Development of Semantic Pattern Dictionary for Non-linear Structures of Complex and Compound Sentences

Satoru Ikehara[*1], Masato Tokuhisa[*1], Jin'ichi Murakami[*1]
and Masashi Saraki[*2]

[*1] Tottori University, Japan.
{ikehara, tokuhisa, murakami} @ike.tottori-u.ac.jp

[*2] Nihon University, Japan.
saraki@st.rim.or.jp

Proceedings of the ESSLLI'05 Workshop on Empirical Challenges and Analytical Alternatives to Strict Compositionality

Heriot-Watt University
Edinburgh, Scotland

Frank Richter and Manfred Sailer (Editors)

2005

http://www.macs.hw.ac.uk/esslli05/

Abstract

Semantically Classified Sentence Pattern Dictionary has been compiled on the basis of *Semantic Typology* in order to develop an *Analogical Mapping Method* for MT. This dictionary includes 221,563 *Semantic Patterns* which have been generated from Japanese compound and complex sentences. The patterns have been made up in the semi-automatic manner using a set of variables (of full words) and functions (expressing aspect, tense, and modality). In the particular pattern, the literal remainders, however, exists including not only functional words but also *non-linear* portions which are untranslatable to the target language in the linear sequence of MT. The dictionary comprises *word-level*, *phrase-level* and *clause-level*. *Non-linear structures* of Japanese sentences having two or three predicates have been extracted from a parallel corpus including a million pairs for Japanese and English sentences. The suitable definition of the *linearity* and *non-linearity* of linguistic expressions has enabled the semi-automatic pattern generalization process and the efficient development of the pattern dictionary. Our experimental evaluations showed that this dictionary semantically covers 74% of compound sentences and 67% of complex sentences, and the development cost was reduced to one-tenth that of a human intensive development.

1 Introduction

Three years ago, we started the 5-year project to develop *Semantically Classified Sentence Pattern Dictionary (SP-dictionary),* in order to realize a new MT method named *Analogical Mapping Method (AM-method).* This project is conducted under the funding of the Japan Science and Technology Agency and have developed the first version of the *SP-dictionary*. This paper will give the outlines of *AM-method* and the report of the process and results in the *SP-dictionary* development.

A huge investment has been made in the research and development of MT technology, resulting in some noteworthy achievements (Nagao, 1996). However, it is more difficult to develop MT systems between languages belonging to different families alienated from each other, such as Japanese and English, and this development of the particular system requires even further effort to improve the quality and accuracy of the output.

One method for solving this problem is *Pattern-based MT* (Takeda, 1996a,b; Watanabe and Takeda, 1998). This problem-solvimg has already been used in many commercial systems combining the *Transfer-method* and *Translation-memory* (Nagao et al., 1998) since they are adequate technique of acceptable translations for matched sentences. However, the number of prepared patterns is too small to cover general expressions so that they are only used in the translations for special fields or for translation help. One of the reasons for this limitation is the high cost of developing large-scale pattern dictionaries, although the major reason is the difficulty of defining semantically consistent sentence patterns. Though there is a lot of research on SP-learning technology (Allmuallim et al., 1994; Güvenir and Cicekli, 1998; Kitamura and Matsumoto, 1996), it is a long way from being actually used.

To address such problem, a *Multi-Level-Translating Method* (MLTM) (Ikehara et al., 1987) has provided an approach for grasping the relationship between structures and meanings in linguistic expressions, which will give a solution for breaking through the limitations of the traditional approach based on the *compositional semantics*. The implementation of the MLTM requires building up an extremely large language knowledge base by which patternized expressions can be accurately defined corresponding to the speaker's cognition of the objective world and his/her subjectivity. In the first step in the constructions process, such a knowledge base as *Goi-Taikei (A-Japanese-Lexicon)*, has already been compiled (Ikehara et al., 1997) resulting in a marked improvement in the translation quality of simple sentences (Kanadechi et al., 2001).

However, the MLTM has two problems (Ikehara, 2001a,b), one of which is that the method does not always produce optimal results of translations since it gives only one output corresponding to the syntactic structure of the target language. Another one is in how it handles the semantic *non-linearity* of complex sentences with multiple coordinate clauses and compound sentences of comprising one or more subordinate clauses.

To solve the above problems, an *AM-method* (Ikehara, 2002) has recently been proposed in which fundamentals thereof can be established by the *Semantic Typology* (Arita, 1987) and *Analogically Equivalent Thinking* (Ichikawa, 1960) theories. In this method, the *non-linear* sentence structures of a source language are semantically mapped into those of a target language using a *SP-dictionary* where one or more *semantic patterns* (SPs) for the target are defined corresponding to a pattern of the source.

2 Principles of *AM-method*

The AM-method[1] provides a problem-solving approach to the aporia in the semantic analysis and semantic understanding based on *compositional semantics*. The method is constructed from two theories: The first is the *Semantic Typology Theory* proposed by Arita (1987), which suggests that conceptual cognition is accompanied by an epistemological framework under the influence of one's mother tongue. The second is the *Analogical Mapping Theory* advocated by Ichikawa (1960). According to Ichikawa, a set of SPs in the source language can be mapped to a corresponding set in the target, with the use of an analogy between them by choosing an adequate common view-point.

With the combination of these two theories, we have brought forth a heuristic

[1] Nagao proposed an *Analogical Translation Method* based on the similarities between syntactic structures and word meanings used in corpus writings (Nagao, 1984; Sato, 1997). This is considered as basis for *Pattern-based MT*. By contrast, our method notices the similarities between the concepts represented by expression structures and goes beyond the similarity in syntactic structures.

approach to semantic analysis of the semantically in-decomposable expressions, the whole meaning of which is not just the simple sums of those of their component words. Such expressions, which are referred to as *non-linearity*, are then classified as SPs under *Logical Semantic Categories* (LSC). Given a Japanese sentence, its SP is determined using pattern matching, and then mapped to the corresponding English pattern, according to which a complete sentence will be generated.

