Non-Compositional Language Model and Pattern Dictionary Development for Japanese Compound and Complex Sentences

Satoru Ikehara, Masato Tokuhisa, Jin'ichi Murakami
Tottori University,
Koyama-Minami Tottori, 680-8552, Japan
{ikehara, tokuhisa, murakami}@ike.tottori-u.ac.jp

Abstract

To realize high quality machine translation, we proposed a *Non-Compositional Language Model*, and developed a sentence pattern dictionary of 226,800 pattern pairs for Japanese compound and complex sentences consisting of 2 or 3 clauses. In pattern generation from a parallel corpus, *Compositional Constituents* that could be generalized were 74% of independent words, 24% of phrases and only 15% of clauses. This means that in Japanese-to-English MT, most of the translation results as shown in the parallel corpus could not be obtained by methods based on *Compositional Semantics*. This dictionary achieved a *syntactic coverage* of 98% and a *semantic coverage* of 78%. It will substantially improve translation quality.

1 Introduction

A wide variety of machine translation (MT) methods are being studied(Nagao, 1996; Brown et al., 1990; Vogel et al., 2003), but to obtain high-quality translations between languages belonging to different families that are alien each other is difficult. Most practical systems still employ a transfer method based on *compositional semantics*. A problem with this method is that it produces translations by separating the syntactic structure from meaning, and is thus liable to lose the meaning of the source text.

Better translation quality can be expected from *pattern-based MT* and *example-based MT* where the syntactic structure and semantics are handled together. However, *pattern-based MT* require immense pattern dictionaries that are difficult to develop(Jung et al., 1999; Uchino et al., 2001).

Meanwhile, *example-based MT*(Nagao, 1984; Sato, 1992; Brown, 1999) obtains translation results by substituting semantically similar elements in structurally matching translation examples, so a pre-prepared pattern dictionary is not needed. However, the capability to substitute a constituent in an example changes from one example to the next, and to automate this judgement is impossible. This problem could be addressed by manually tagging each example beforehand to specify which constituents can be substituted, but the resulting method would be just another *pattern-based* translation method.

Attention has been focused on the use of *cognitive grammar*(Langacker, 1987) and *construction grammar*(Fillmore, 1988) in the search to find methods that might help to resolve this problem. However, the standards for determining the structural meaning units and the granularity needed for meaning analysis have not been clarified.

As a method in which the syntactic structure and meaning are dealt with as an integral whole, a sentence pattern (SP)-dictionary called *A-Japanese Lexicon* has already been developed for Japanese simple sentences(Ikehara et al., 1997). This dictionary includes 14,800 valency patterns. The translation quality of Japanese simple sentences into English was 90% and this could be improved up to 97% by additional pattern pairs(Kanadechi et al., 2003).

Therefore in this study we developed an SP-dictionary for translating compound and complex

sentences. First we proposed a *Non-compositional Language model* (*NL-model*) and a method for creating sentence patterns. Based on these, we built a large scale SP-dictionary from a parallel corpus through the generalization of compositional constituents.

2 Language Model

Conventional MT methods are based on the concept of *compositional semantics*. However, real languages have many expressions to which this concept cannot be applied. Solving this problem requires finding a mechanism acquiring the meaning of entire expressions before their constituents are analyzed.

2.1 Expressions and Constituents

Arita(1987) has pointed out that humans employ a framework of expressions (*semantic structure*) in their mother tongue in the process of conceptualizing objects. In the semantic structure that come to mind during the process when a speaker is forming a concept, two types of constituents to be considered those that cause the overall meaning to be lost when other constituents are substituted for them, and those that do not cause the overall meaning to be lost when an alternative constituent is substituted for them. Based on this idea, we derived the following definitions.

Definition 1: Types of constituents:
A compositional constituent (*C-constituent*) is defined as the constituent for which there are one or more alternative constituents and for which *the meaning of a semantic structure* does not change when this constituent is substituted. Any other constituent is defined as non-compositional constituent (*N-constituent*).

