuracy have been presented.

In our approach, the information required for the generation of numeral classifiers is extracted automatically from a corpus using phrase alignment between bilingual sentence pairs. This provides a uniform approach for various numeral classifier constructions.

4 Phrase Alignment

The term *phrase alignment* refers to the extraction of equivalent partial word sequences between bilingual sentences. Not only single words but also more complex grammatical constituents like noun or verb phrases can be aligned based on the syntactic structure of each sentence. Equivalent phrases indicate corresponding expressions between two languages. Based on the assumption that the word sequences of bilingual sentences have the same semantic information, knowledge of specific features of one language that do not exist in the other language can be extracted from these phrasal alignments.

Various phrase alignment methods have been proposed, such as (Kaji et al., 1992), (Matsumoto et al., 1993), (Kitamura and Matsumoto, 1995), (Meyers et al., 1996), (Yamamoto and Matsumoto, 2000), (Imamura, 2001), and (Richardson et al., 2001).

An example of a phrase alignment between bilingual sentences is given in Figure 1. The circled words and connecting lines mark word alignments. Based on these restrictions, both sentence structures can be compared, and equivalent phrases can be extracted. The equivalent phrases (1)NP and (3)NP of our example align the English numerals to the corresponding Japanese numeral classifier expressions. Therefore, we can interpret the English phrase "at eight" as a time expression ("X *ji*" $\leftrightarrow$ "X o'clock") and "two" as the number of people of the reservation ("X *ri*" $\leftrightarrow$ "X person(s)").

[1] X - numeral, C - classifier, N - quantified noun, m - case marker

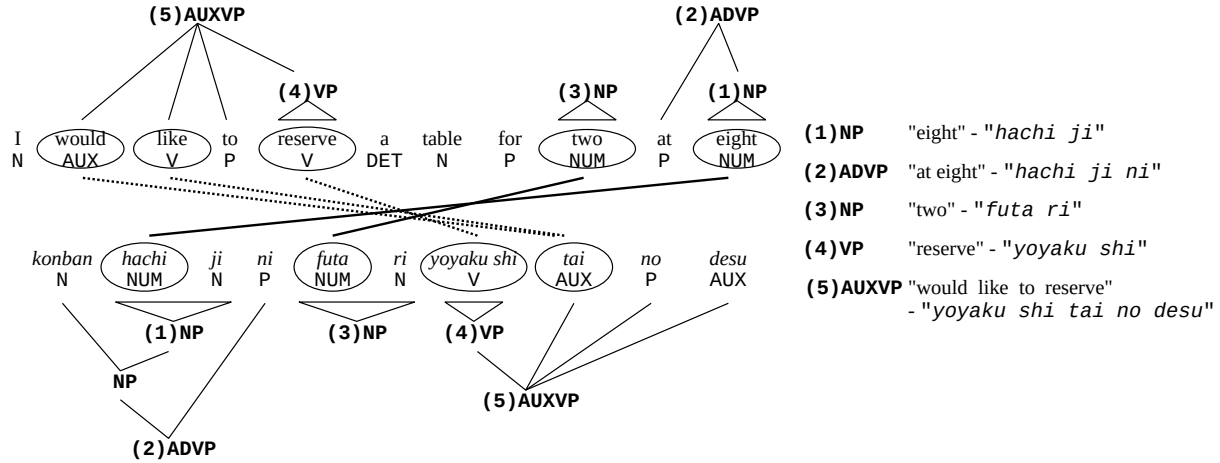

Figure 1: Phrase Alignment

5 Generation of Numeral Classifier

In our experiments described in Section 6, we used English as a representative of languages without a numeral classifier system and Japanese as a representative of numeral classifier languages.

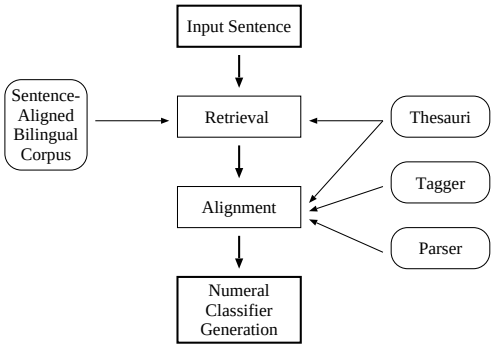

Figure 2: Configuration

The proposed method illustrated in Figure 2 takes an English utterance, containing numerals without a quantified noun, as the input and outputs information (classifier word and realization type) required for the generation of the corresponding Japanese numeral classifier.