(1) Theory of *Analogical Mapping*

Ichikawa (1960) formulated the analogical reasoning in scientific discovery and then proposed his *Analogical Mapping Theory* in "*Creative Thinking*", referred to as *Theory of Equivalent Transformation*, in 1960, stating that analogical thinking lies at the core of human creativity. This theory presented a sort of model of the creative process for problem-solving, provided that different systems have a commonality, ϵ, in their events or phenomena under a certain condition C, as shown in the following equation:

$$C(A_\alpha \overset{\epsilon}{=} B_\beta) \tag{1}$$

where C is a condition, ϵ is a commonality, A_α is an event in system α, and B_β is an event in system β.

Analogical thinking refers to the process according to above equation where given an event A_α(source) in system α, a human being develops in their mind an event B_β(target) in system β which has a commonality ϵ under a condition C.

(2) *AM-method* in MT

Technical difficulties arise when the numberless individual linguistic expressions of a language are mapped onto those of another language with their meanings correctly translated. However, these numberless expressions can be reduced to a finite number of semantic units by applying above equation.

In translating expression A_α in language α into an expression B_β in language β, language β must have expression B_β which implies a concept represented by the expression A_α. This logic provides the grounds for implementing the translations between different languages based on their meanings when the commonality ϵ is considered as a concept existing in both the source and target languages.

This technique is called the *AM-method* that uses *semantic types*. The following equation (2) shows the principles of the method:

$$A_\alpha \Rightarrow C(A_\alpha) \Rightarrow \epsilon \Rightarrow C(B_\beta) \Rightarrow B_\beta \tag{2}$$

Where ϵ is a *true item* (a collection of common concepts, i.e. a member of a LSC), and C is a function to typify a linguistic expression as an appropriate basic *semantic type*.

The equation (2) is applied to a translation when $\alpha \neq \beta$, and for rewording in the same language if $\alpha = \beta$.

(3) LSC (*Logical Semantic Category*)

The *semantic types* of the two languages are mapped via the LSC. This category is a set of concepts, each of which is usually represented by a *semantic type* (a unit of an expression categorized by its meaning). The category contains a set of *true items*. *True items* constitute two types: *true items* for simple concepts (represented by single word) and those for composite concepts (represented by multiword expressions). The categories and items are based on the *Semantic Attributes* of the *Valency Patterns* defined in "*A-Japanese-Lexicon*" (Ikehara et al., 1997).

(4) Mapping of *Semantic Types*

The *semantic types* formulated in the form of patterns, named as SPs, are classified in accordance with the *true items* stored in the LSC. Thus, the SPs of the source language can be semantically corresponded to those of the target language via the same *true items*. However, some SPs relating to complex concepts will be classified into several groups. Figure 1 and 2 show an application example of *AM-method* for Japanese to English MT system.

In the translation process, the most appropriate SPs of the target language are selected from the one or more instances that semantically correspond to the SP of the source language. The most appropriate, i.e. most similar in meaning, SP is dynamically selected during translation. To achieve this goal, the *SP-dictionary* provides contextual conditions concerning intra-sentences, inter-sentences, and contexts. Next, the retrieved Japanese SP is mapped to the corresponding English SP by means of an analogical mapping mechanism provided by the LSC. Finally, the English SP is processed to generate the translated equivalent. In this process, the Japanese components stored in the *linear component list* are translated by conventional methods and allocated to the appropriate variables of the English SP.

3 SP Generation for *Non-linear Expression*

An SP is considered as an epistemological framework for conceptual cognition and is individual to each language. In many cases, the structure of this framework does not satisfy the conditions of the *semantic composition*. SPs are defined from the view point of the *linearity* and *non-linearity* of expressions as will be described in the following sections.

3.1 Method of Judging *Non-linearity*

(1) Definitions of *linearity* and *non-linearity*

The development of conventional natural language processing technologies has been supported by the principle of *semantic composition*. There have been many studies and discussions among the adherents of *compositionality* and *contextuality* (Allen, 1995; Larson and Segal, 1995; Carpenter, 1998; Platts, 1997; Green et al., 2002; Cruse, 2004; Partee, 2004; Szabó, 2005). The compositional principle is known as Frege's definition of "*The meaning of a complex expression is determined*

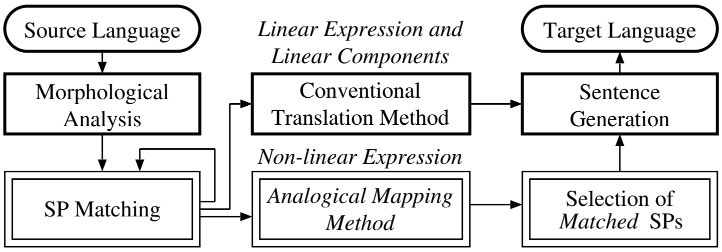
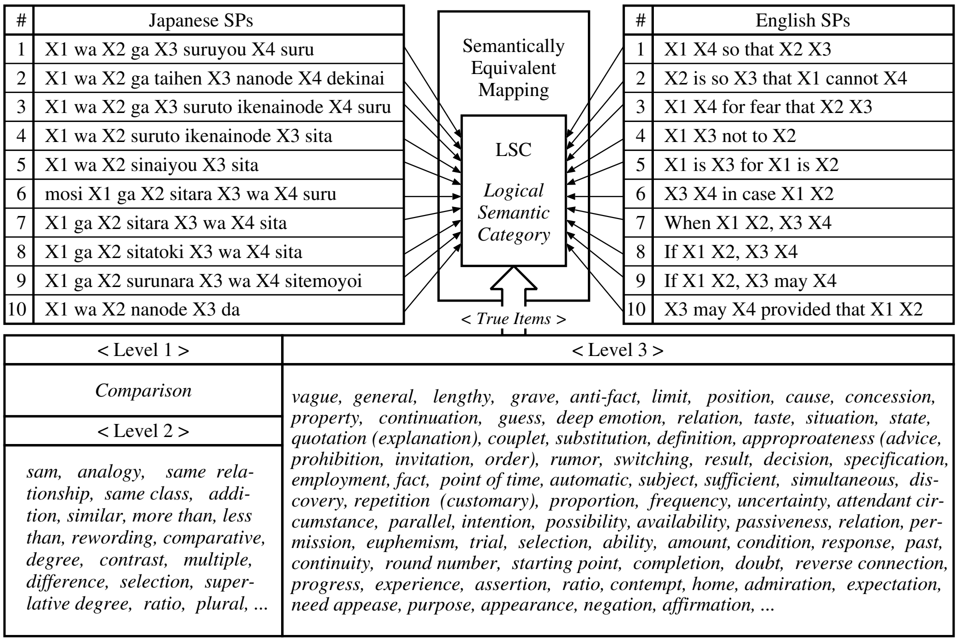