Definition 2: Types of expressions:
A compositional expression (*C-expression*) is defined as an expression that consists entirely of C-constituents, and a non-compositional expression (*N-expression*) is defined as an expression that has one or more N-constituents.

Before we applying these definitions to actual linguistic expressions, we need to clarify what we mean by "*the meaning of a semantic structure*." This is very important problem for semantic analysis, because the guranularity needed for semantic analysis is determined by the way of the meaning defiition.

In this study, considering applications to Japanese-to-English MT, the meaning of Japanese semantic structures defined in terms of English semantic structures.[1]

Figure 1 shows an example. The source text is a Japanese expression expressing a relationship between two events: "*directly after some event happened, somebody performed some action*," and this meaning is defined by the English expression. For individual constituents such as "she" and "college," there are domains of substitutable constituents with which they can be substituted without changing the English semantic structure, so these constituents are classified into C-constituents.

2.2 Characteristics of C-constituents

From the above definitions, we can see that a C-constituent possesses the following characteristics. From these characteristics, possible guidelines for pattern-forming can be obtained.

#1: The number and the scope of C-constituents depends on the language used for defining the meanings of expressions.

#2: C-constituents need to be independent of each other.

#3: The domain of alternatives for a C-constituent is syntactically and semantically limited.

#4: Whether a constituent is compositional depends on the way it is articulated.

#5: C-constituents are defined in relation to the entire expression. Many times these expressions consist of plural words, and some of them are N-expressions.

2.3 Non-compositional Language Model

According to definition 1, any linguistic expression consists of zero or more C-constituents and one or more N-constituents. The scope of these constituents can be arbitrarily selected. Then, we assume that C-constituents are extracted from expressions with a meaningful range (e.g., a word, phrase or clause). The C-constituent extracted in

[1] In this way, when the meaning of a linguistic expression is defi ned in another natural language, semantic ambiguity occurs in the language used in the defi nition. However, in the case of MT, the meaning of the translation results is understood by a speaker of the target language, so it is not thought to constitute a problem.

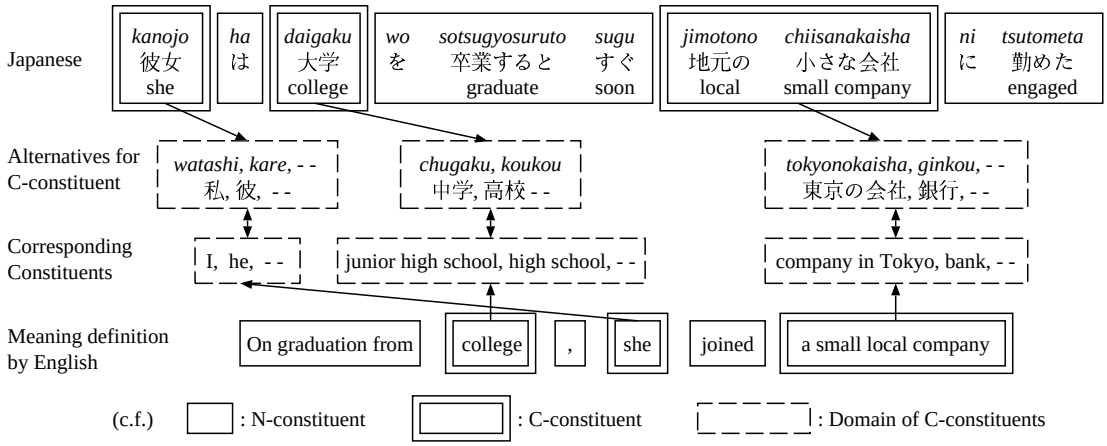

Figure 1: Corresponding Relationships of C-constituents between Japanese and English

this way may itself also be a N-expression according to characteristic #5, so a linguistic expression can generally be expressed with the language model designated as the *NL-model* in Figure 2.