Our resources are (1) a bilingual corpus, in which sentences are aligned beforehand, (2) thesauri of both languages, which are used in aiding word alignment and incorporating the semantic distance between words, and (3) the tagger and parsers of the alignment module.

5.1 Sentence-Retrieval

Given an input sentence containing numerals, we retrieve the most similar examples in the training corpus by utilizing the method proposed in (Sumita, 2001). This method carries out DP-matching of the input sentence and example sentences while measuring the semantic distance of the words. The source parts of all example sentences in the bilingual corpus are examined. By measuring the distance between the word sequences of the input and example sentences, it retrieves the examples with the minimum distance, provided the distance is smaller than the given threshold. Otherwise, the retrieval step fails with no output. The distance is calculated by a standard *dynamic programming* technique (Cormen et al., 1996) as

$$dist = \frac{I + D + 2 * \sum \frac{K}{N}}{L_{input} + L_{example}} \qquad (1)$$

The counts of the Insertion (I), Deletion (D), and Substitution (S) operations are summed up, and the total is normalized by the sum of the length of the source (L_{input}) and example sequences ($L_{example}$). Substitution considers the semantic distance between two substituted words and is defined as the division of K, the level of the least common abstraction in the thesaurus of two words, by N, the height of the thesaurus (Sumita and Iida, 1991).

5.2 Numeral Phrase Alignment

In order to identify Japanese numeral classifiers corresponding to numeral expressions of the matched English utterances, the equivalent Japanese utterances of the training corpus are extracted, and each of the bilingual lan-

guage pairs is aligned by utilizing the *hierarchical phrase alignment* algorithm proposed in (Imamura, 2001).

This method aligns bilingual texts phrase-by-phrase from partial parse results as illustrated in Figure 1. First, both sentences are analyzed morphologically and parsed using a chart parser resulting in (possibly partial) sentence structures. In the second step, links between single words are established. Finally, corresponding phrases are identified according to the similarity of syntactic categories of the nodes in both parse trees. If a sentence cannot be parsed completely, the system uses the combination of partial tree results for the alignment process. Assignment ambiguities are resolved by using structural similarities between the languages.

5.3 Numeral Classifier Assignment

All aligned phrases containing numeral expressions are then mapped to the input sentence, i.e. the surface words of the phrase alignments are replaced with those of the input sentence by utilizing the sequence of insertions (I), deletions (D), and substitutions (S) of the DP match as mapping rules as illustrated in Figure 3.

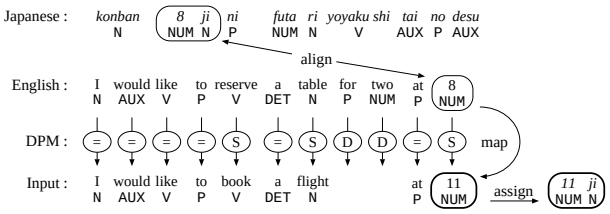

Figure 3: Classifier Assignment

Thus, we obtain a list of possible Japanese numeral classifier assignments for each numeral of the English input. Based on the frequency of matched utterances and aligned phrases of the numerals, we choose the most frequent one for the generation of the assigned numeral classifier expression.

In addition, the type of numeral classifier is obtained by a set of heuristic rules matching the patterns given in Table 1 in each Japanese utterance obtained by the DP match. These rules verify the existence and order of the pattern constituents in the respective sentences and the most frequent type is selected for the generation of the numeral classifier.

6 Evaluation

To evaluate our approach, we used a collection of Japanese utterances and their English translations, which are usually found in phrasebooks for tourists going abroad (Takezawa et al., 2002). The translations were made utterance-by-utterance, resulting in an utterance-aligned corpus. Its size is about 200K, and the average length for Japanese and English utterances is 7.7 and 5.5, respectively.