Figure 1: Translation process by *AM-method*



#	Japanese SPs
1	X1 wa X2 ga X3 suruyou X4 suru
2	X1 wa X2 ga taihen X3 nanode X4 dekinai
3	X1 wa X2 ga X3 suruto ikenainode X4 suru
4	X1 wa X2 suruto ikenainode X3 sita
5	X1 wa X2 sinaiyou X3 sita
6	mosi X1 ga X2 sitara X3 wa X4 suru
7	X1 ga X2 sitara X3 wa X4 sita
8	X1 ga X2 sitatoki X3 wa X4 sita
9	X1 ga X2 surunara X3 wa X4 sitemoyoi
10	X1 wa X2 nanode X3 da

#	English SPs
1	X1 X4 so that X2 X3
2	X2 is so X3 that X1 cannot X4
3	X1 X4 for fear that X2 X3
4	X1 X3 not to X2
5	X1 is X3 for X1 is X2
6	X3 X4 in case X1 X2
7	When X1 X2, X3 X4
8	If X1 X2, X3 X4
9	If X1 X2, X3 may X4
10	X3 may X4 provided that X1 X2

< Level 1 >	< Level 3 >
Comparison	*vague, general, lengthy, grave, anti-fact, limit, position, cause, concession, property, continuation, guess, deep emotion, relation, taste, situation, state, quotation (explanation), couplet, substitution, definition, approproateness (advice, prohibition, invitation, order), rumor, switching, result, decision, specification, employment, fact, point of time, automatic, subject, sufficient, simultaneous, discovery, repetition (customary), proportion, frequency, uncertainty, attendant circumstance, parallel, intention, possibility, availability, passiveness, relation, permission, euphemism, trial, selection, ability, amount, condition, response, past, continuity, round number, starting point, completion, doubt, reverse connection, progress, experience, assertion, ratio, contempt, home, admiration, expectation, need appease, purpose, appearance, negation, affirmation, ...*
< Level 2 >	
sam, analogy, same relationship, same class, addition, similar, more than, less than, rewording, comparative, degree, contrast, multiple, difference, selection, superlative degree, ratio, plural, ...	

Figure 2: *Semantically Equivalent Mapping* of *SPs* via *True Items*

by the meanings of its parts, and the way in which those parts are combined".

The most typical example based on the principle will be *Transfer-method* for conventional MT system. In this method, the partial meanings of the whole of an original structure are directly expressed in the converted lexical structure in the target language and then combined together with each other to generate the target language expression, assuming that the meanings of parts are given by lexicon and the combination way is given by syntax.

However, this method has reached the limits. The original meanings in a sentence in the source language are lost during the translation process and high quality translation cannot be obtained, especially in the translation between the languages of different families.

We propose pattern based method for determining the meaning of the whole expression in advance, assuming that the meaning of the whole expression cannot be determined by the parts and but the meanings of the parts can be determined by the meaning of whole expression.

Linguistic expression is a means of representing speaker's conceptual cognition. A speaker first selects the most suitable expression structure from options occurred in his/her mind to represent his/her cognition and then specifies partial expressions for each component to complete the sentence while keeping the total meaning in his/her mind.

In this process, there are two types of components: One is the components which can be replaced by alternatives in a domain without changing the entire meaning. Another is the component which cannot be replaced by any other components. Then, we discriminate the former as a *linear components* and the latter as a *non-linear components*. The *linearity* and *non-linearity* of a component and an entire expression are defined in detail as follows:

Definition 1 : *Linearity* of components

A *linear component* of an expression is a component which can be replaced by an equivalent component with no change in the meaning of the expression itself.

Definition 2 : *Linearity* of an expression

An expression composed of only *linear components* is defined as a *linear expression*. Meanwhile, an expression comprising one or more *non-linear components* is defined as a *non-linear expression*.

Definition 3 : SP (*semantic pattern*)

SP is defined as an expression in a *non-linear expression*.

From the Definition 2 and 3, it can be understood that the principle of *semantic composition* holds for *linear expressions*. Our definitions is compatible to the Frege's explanation. According to the Frege's theory, the feature of *compositionality* of logical expressions is that if any part of an equation is replaced by another equivalent component, the total value, which is the meaning of the entire expression, does not change (Allwood et al., 1977). *Linear components* correspond to *compositional components* since they are replaceable with another equivalent components without changing the meaning, but the determination of whether *decomposable components* or not cannot be made without checking it's inner structure. In contrast to this, *non-linear components* cannot replaced with other components without changing the entire meaning so that they cannot said as *compositional component*.

It is very important to notice that there is no need to develop SPs for *linear expressions*, since such expressions can be processed by the conventional method based on *semantic composition*.

(2) Definition of Meaning for Linguistic Expressions

The meaning of SP needs clarification for the application of the above definitions to actual sentences. Considering the practical way of defining the meaning for an actual expression, a description has no more significance to a computer more than a symbol, so that any description will do in so far as it is systematically defined. Hence, we describe the meaning of expressions for a source language by the expressions for a target language. This is easy and convenient way in designing a MT system.

From this definition it is assured that the *linear components* of the source expression have a semantically corresponding component in the target expression and the corresponding relationship of the entire expression does not vary with the replacement of these kinds of components. This matter establishes the principle for judging whether *linearity* or *non-linearity* with regard to an expression component. When the corresponding structure of the target expression does not change when a component of the source expression (i.e. word, phrase or clause) is replaced by alternatives, the component is judged as *linear*. Otherwise it is judged as *non-linear*.