As this figure shows, when C-constituents are repeatedly extracted from an N-expression, the end result is an N-expression that contains no C-constituent. Although the resulting N-expression may just be a single word, it could also be an idiomatic phrase that has no substitutable constituents. Thus, in *NL-model*, linguistic expressions can finally be articulated into N-constituents and N-expressions.

The difference from conventional *Compositional Language model* (*CL-model*) is in that *NL-model* does not assumes that all of expressions can be articulated into only C-constituents[2].

3 Pattern Forming

3.1 Principles of Pattern-forming

Not only words but also meaningful expression units such as phrases and sentences represent concepts, and the semantic structures of these expressions represent higher order concepts(Ikehara, 2003). Then, if we develop a pattern dictionary of semantic structures for expressions, higher order concept can be taken out of each expression by pattern matching.

An important aspect of the *NL-model* is that the N-expressions that appear at each stage of the decomposition are meaningful expression units. In this process, loss of the original meaning can be avoided by using a semantic dictionary for N-expressions at each stage. For example, if linguistic expressions are classified into sentences, clauses and phrases, and pattern dictionaries are prepared for N-expressions at each of these levels, then this would provide a mechanism for scooping up the meaning of entire sentences.

We think that patterns are a suitable framework for expressing the structure of N-expressions, because:

(a) N-constituents cannot be replaced by any other constituent, so a literal description is appropriate,

(b) the order of constituents is not flexible and often fixed.

Therefore, in this study we use a pattern- forming approach for N-expressions.

The Japanese-to-English parallel corpus is a typical example in which the meaning of Japanese expressions is defined by English expressions. Pattern pairs were therefore produced for N-expressions in the parallel corpus by extracting C-constituents and generalizing them.

When a parallel corpus is used, the following two types of constituents need to be considered as C-constituents:

(1) the constituent to which there is a semantically corresponding constituent in the English expression,

(2) the constituent to which there is no corresponding constituent in English, but deleting

[2]Translations by matching larger text unit are used in conventional MT including recent phrase-based SMT. But it is only that they regard phrases as C-constituents. The fact remains that entire expression is regarded as C-expression.

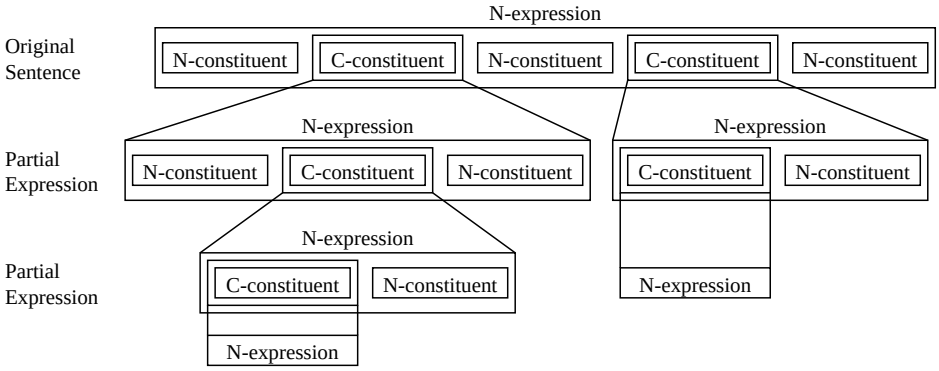

Figure 2: Non-compositional Language Model (*NL-model*)

this constituent from the Japanese expression does not cause any change in the corresponding English expression.

3.2 Pattern Description Language

A pattern description language was designed to achieve the following aims:

a) patterns can be semi-automatically generated from the result of morphological analysis of a parallel corpus, and

b) patterns can be defined according to the degree of generalization.

Here, a) is important in that it allows the development of large-scale SP-dictionaries. This condition also means that there is no need for syntactic or semantic analysis when comparing the input text with the SP-dictionary. If these analyses were required, problems would occur due to vagueness of interpretation, and the significance of the pattern method would be halved.

The descriptors used in this language are shown in Table 1. They are divided into four classes: *literals*, *variables*, *functions*, and *symbols*. As a rule, C-constituents are declared using *variables*(independent constituents), *functions*(subordinate constituents), and *symbols*(structural constituents), while N-constituents are declared by using the *literal* of words.