Moreover, we used thesauri whose hierarchies are based on the Kadokawa Ruigo-shin-jiten (Ohno and Hamanishi, 1984) for distance calculation and word alignment.

6.1 Numeral Classifier Data

In the corpus, 15.7% of the utterances contain numeral expressions, of which 16% (3% of the entire corpus) are numerals without a quantified noun. We split the obtained 6000 utterances randomly into a test set of 300 utterances, used for the evaluation of our approach, and a learning set, used for the retrieval and alignment steps. On average, 1.1 numerals were contained in the utterances of the test data with an average length of 9.8. Table 2 summarizes the frequency of classifiers corresponding to the numerals of the learning set and illustrates the large variety of the different numeral classifiers that have to be assigned.

Table 2: Numeral Classifier Frequency

Classifier	Referent	%
none	uncountable	20.1
–ji	time expression	17.8
–nin	human being	12.4
–ji–hun	time expression	12.2
–tsu	general object	8.5
–mei	human being	3.4
–doru	dollar	3.1
–mai	flat object	2.9
–bin	flight	2.5
–goshitu	room	2.2
–ko	concrete object	1.3
–hai	cup, glass	1.2
others	26 classifiers	12.4

The distribution of the syntactic positions of these classifiers is given in Table 3.

6.2 DP Matching

The coverage of the DP matching method reported in (Sumita, 2001) was 89.2% when ap-

Table 3: Distribution of Syntactic Position

type	%
anaphoric	43.6
floating	13.1
appositive	12.4
predicative	10.0
prenominal	9.1
attributive	7.6
partitive	4.2

plied to 500 randomly selected samples of the entire corpus. More specifically, 46.4% of the examples with an average length of 5.6 were matched exactly, whereas 42.8% of the utterances with an average length of 7.7 were matched approximately. No utterance was retrieved when there was no examples whose distance is within the given threshold of 0.3. The average length of 11.0 for *no-match* utterances shows clearly that the DP match did not perform well for long utterances due to the lack of an explicit step decomposing an input utterance into sub-sentences. However, the shorter the utterance, the higher is the possibility that there exists a similar one in the example database. On average, three example utterances were retrieved from the database.

6.3 Hierarchical Phrase Alignment

The accuracy levels of the Japanese and English parser reported in (Imamura, 2001) were 52% and 44%, respectively, and the accuracy of the extracted equivalent phrases was about 86%.

The algorithm succeeded in aligning 2.6 phrases, on average, when applied to the utterance pairs obtained by the retrieval step.

6.4 Numeral Classifier Generation

For the evaluation of the numeral classifier assignment, we provided the English input utterance, its corresponding Japanese utterance from our bilingual corpus, the extracted numerals, and the respective classifier information assigned by our system to a native Japanese, who assigned one of the following ranks:

Rank	Evaluation
A	same classifier/type as in corpus
B	different classifier/type, but acceptable
C	incorrect classifier/type
D	no output

We used the *recall* and *precision* measures defined below for the evaluation of our system, where ranks A and B are considered correct assignments.

$$\text{recall} = \frac{\text{number of correctly assigned numeral classifiers}}{\text{number of numerals of test set}} = \frac{A+B}{A+B+C+D}$$

$$\text{precision} = \frac{\text{number of correctly assigned numeral classifier}}{\text{number of assigned numeral classifier}} = \frac{A+B}{A+B+C}$$

Figure 4 summarizes the *recall* and *precision* results according to the distance threshold of the DP matching step.

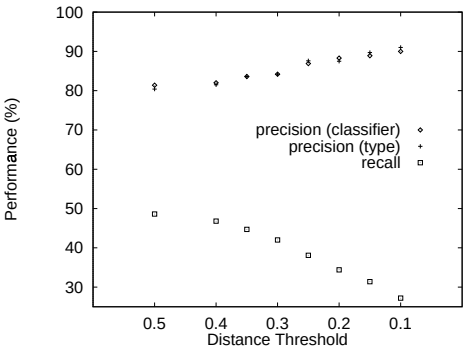

Figure 4: Recall/Precision Performance

The smaller the distance threshold, the fewer examples can be retrieved, resulting in a decrease in the system's recall. However, the more similar the examples, the more accurate the numeral classifier can be selected, leading to an increase in precision. If we select the DP matching threshold used in (Sumita, 2001), namely $dist < 0.3$, the recall is 42.0% due to a lack of similar examples for the retrieval step or to a failure of the alignment module. The detailed numbers of each rank are given in Table 4.