(3) Characteristics of *linear components*

Figure 3 shows the example of *linear components*. Important aspects of the *linear component* defined above are as follows. First, although the replaceable component is defined as *linear*, it does not mean it is an unbounded replacement. It has a syntactically and semantically limited domain as shown in Figure 3.

Second, when all components are *linear*, the entire expression is defined as *linear*. However, the determination of whether *linearity* or not is dependent on the suitable selection of a component, and thus the *linearity* of the entire expression is dependent on the way in which the expression is divided into components.

Third, the *linear component* is defined in relation to the entire expression. This does not mean the *linearity* of itself. The internal structure of the *linear component* can be *non-linear* as shown in Figure 4.

Thus, the *linear components* can be separated again into *linear* and *non-linear components*, when the total expression has been separated into *linear components* and/or *non-linear*.

Above mentioned linguistic model is consistent with the "*Construction Grammar*" proposed by Fillmore (Fillmore et al., 2005). The importance of the information presented by patterns was also pointed out for the analysis of Multiword-Expressions (Baldwin and Bond, 2002; Sag et al., 2002).

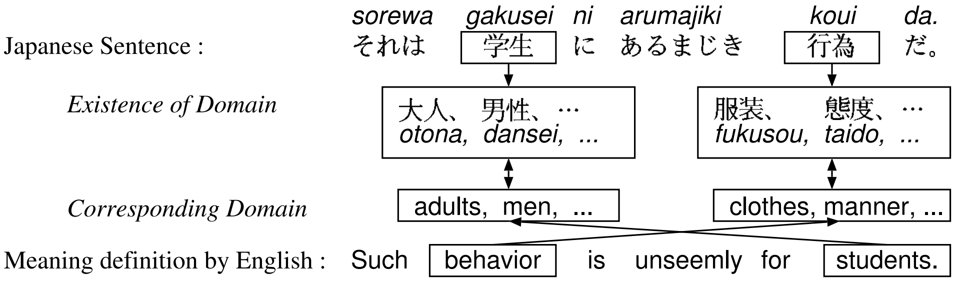


Figure 3: Example of linear components

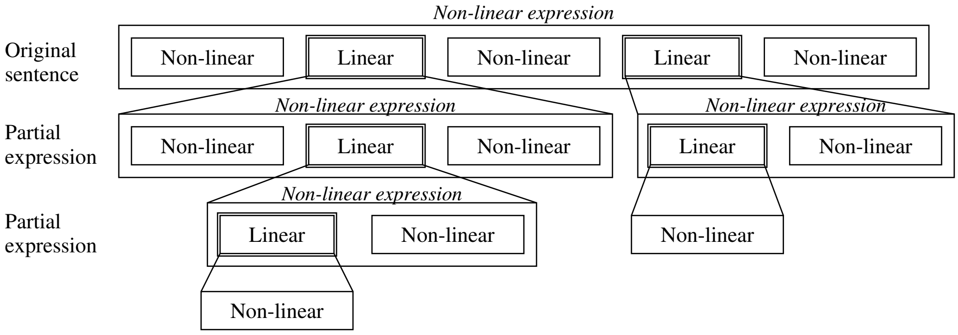


Figure 4: Recursive structure of *non-linear expressions*

3.2 Framework for defining SP

(1) SPs representing *non-linearity*

The SPs can be extracted by elimination of the *linear components* from the expressions while holding the intrinsic meaning of them. As a result of this abstraction, the *non-linear components* are retained but the *linear components* are replaced with arbitrary factors. These SPs are language-dependent. Japanese and English, for example, have their respective SPs.

The number of SPs would be finite in practice, although there are infinite variations of expressions in text and conversational speech, because a language does not have so many linguistic norms supporting the generation of SPs[2]. Therefore, it is feasible that a finite number of SPs are defined, to which the specific expressions in both languages are linked to implement the MT.

(2) *SP-Description Language*

In the development of an *SP-dictionary*, it is very important to obtain high coverage for actual expressions and semantic exclusiveness among the SPs. *SP-Description Language* (SP-DL) was developed to semi-automatically generate an *SP-dictionary* from a large-scale parallel corpus and to conduct matching *SP-dictionary* with input sentences using only morphological analysis results. Table 1 shows the constituents of SPs. The framework for the SP-DL will be described as follows:

SPs are defined using *essential* and *optional components*. The *essential* consist of *linear* and *non-linear components*: the *linear* are converted to abstract structure of *variables* and *functions*, whereas the *non-linear* are described by the same as literals in the original sentence. *Optional components*, on the other hand, are described by *symbols*. They are separated into "*hidden components*" and "*specified components*". In SPs, only the positions are defined for the former, but concrete

[2] SPs represent non-linear expressions that must be memorized to use them. Then, if the number of them is infinite, humans cannot use them freely because of their limited memory capacity. Our linguistic model will yield the answer to Plato's problem. The answer is that almost infinite linguistic expressions are generated from the recursive structure by combining the finite non-linear components as shown in the last section of this paper

Table 1: Elements for defining SPs

Classification		Explanations	
Literals	Japanese Character	Kanji, Hiragana, Katakana, Numerals, Alphabet	
	English Character	Alphabet, Numerals	
Variables (15 types)	Word Variable (9 types)	Represents *linear* full words: nouns, verbs, etc.	
	Phrase Variable (5 types)	Represents *linear* phrases: noun / verb phrases, etc.	
	Clause Variable (1 type)	Represents *linear* clauses	
Functions (107+α types)	Variable Function (8 types)	Change the syntactic attribute of variables	
	Literal Function (arbitrary types)	Check whether the literals of function name are included in the argument expression	
	Extract Function (2 types)	Subject and object extraction from phrases or clauses substituted in variables	
	Form Function (67 types)	Word Form(18 types)	Conjugation, etc.
		Others (49 types)	Tense, aspect and modality
	Sentence Generator (27 types)	Compose English sentence structure from one or more phrases or clauses	
	Macro Function (3 types)	Substitute a sentence structure with variables to an upper type variable	
Symbols (7 types)	Separator	Represents the positions for optional components	
	Continuation Mark	Represents the positions forbidding optional components	
	Component Selector	Represents a selectable component group	
	Optional Mark	Represents optional components	
	Permutation Mark	Represents permutable components	
	Changeable Position Mark	Represents removable components and positions	
	Supplementation Mark	Supplementation of erased subjects and objects	

expressions are defined for the latter.