Constituents of a pattern are classified as either essential or optional. An essential constituent is one without which it would be impossible to define a corresponding relationship with the English pattern. An optional constituent is one that allows a corresponding English pattern to be defined even when it is omitted.

Optional constituents are further classified into two types. One type contains constituents that can be omitted from both the English and Japanese expression. The other contains constituents that do not appear in either the English or the Japanese expression, but these constituents can be inserted into those expressions.

3.3 Generalization of C-constituents

(1) Generalization by Variables

According to the characteristics of #3, when generalizing C-constituents by *variables*, the domain of a *variable* needs to be specified by syntactic attributes and semantic attributes. A syntactic attribute is represented by a variable name and a semantic attribute is represented by an argument attached to the variable. Attentions need to be paid to the following points in this generalization.

(a) A C-constituent may itself be an N-expression.

(b) Corresponding English constituents do not necessarily have the same syntactic attributes.

(2) Generalization by Functions

Conjugated forms of words that have been converted into *variables* are specified by *a word form function* when necessary. Also, particles and subordinate words such as auxiliary verbs are specified using *a Tense aspect modality function* when necessary.

(3) Generalization by Symbols

Fluctuations in expressions are also C-constituents. These are declared using *symbols*.

Symbols are also used to specify omitted subjects and objects when necessary. The optional constituents mentioned in section 3.2(2) are also C-constituents and word order that do not change

Table 1: Descriptors of Pattern Description Language

#	Classification	Types (The number of descriptors)
1	*Literal*	Japanese Character, English Character
2	*Variable*(17)	*Word variable*(11), *Phrase variable*(5), *Clause variable*(1)
3	*Function*(151)	*Word form function*(33), *Tense aspect modality function*(56), *Part of speech transfer function*(8)，*Macro function*(20)，*Group function*(9), *Extraction function*(2), *Literal function*(2), others
4	Symbol(10)	*Insertion mark*, *Optional mark*, *Permutation mark*, *Changeable position mark*, *Supplementation mark*, Others

c.f. For futher details, look at pattern pair examples shown in Table 2.

Table 2: Examples of Generated SPs

Word-level SP		
Example sentence	J	*ukkarishite* *teikikenwo* *ieni* *wasuretekita* うっかりして 定期券を 家に 忘れてきた。 by mistate season ticket at home left
	E	I was so careless as to leave my season ticket at home.
Sentence Pattern	J	*ha* *te* *wo* *ni* #1[N1(4)は]/V2(3003)て/N3(932)を/N4(447)に/V5(1809).tekita。
	E	I was so AJ(V2) as to V5 #1[N1_poss] N3 at N4.
	c.f.	(a) N1, N3, N4: *Noun Variables*, (b) V2, V5: *Verb Variables*, (c) (4), (3003): Semantic attribute numbers specifying semantic constratints on a variable, (d) #1[...]: Omissible constituents (*Optional mark*) (e) /: constituents that is able to appear in the input sentence (*Insertion mark*), (f) .tekita: Function for specifying a predicate suffix (*Tense aspect modality function*), (g) AJ(V2): Adjective form of the value of *verb variable* V2 (*Part of speech transfer function*), (h) N1_poss: Value of N1 transformed into possessive case (*Word form function*)
		Phrase-level SP
Example sentence	J	*sonoketsuronha* *ayamattazenteini* *motoduiteiru* *nodakara* *ayamaridearu* その結論は 誤った前提に 基づいている のだから 誤りである。 the conclusion false premise based on because be wrong
	E	The conclusion is wrong in that it is based on a false premise.
Sentence Pattern	J	*ha* *ni* *nodakara* NP1(1022)は/V2(1513).ta/N3(2449)に/V4(9100).teiruのだから/N5(1453).dantei。
	E	NP1 is AJ(N5) in that it V4 on AJ(V2) N3.
	c.f.	(a) NP1: *Noun phrase variable*
		Clause-level SP
Example sentence	J	*soreha* *kiwameteyuudokudearu* *node* *siyouni* *atatteha* *junibunni* それは 極めて有毒である ので 使用に 当たっては 十二分に that extreamely harmfull because use upon fully *chuuisi* *nakutehanaranai* 注意し なくてはならない。 pay attention must
	E	It is significantly toxic so that great caution must be taken with its use.
Sentence Pattern	J	*node* *niatatteha* CL1(2492).tearuので、N2(2005)に当たっては/VP3(3901).gimu。
	E	so+that(CL1, VP3.must.passive with subj(CL1)_poss N2)
	c.f.	(a) CL1: *Clause variable*, (b) so+that(..., ...): *Macro function* that generates a "so that" sentence structure, (c) subj(CL): An *extraction function* that extracts the subject from the value of a clause variable