Table 4: Ranking Results for $dist < 0.3$

Rank	Classifier	Type
A	121	115
B	18	24
C	26	26
D	166	166
total	331	

Therefore, we succeeded in generating numeral classifiers correctly for 84.2% of the successfully matched and aligned test utterances. The precision of exact-match samples (dist=0.0) is 90%, thus imposing an upper boundary for our method that conforms with the results described in (Bond and Paik, 2000).

7 Discussion

The algorithm proposed in this paper uses bilingual utterance pairs and phrase alignment for the extraction of language-specific knowledge of the target language not available in the source language. We applied this method to the analysis of numeral expressions. Due to the example-based knowledge extraction, the problems concerning non-conceptual or one-to-many conceptual representations of the rule-base approaches mentioned in Section 3 were avoided, providing a uniform approach for the assignment of various Japanese numeral classifier constructions to corresponding numerals of the English source utterances.

In this section we focus on the limitations of our approach and propose some future directions on how to overcome these problems.

Like all example-based approaches, we suffer from the limited size of our database. If a similar example within the given threshold does not exist, we cannot apply our method. An increase in the corpus size lessens this deficiency, but the dominant problem for the retrieval of examples is in dealing with relatively longer utterances due to the lack of a preprocessing step for splitting utterances into smaller chunks. However, the identification of clauses containing numeral expressions and the retrieval of decomposed utterance fragments would lead to an increase in recall.

On the other hand, if the utterances are too short, there might not be enough contextual information to allow the disambiguation of the correct numeral classifier assignment. For example, utterances like "*three please*" trigger the retrieval of a large number of examples, but in order to select a classifier in a reliable way, context information beyond this utterance is required. Another example for the need of extra-sentential knowledge is the occurrence of mensural classifiers, like currencies (dollar vs. yen) or linear measures (mile vs. kilometer).

The phrase alignment module suffers from the low accuracies of the utilized parsers. However, the incorporation of parsers with better performance would allow the comparison of complete sentence structures triggering improved alignment results. Moreover, the lack of numeral word correspondences caused the failure of the phrase alignment step, e.g. the non-numeral use of "*one*" in "*a cheaper one*" (anaphoric) or "*one of your guests*" (determiner), could not be aligned in an appropriate way. Therefore, equivalent phrases could be better extracted by using a word alignment method with a high recall rate.

Numeral classifiers judged acceptable, i.e. those of rank B, can be divided into three groups: (1) selection of the general classifier *tsu* instead of more specific classifiers like *ko* (concrete objects), *mai* (flat objects), or *hon* (long, thin objects); (2) under-specification, i.e. the omission of classifiers, which might be acceptable for classifiers like *goshitu* (rooms) or *bin* (flights); (3) politeness, i.e. a classifier of the same meaning but used at different levels of politeness, like *meisama*, *mei*, and *nin* (human beings).

In the experiments described above, the obtained classifiers were accepted blindly, i.e. we did not verify the characteristics of the numeral expression of the input against those of the selected classifier. For example, the classifiers *tsu* (general) and *ri* (person) are used only in connection with specific numbers[2] but are not acceptable for larger numbers. Moreover, the assignment of classifiers corresponding to a sequence of numerals is limited to classifiers like *goshitu* (rooms) or *ban* (phone numbers), and it is not possible for sortal classifiers like *ko* (concrete objects).

8 Conclusions

In this paper we proposed a new, uniform method using phrasal alignments for the analysis of corresponding numeral phrase constructions in bilingual utterance pairs. The precision and recall rates of our method are 84.2% and 42.0%, respectively.

The quite low recall is due to the deficiency of the current implementation of retrieving examples that are similar to the input by dealing with relatively longer utterances. However, the decomposition of the input utterances would improve the coverage of our system. We plan to retrieve examples based on DP matching of phrases instead of complete sentences.