In order to describe SPs generalized by *word-level*, *phrase-level* and *clause-level*, three kinds of variables, *word-variables* (9 types), *phrase-variables* (5 types) and *clause-variables* (1 type) are defined. Domains for these variables are semantically defined using *semantic attributes*. In the matching process with an input sentence, the matched component of the sentence is substituted to the corresponding variable. To represent synonymous words or expressions, symbols grouping the expressions with the same meaning and many different functions were prepared. The former is used not only for identifying different forms of a word but also for phrases equivalent to particles. The latter is used mainly to represent tense, aspect and modality.

The sequence of components in the matched SPs needs to be the same as those of the input sentence, in principle. However, word order for Japanese sentences is not firm. In many ways it can be permuted without changing the meaning. Therefore, a *description of arbitrary word orders* and a *description of changeable position words* were introduced.

4 SP Generations

4.1 Generation Method

(1) Examples of sentence pairs

The *SP-dictionary* has been developed for processing Japanese compound and complex sentences having two or three predicates. The reason for targeting such kinds of sentences will be described as follows:

The translation using the pattern dictionary has been achieved to the high degree (accuracy: 90%, limit of method: 98%) (Ikehara, 2001a) for simple sentences by the realization of "Goi-Taikei: *A-Japanese-Lexicon*" (Ikehara et al., 1997). But there is no semantic knowledge base for the *non-linear structures* of complex and compound sentences and translation quality still remains low.

The reason for restricting the number of predicates is as follows: In the case of sentences with 4 or more clauses, all clauses are merely *non-linear*. Many times, these sentences can be translated by separating them into plural sentences with 2 or 3 clauses.

A parallel corpus of a million sentence pairs was collected from 30 kinds of documents such as word dictionaries, handbooks for letter writing, Japanese text books for foreigners, and test sentence sets prepared for MT. A set of 128,713 applicable sentence pairs were semi-automatically extracted from them and used as example sentence pairs. Table 2 shows the types of component of speech and their number of appearance in the example sentences. The average number of words in Japanese sentences is 12.2 words.

(2) SP Generation

The example sentences are segmented by the morphological analyzer of ALT-JAWS (NTT, 2002) and the segmentation words and partial expressions of a Japanese sentence are semantically and semi-automatically brought into correspondence with those of an English sentence by using Japanese to English dictionaries. In this

Table 2: Word Appearances in Example Sentences

#	Part of Speech	Total Frequency	Different Words	Frequency / Word
1	Noun	417,886	56,861	7.4
2	Real Verb	223,178	10,324	21.6
3	Pseudo Verb	51,918	271	191.6
4	Adjective	31,681	915	34.6
5	Adjective Verb	19,587	2,562	7.6
6	Adverb	39,051	3,191	12.2
7	Adnominal	32,585	731	44.6
8	Conjunction	3,146	77	40.9
9	Interjection	147	60	2.5
10	Prefix	1068	110	9.7
11	Suffix	1749	336	5.2
12	Auxiliary Verb	165,251	236	700.2
13	Particle	465,811	349	1334.7
14	Symbol	121,555	32	3798.6
–	Total	1,574,613	76,055	20.7 / word

process, synonymous words and/or expressions are checked out by the ALT-JAWS and automatically rewritten into canonical forms. For the semantic constraints for *variables*, 2,718 types of *semantic attributes* registered in *Goi-Taikei* (Ikehara et al., 1997) and *Ruigo Daijiten* (Shibata and Yamada, 2002) are used. A newly designed semantic attribute system is used for declinable words (verbs, adjectives, etc.).

The SPs were generated in the order of *word-level* SPs, *phrase-level* SPs and *clause-level* SPs as shown in Table 3. Examples of SPs are shown in Figure 5.

It was necessary to have 13.6 person-years of analysts for the development of the *SP-dictionary*. According to the partial experiments of writing patterns by human, the cost of developing this dictionary was estimated to have reduced to one-tenth compared to the cost necessary for a solely manpower based development.

Table 3: Generalization Levels of SPs

Level	Processes of Generalization
word-level	(1) Marking of optional, **(2) Replacement of *linear words* by variables,** (3) Replacement of predicate ending by functions, (4) Designation of equivalent component groups.
phrase-level	**(1) Replacement of *linear phrases* by variables and word variables by phrase variables,** (2) Normalization of polite expressions, (3) Expansion of functional words.
clause-level	**(1) Replacement of *linear clauses* by variables,** (2) Application of the functions which transform Japanese clauses to English phrases, (3) Application of the functions creating English sentence structures.

word-level SP	
Japanese SP	#1 [$N1$(G4) は]/$V2$(R3003) て/$N3$(G932) を/$N4$(G447) に/$V5$(R1809).*tekita*。 ha te wo ni
Example	うっかりして 定期券を 家に 忘れてきた。 ukkarisite teikikenwo ieni wasuretekita
English SP	I was so $AJ(V2)$ as to $V5$ #1[$N1$_*poss*] $N3$ at $N4$.
Example	I was so careless as to leave my season ticket at home.

phrase-level SP	
Japanese SP	$NP1$(G1022) は / $V2$(R1513).ta / $N3$(G2449) に / ha ni $V4$(R9100).*teiru* のだから / $N5$(N1453).*dantei*。 nodakara
Example	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。 sonoketsuronwa ayamattazenteini motoduite irunodakara ayamaridearu
English SP	$NP1$ is $AJ(N5)$ in that it $V4$ on $AJ(V2)$ $N3$.
Example	The conclusion is wrong in that it is based on a false premise.

clause-level SP	
Japanese SP	$CL1$(G2492).*teiru* ので、$N2$(G2005) に当たっては/$VP3$(R3901).*gimu* node niatatteha
Example	それは 極めて 有毒であるので、 使用に当っては sorewa kiwamete yuudokude arunode siyouniatattewa 十二分に 注意しなくてはならない。 juunibunni chuuisinakutehanaranai
English SP	*so*+*that*($CL1$,$VP3$.*must.passive* with subj($CL1$)_*poss* $N2$)
Example	It is significantly toxic so that great caution must be taken with its use

c.f. G#:Semantic Attribute Number defined by *A-Japanese-Lexicon* (Ikehara et al., 1997).
R#:Semantic Attribute Number defined by *Ruigo Daijiten* (Shibata and Yamada, 2002).