the entire meaning are also considered as structural C-constituents, and are declared with a *symbol*.

(4) Higher-level Generalization

Fine-grained generalization is performed by, for example, using *equal-value group functions* and *corresponding group functions* whereby multiple functions are grouped together, or by using *literal functions* to apply fine constraints to the domains of *variables*. The declaration of English patterns is also simplified by using *macro functions* to synthesize a wide variety of English constructions such as *so-that* constructions.

4 Development of SP-dictionary

(1) Pattern Generation Process

A parallel corpus for basic compound and complex sentences with two or three clauses was prepared, and this corpus was generalized to produce a SP-dictionary as follows.

Step 1: Creation of parallel corpus:

A parallel corpus consisting of 1 million sentence pairs was created from about 30 types of documents.

Step 2: Extraction of translation examples:

From the results of morphological analysis of this corpus, we extracted 150,000 target example sentences. Analytical errors were corrected manually.

Step 3: Pattern generation:

Using resources such as Japanese-English word dictionaries, semantically corresponding relationships were found and converted into *variables*, *functions*, and *symbols* in the following three steps:

(a) *Word-level* generalization:
Compositional independent words are converted into *word variables*.

(b) *Phrase-level* generalization:
Compositional phrases are converted into *phrase variables*.

(c) *Clause-level* generalization:
Compositional clauses are converted into *clause variables*.

For C-constituents that can be automatically recognized, the generalization is performed automatically, while cases that cannot be judged automatically are entrusted to a language analyst. Example of patterns produced in this way are shown in Table 2.

(2) Number of Generated Patterns

Table 3 shows the number of different patterns in the resulting SP-dictionary. The number of patterns was largest for *word-level* patterns, followed by *phrase-level* and *clause-level* patterns. The number of patterns created at the *clause level* was particularly small.

Table 3: Number of Generated SPs

Type of SPs	*Word Level*	*Phrase Level*	*Clause Level*	Total
Compound	61,171	39,243	18,173	118,587
Complex	48,123	32,049	5,778	85,950
Mixed Type	12,510	8,146	1,524	22,280
Total	121,904	79,438	25,475	226,817

Table 4: Ratio of C-constituents

Constituents	No. of constituents	No. of variables	Ratio of C-constituent
Words*	734,528	542,925	73.9 %
Phrases	463,636	111,359	24.0 %
Clauses	267,601	39,718	14.8 %

* : Independent words such as nouns and verbs.

(3) Ratio of C-constituents

Table 4 shows the number of constituents converted into *variables* at each level of generalization. The ratio of generalized C-constituents was 74% at the *word level* and 24% at the *phrase level*, but just 15% at the *clause level*. This means that most of the clauses in the parallel corpus are N-constituents, which are impossible to generalize.

Acoodingly a semantically suitable translation as found in a parallel corpus cannot be obtained when N-constituents are extracted, translated, and incorporated into an original sentence.

5 Coverage of SP-dictionary

In the Pattern method based on NL-model, it is very important to know whether a pattern dictionary that cover most of the semantic structures can be developed.

(1) Experimental Conditions

We produced a program (pattern parser) to compare input sentences against the SP-dictionary, and used it for evaluations.