One of the main reasons for incorrect classifier selection is the lack of contextual knowledge

[2] *tsu* (general) used for $1 \leq num \leq 9$; *ri* (person) used for $1 \leq num \leq 2$

within short utterances, which leads to ambiguities that cannot be solved without the context in which the sentence is uttered. Moreover, it is necessary to verify the characteristics of numerals and quantified nouns to avoid incompatible assignments.

Despite these shortcomings, our approach is widely applicable in various natural language applications. Furthermore, it is not limited to the task evaluated here but can be extended to the word disambiguation of numeral noun phrases or even other linguistic features, like *number*, *tense*, or *modality*, that are not explicitly expressed in one language but obligatory in the other language.

Acknowledgments

The authors would like to thank Kadogawa-Shoten for providing us with the Ruigo-Shin-Jiten thesaurus. The research reported here was supported in part by a contract with the Telecommunications Advancement Organization of Japan entitled, *"A study of speech dialogue translation technology based on a large corpus."*.

References

K. Adams and N. Conklin. 1973. Towards a theory of natural classification. In *Proc. of the 9th meeting of the Chicago Linguistic Society*, pages 1–10, University of Chicago.

Y. Asahioka, H. Hirakawa, and S. Amano. 1990. Semantic classification and an analyzing system of Japanese numerical expressions. *IPSJ SIG Notes*, 90-NL-78:129–136.

F. Bond and K. Paik. 2000. Reusing an ontology to generate numeral classifiers. In *Proc. of the 18th COLING*, pages 90–96, Germany.

F. Bond, K. Ogura, and S. Ikehara. 1996. Classifiers in Japanese-to-English machine translation. In *Proc. of the 16th COLING*, pages 125–130, Copenhagen, Denmark.

H. Cormen, C. Leiserson, and L. Rivest. 1996. *Introduction to Algorithms*. MIT Press.

W. Croft. 1994. Semantic universals in classifier systems. *Word*, 45:145–171.

P. Downing. 1996. *Numeral Classifier System: The Case of Japanese*. John Benjamines, Amsterdam.

K. Imamura. 2001. Hierarchical phrase alignment harmonized with parsing. In *Proc. of NLPRS'01*, pages 377–384, Tokyo, Japan.

H. Kaji, Y. Kida, and Y. Morimoto. 1992. Learning translation templates from bilingual text. In *Proc. of the 14th COLING*, France.

M. Kitamura and Y. Matsumoto. 1995. A machine translation system based on translation rules acquired from parallel corpora. In *Proc. of Recent Advances in NLP*, pages 27–36.

Y. Matsumoto, H. Ishimoto, and T. Utsuro. 1993. Structural matching of parallel texts. In *Proc. of the 31th ACL*, pages 23–30.

A. Meyers, R. Yarngaber, and R. Grishman. 1996. Alignment of shared forests for bilingual corpora. In *Proc. of the 16th COLING*, pages 460–465, Copenhagen, Denmark.

S. Ohno and M. Hamanishi. 1984. *Ruigo-Shin-Jiten*. Kadokawa.

R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, Essex.

S. Richardson, W. Dolan, A. Menezes, and J. Pinkham. 2001. Achieving commercial-quality translation with example-based methods. In *Proc. of the Machine Translation Summit VIII*, pages 293–297, Santiago de Compostela, Spain.

V. Sornlertlamvanich, W. Pantachat, and S. Meknavin. 1994. Classifier assignment by corpus-based approach. In *Proc. of the 15th COLING*, pages 152–159, Kyoto, Japan.

E. Sumita and H. Iida. 1991. Experiments and prospects of example-based machine translation. In *Proc. of the 29th ACL*, pages 185–192.

E. Sumita. 2001. Example-based machine translation using DP-matching between word sequences. In *Proc. of the 39th ACL, Workshop: Data-Driven Methods in Machine Translation*, pages 1–8, Toulouse, France.

T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto. 2002. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Proc. of the 3rd LREC*, pages 147–152, Las Palmas, Spain.

K. Yamamoto and Y. Matsumoto. 2000. Acquisition of phrase-level bilingual correspondence using dependency structure. In *Proc. of the 18th COLING*, pages 933–939, Saarbruecken, Germany.