Figure 5: Examples of Generated SPs

5 Statistics of SP-dictionary

5.1 Quantity of Generated SPs

The number of different SPs are shown in Table 4. The original number of SPs was 245,721 in total but they include 24,158 of the same SPs. The ratios of the same SPs were 5%, 16% and 12% for each level. Then, the number of different SPs was reduced to 221,563. The ratios of the numbers of *word-level*, *phrase-level* and *clause-level* SPs to the example sentences are 99.5%, 81.3% and 10.1%.

The number of *clause-level* SPs is much smaller than that of the example sentences. This smaller number means that most of the clauses in the example sentences have *non-linearity* which makes much difficult to convert the expression to the target language. Hence the MT methods based upon *compositional semantics* cannot deliver the expected results of high quality translations as shown in the example.

Table 4: The Number of Different SPs

Sentence Type	No. of Predicates	Explanation	No. of Example Sentence	Generated Sentence Patterns			
				word level	*phrase level*	*clause level*	Total
Type 1	2	1 conjugation	57,235	53,578	37,356	5,521	96,455
Type 2	3	2 conjugation	6,196	6,080	4,952	417	11,449
Type 3	2	1 embedding	46,907	44,008	30,932	3,185	78,125
Type 4	3	2 embedding	5,986	5,889	5,084	811	11,784
Type 5	3	1 conj. + 1 emb.	12,389	12,174	10,025	1,551	23,750
—	–	Total	128,713	121,729	88,349	11,485	221,563

5.2 The Ratio of *Linear* and *Non-linear Components*

(1) Frequency of Variables

Table 5 shows the types and the frequency of the variables used in SPs.

The analysis of the frequency of variables will be described as follows: The total number of full words in the example sentences was 763,968. Out of those, there were 472,521 *word variables*. The ratio of the full words replaced by variables was 62%. Out of 5.9 words per sentence, 3.7 full words were replaced by *word variables* as *linear components*, and thus 2.2 full words kept literals as *non-linear components*. Meanwhile the number of phrases replaced by *phrase variables* was 102,000. In contrast to the word and phrase variable replacements, the number of clauses replaced by variables was only 11,580 (4.3%) out of 267,601 clauses.

Compared to full words and phrases, the *linearity* of clauses was extremely low. This fact shows that a Japanese complex or compound sentence are often translated into simple English sentences. Therefore, high-quality translations, as shown in the example, cannot be expected using conventional MT methods based on *compositional semantics*.

Table 5: Frequency of Variable used in SPs

Type of Variables	Type of SP			
	word-level		*phrase-level*	*clause-level*
Noun (N)	303,319	(417,886)	138,033	10,135
Time Noun ($TIME$)	8,527		5,187	529
Numeral (NUM)	6,036		2,314	189
Verb (V)	101,484	(223,178)	48,036	4,254
Adnominal (REN)	21,241	(32,585)	2,158	127
Adverb (ADV)	11,491	(39,051)	7,631	603
Adjective (AJ)	10,950	(31,681)	6,193	425
Adjective Verb (AJV)	9,473	(19,587)	6,273	434
Sub-total for Word Var.	472,521	(763,968)	215,825	16,696
Verb Phrase (VP)	—		58,908	2,838
Noun Phrase (NP)	—		40,629	1,985
Adjective Phrase (AJP)	—		1,341	78
Adjective Verb Phr. ($AJVP$)	—		935	37
Adverb Phrase ($ADVP$)	—		117	8
Sub-total for Phrase Var.	—		101,930	4,946
Clause (CL)	—		—	11,580 (267,601)
Total	472,521		317,755	21,942
No. of SPs	121,729		88,349	11,485
No. of variables / SP	3.88 / SP		3.60 / SP	1.91 / SP

c.f. (nn,nnn) = No. of appearance of words in the original sentence

Table 6: Average number of the functions used in SP

Type of Function	*word-level*	*phrase-level*	*clause-level*	Total
Tense	33,660	33,675	5,798	73,133
Aspect	13,642	15,598	3,183	32,423
Modality	38,952	38,923	6,514	84,389
Total	86,254	88,196	15,495	189,945
No. of SPs	121,729	88,349	11,485	221,563
No. of Functions / SP	0.709/SP	1.00/SP	1.35/SP	0.86/SP

(2) Frequency of Functions

The average number of the functions used in SP is shown in Table 6. The frequency of function use in the three levels were 86,295, 88,193 and 15,495 respectively. This corresponds to 0.7, 0.95 and 1.5 per SP. It can be observed that generalization has progressed with the level of SPs.

5.3 Discussion

Out of the example sentence pair, 302 sentences (0.23%) had not any *linear component* to be replaced by a variable or a function and most of the example sentences (more than 99%) had one or more *linear components*. The former sentence pairs were kept as literal patterns.

On the other hand, 15 SPs in *word-level*, 401 SPs in *phrase-level* and 155 SPs in *clause-level* had no literal element. Only these are SPs for *linear sentences* defined by 3.2 (2) (see "definition 2"). Then it can be seen that most of complex and compound Japanese sentences are non-linear expressions that are difficult to translate into English by the method of *Semantic Composition*.

But, it is very important to notice that most of these sentences have one or more *linear components* (on average 4-5 components). This implies the capability of developing the *SP-dictionary* with high coverage. Pattern translation method will be expected to overcome the limitation of *Example-based* MT.

6 Evaluation of Coverage and Precision

The most important parameters for evaluating *SP-dictionary* will be coverage for input sentences and semantic exclusiveness of the SPs retrieved from the dictionary. In this section, we will evaluate *Matched Pattern Ratio* and *Precision* for the matched SPs.