Ten thousand sentences that we randomly selected from the example sentences were used for creating the patterns and used in experiments. Since the input sentences will always match the patterns from which they were created, experiments were conducted in the manner of *cross-validation*.

Normally more than one pattern matches to an input sentence, and not all of them are necessarily

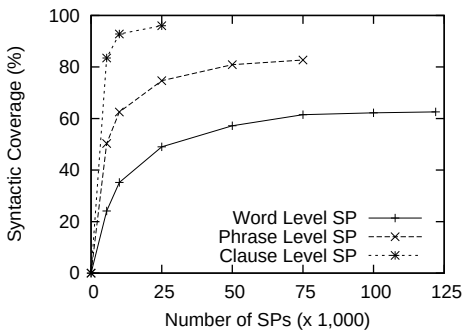

Figure 3: Saturation of Syntactic Coverage

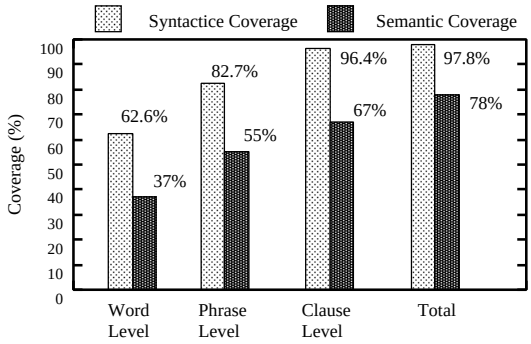

Figure 4: Syntactic and Semantic Coverage

correct in semantics. Therefore, the coverage was evaluated according to the following two parameters:

- **Syntactic coverage:** The ratio of input sentences that are matched to at least one pattern.
- **Semantic coverage:** The ratio of input sentences for which there is one or more semantically correct pattern.

(2) Saturation of Syntactic Coverage
Figure 3 shows the relationship between the number of patterns and the *syntactic coverage* at the *word*, *phrase*, and *clause levels*. As the number of patterns increases, *syntactic coverage* saturates rapidly.

In the case of simple sentences, it was reported that the number of valency patterns required to cover all of them more or less completely was estimated to be 25 thousands(Shirai et al., 1995).

When the patterns are rearranged in order of their frequency of matching, the saturation speed in Figure 3 becomes about five times faster. Then, we would expect the number of required patterns for compound and complex sentences to be somewhere in the tens of thousands or thereabouts as is the case of simple sentences.

(3) Evaluation of Coverage
Figure 4 shows the results of our evaluation of the coverage of the SP-dictionary. The results show that the whole dictionary covered almost all input sentences. However, many cases of matches to semantically inappropriate patterns occurred, and the *semantic coverage* decreased to 78% when these were eliminated.

(4) Semantic Coverage
The applicable range of the patterns is *level* smaller in the order of *word*, *phrase* and *clause level*. Accordingly, for an input sentence that matches patterns on multiple levels, to select and use the most semantically appropriate pattern based on this sequence is probably preferable.

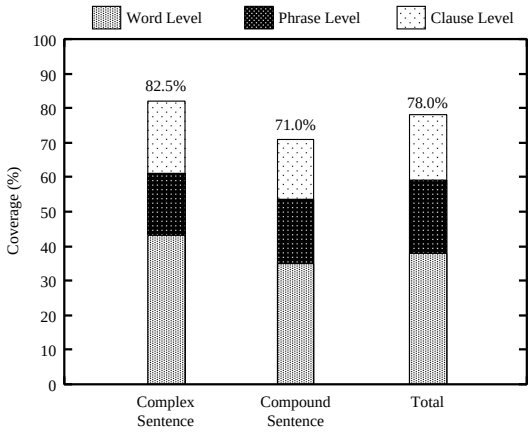

Figure 5: *Semantic Coverage* by 3 Level Patterns

Figure 5 shows the ratio of patterns that are used when they are selected based on this sequence. As this figure shows, overall *semantic coverage* is considered to be high enough for practical use. The achievement of such a large-scale SP-dictionary is unprecedented in the world.

6 Summary

We proposed a *NL-model* and a pattern forming method based on the model. We also developed an SP-dictionary of basic Japanese compounds and complex sentences that contains 227,000 pattern pairs.

The important aspects of NL-model is in that this model provides the standards for determining the structural meaning units and granularity needed for meaning analysis.

According to our evaluation, the ratios of C-constituents that could be generalized by variables were 74% for independent words and 24% for

phrases, while that for clauses was only 15%. This means that in Japanese to English MT, high quality translations for compound and complex sentences as found in a parallel corpus cannot be obtained with methods based on *Compositional Semantics*. We also found that the *syntactic coverage* of the SP-dictionary was 98%, and the *semantic coverage* was 78%. These coverages are considered to be high enough for practical use.

The SP-dictionary is aimed at the translation of N-expressions. For C-expressions, conventional MT methods can be applied, so we expect that translation quality will be substantially improved by incorporating this dictionary into conventional MT systems

The patterns can be used as a mesh for scooping up the entire meaning of expressions. We also expect that they will be used for semantic analysis in a wide range of applications other than MT.

Acknowledgements

This study was performed with the support of the Core Research for Evolutional Science and Technology (CREST) program of the Japan Science and Technology Agency (JST). Our sincere thanks go out to everyone concerned and to all the research group members who cooperated with this study.

References

Arita, Jun. 1987. *Lecture on German Language*, volume II. Nankodo Publisher.

Brown, P. F., R. John, S. D. Pietra, F. Jelinek, J. D. Lfferty, R. L Mercar, and P. S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Brown, R. D. 1999. Adding linguistic knowledge to a lexical example-based translation system. In *TMI 99*:22–32.

Fillmore, C. 1988. The mechanics of construction grammar. *Berkeley Linguistics Society*, 14:35–55.

Ikehara, Satoru, Masahiro Miyazaki, Satoshi Shirai, Akio Yokoo, Hiromi Nakaiwa, Kentaro Ogura, Yoshifumi Ooyama, and Yoshihiko Hayashi. 1997. *A-Japanese-Lexicon*. Iwanami Book Store.

Ikehara, Satoru. 2003. Concepts represented by linguistic expressions and translation. *IEICE-SIG-TL*, TL2003-25:7–12.

Jung, H., S. Yuh, T. Kim, and S. Park. 1999. A pattern-based approach using compound unit recognition and its hybridization with rule-based translation. *Computational Intelligence*, 15(2):114–127.

Kanadechi, Masato, Masato Tokuhisa, Junichi Murakami, and Satoru Ikehara. 2003. Translation quality evaluations of verbs and nouns by valency pattern dictionary. *ISPGJ-SIG-NL*, 2003-NL-153:119–124.

Langacker, R. W. 1987. *Foundation of Cognitive Grammar*. Stanford University Press.

Nagao, Makoto. 1984. A framework of a mechanical translation between japanese and english by analogy principle. In Eithorn, A. and R. Barneji, editors, *Artifi cial and Human Intelligence*:173–180. North-Holland.

Nagao, Makoto. 1996. *Natural Language Processing*. Iwanami Book Store.

Sato, Satoshi. 1992. An example based translation and system. In *COLING-91*:1259–1263.

Shirai, Satoshi, Satoru Ikehara, Akio Yokoo, and Hiroko Inoue. 1995. The quantity of valency pattern pairs required for Japanese to English MT and their compilation. In *NLPRS'95*, volume 1:443–448.

Uchino, Hajime, Satoshi Shirai, Akio Yokoo, Yoshifumi Ooyama, and Osamu Furuse. 2001. Japanese to english machine translation system for news flash - alt-flash. *IEICE Transaction*, J84-D-II(6):1168–117.

Vogel, S., Y. Zhang, F. Huang, A. Tribble, A. Venugopal, B. Zhao, and A. Waibel. 2003. The cmu statistical machine translation system. In *MT Summit IX*:402–409.