6.1 Evaluation Conditions

As one of the method to realize semantic exclusiveness, selectional restriction has been realized. The domains of *variables* are restricted by using semantic attribute system. But, there are many ways to select the correct SPs for input sentences when retrieved SP candidates for an input sentence contain one or more correct SPs. Our experiments showed that correct SPs can be find by the accuracy of more than 90% by using *Multivariate Analysis*. Then, the experiments were conducted neglecting semantic attributes given to variables and coverage were obtained.

The experiments were conducted in the manner of *Cross Validation*. 10,000 input sentences were randomly selected from the original example sentences, so that any input sentence is assured to match the pattern that had been obtained from itself. Therefore such pattern were excluded from matched patterns and coverage for the *SP-dictionary* was evaluated using a *Matched Pattern Ratio* and *Precision* as follows.

***Matched Pattern Ratio* ($P0$):** The ratio of input sentences that have one or more matched SPs

***Precision* ($P1$):** Semantically-correct ratio of the matched SPs (corresponding to a random selection method)

***Accumulative Precision* ($P2$):** The ratio of matched SPs containing one or more semantically-correct candidates (corresponding to the most suitable candidate selection method)

Matched Pattern Ratio means syntactic coverage. Matched SPs yield the results of syntax analysis but do not always yield semantically-correct translations. Semantically correct candidates, on the other hand, assure semantically-correct translations. Thus, $P0 \times P2$ represents semantic coverage of the *SP-dictionary*.

6.2 Evaluations of *Matched Pattern Ratio*

(1) Saturation of Coverage

The relationship between the *Matched Pattern Ratio* ($P0$) and the number of SPs were evaluated (Figure 6). $P0$ tends to saturate in the tens of thousands of SPs. Effective coverage cannot be obtained by less than ten thousand SPs. Several tens of thousands of SPs will be necessary for an actual use.

(2) Coverage of *SP-dictionary*

$P0$ for *word-level*, *phrase-level*, and *clause-level* SPs are shown in Table 7. In this table, "*entire match*" means the ratio that one or more entirely matched SPs were found for an input sentence. "*Partial match*" means the ratio that there were one or more patterns, the matching conditions of which were satisfied by the input sentence but there were additional components in it.

In the case of *word-level* SPs, *entire match* ratio is low compared with that of "*partial match*". Coverage of *phrase-level* SPs is the highest and most promising. Compared to this, that of *clause-level* SPs is not high. This is because of the low number of SPs.

(3) Number of Matched Patterns

Many times, one or more SP matched to an input sentence. Also, the way a SP matches the input sentence is not always limited to one. The number of matched SPs per input sentence is shown in Table 8.

From this table, it is found that many SPs matched to an input sentence and also there are some matching ways for a SP. These are remarkable for *phrase-level* SPs.

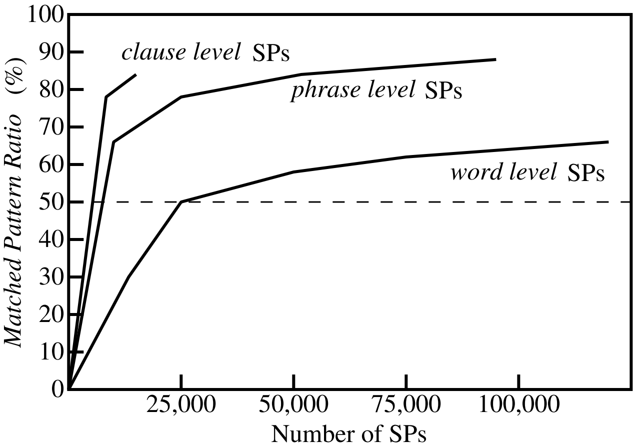


Figure 6: Saturation of *Matched Pattern Ratio* ($P0$)

Table 7: *Matched Pattern Ratio* of *SP-dictionary*

Level of SP	*entire match*	*partial match*	*Matched Pattern Ratio* ($P0$)
word-lv.	15.1 %	50.9 %	66.0 %
phrase-lv.	50.0 %	40.0 %	89.9 %
clause-lv.	44.2 %	40.3 %	84.5 %
Total	56.2 %	35.6 %	91.8 %

Table 8: Number of Matched Patterns per input Sentence

Level of SP	No. of Matched SPs	No. of Total Matchings	Matchings per SP
word-lv.	17.1	31.9	1.9
phrase-lv.	68.4	283.8	4.1
clause-lv.	12.1	57.9	4.8

(For the case of input sentences which have matched SPs)

6.3 Evaluations of Precision

(1) Evaluation Results

The results of $P1$ and $P2$ are also shown in Table 9. Compared to $P1$, $P2$ is a few times higher. This means that the matched SPs contain many incorrect candidates.

(2) Capability of Correct Translations

Although *word-level* SPs will assure high-quality translations, the coverage is small because of the high individuality. Meanwhile, the coverage of *phrase-level* SPs and *clause-level* SPs are high, but their translation quality will not be as accurate compared to *word-level* SPs. Then, *word-level*, *phrase-level* and *clause-level* order will be suitable to use for the matched SPs of an input sentence. The ratios for each level of SP used for the translation are shown in Figure 7.

This figure shows that 67-74% of input sentences can be translated directly using the *SP-dictionary*. As previously mentioned, SPs are defined for *non-linear sentence structures*, in principle. If we leave the translation of *linear sentence structures* to a conventional MT method, a 67-74% semantic coverage will be very effective. However, there are many possibilities of a further improvement in the semantic coverage. We are now going to try a further generalization for tense, aspect and modality to achieve a semantic coverage of 80-90%.

Table 9: Evaluation Results for Precision

Level of SP	*Precision* ($P1$)	*Accumurative Precision* ($P2$)
word-lv.	30.5%	69.0%
phrase-lv.	24.4%	66.2%
clause-lv.	13.8%	52.2%

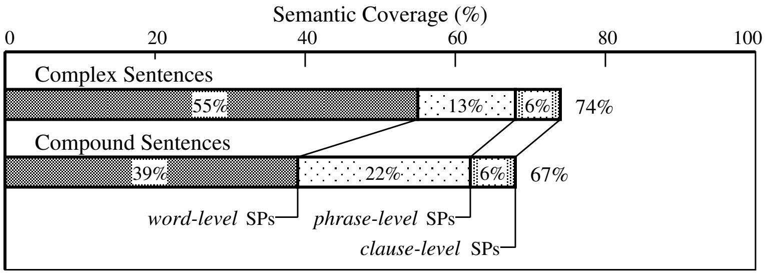


Figure 7: Semantic Coverage of *SP-dictionary*

7 Conclusion

In order to realize the *AM-method* for MT, the *SP-dictionary* for complex and compound sentences was developed and the quality was evaluated. This dictionary includes 221,563 SP pairs consisting of three kinds of SPs: *word-level* (121,729 pairs), *phrase-level* (88,349 pairs) and *clause-level* (11,485 pairs).

This dictionary was semi-automatically generated from 128,713 example sentence pairs, which were extracted from a one million sentences parallel corpus of Japanese-to-English translations.

The suitable definition of the *linearity* and *non-linearity* of linguistic expressions has enabled the semi-automatic pattern generalization process. Thus, the development cost was reduced to one-tenth that of a human intensive development. From the analysis of these SPs, it was clarified that the ratios for *linear components* were 62% for full words, 22% for phrases, and 4.3% for clauses.

These results shows the following concluding remarks: many *non-linear components* exsist in actual sentences and most of clauses are *non-linear*, which means that high-quality translations cannot be expected by using conventional MT methods based on *compositional semantics* and thus that it is very important to develop the method for dealing with *non-linear expressions*.

Matched Pattern Ratios of SPs were 66.0% for *word-level*, 89.9% for *phrase-level*, and 84.5% for *clause-level* SPs. It was also found that 74% of complex sentences and 67% of compound sentences are expected to be translated directly by the *SP-dictionary*. This dictionary leaves room for further generalization particularly for tense, aspect and modality.

We will report the evaluation results for the *AM-method* in the near future.

Acknowledgements

The *Japan Science and Technology Agency* (JST) conducted this research as the *Core Research for Evolutional Science and Technology* (CREST) project. We wish to thank the members of the research group and the language analysts for their assistance.

References

Allen, J. 1995. *Natural Language Understanding*. The Benjamin/Cummings Publishing Company.

Allmuallim, I., Akiba, Y., Yamazaki, T., Yokoo, A. and Kaneda, S. 1994. Two Methods for Learning ALT-J/E Translation Rules from Examples and a Semantic Hierarchy. In *Proc. of COLING*, volume 1, pages 57–63.

Allwood, J., Andersson, L. and Dahl, O. 1977. *Logic in Linguistics*. Cambridge University Press.

Arita, J. 1987. *Lecture of Germany*, pages 48–56. Nan-undo Publisher.

Baldwin, T. and Bond, F. 2002. Multiword Expressions: Some Problems for Japanese NLP. In *Proc. of Annual Meeting of the Association for Natural Language Processing*, pages 379–382.

Carpenter, B. 1998. *Type-Logical Semantics*. MIT Press.

Cruse, A. 2004. *Meaning in Language: An Introduction to Semantics and Pragmatics*. Oxford University Press, second edition.

Fillmore, C., Kay, P., Michaelis, L. A. and Sag, I. A. 2005. *Construction Grammar*. Stanford Univ Center for the Study.

Green, R., Bean, A. A. and Myaeng, S. H. (eds.). 2002. *The Semantics of Relationships An Interdisciplinary Perspective*. Kluwer Academic Publishers.

Güvenir, H. A. and Cicekli, I. 1998. Leaning Translation Templates from Examples. *Information Systems* 23(6), 353–363.

Ichikawa, K. 1960. *Methodology for Creative Research*. Sanwa Shobo.

Ikehara, S. 2001a. Challenge to the Fundamental Problems on Natural Language Processing. *Japanese Society of Artificial Intelligence* 16(3), 422–430.

Ikehara, S. 2001b. Meaning Comprehension Using Semantic Patterns in a Large Scale Knowledge-Base. In *Proc. of PACLING*, pages 26–35.

Ikehara, S. 2002. Toward the Realization of Ultimate MT Method = MT Method based on Analogical Thinking =. *AAMT Journal* (32), 1–7.

Ikehara, S., Miyazaki, M., Shirai, S. and Hayashi, Y. 1987. Recognitions and Multilevel Machine Translation Method based on It. *Journal of IPSJ* 28(12), 1269–1279.

Ikehara, S., Miyazaki, M., Shirai, S., Yokoo, A., Nakaiwa, H., Ogura, K., Ooyama, Y. and Hayashi, Y. 1997. *Nihongo Goi-Taikei (A-Japanese-Lexicon)*. Iwanami Publisher.

Kanadechi, M., Ikehara, S. and Murakami, J. 2001. Evaluation of English Word Translations for Japanese Verbs using Valency Patterns. In *Proc. of Annual Meeting of IPSJ*, volume 2, pages 267–268.

Kitamura, M. and Matsumoto, Y. 1996. Automatic Extraction of Word Sequence Correspondence in Parallel Corpora. In *Proc. of Workshop on Very Large Corpora*, pages 79–87.

Larson, R. and Segal, G. 1995. *Knowledge of Meaning - An Introduction to Semantic Theory -*. MIT Press.

Nagao, M. 1984. *A Framework of a Mechanical Translation between Japanese and English by Analogy Principle*, pages 173–180. North-Holland.

Nagao, M. 1996. *Natural Language Processing*. Iwanami Publisher.

Nagao, M., Kurohashi, S., Sato, S. and Nakano, H. 1998. *Sience of Natural Language*, volume 9, Chapter Linguistic Information Processing. Iwanami Publisher, Tokyo.

NTT, Communication Science Lab. 2002. ALTJAWS: Japanese Automatic Word Separator. (http://www.kecl.ntt.co.jp/icl/mtg/resources/altjaws.html).

Partee, B. H. 2004. *Compositionality in Formal Semantics*. Blackwell Publishing.

Platts, M. B. 1997. *Ways of Meaning - An Introduction to a Philosophy of Language* -. MIT Press, second edition.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A. and Flickinger, D. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Proc. of CICLing*, pages 1–15.

Sato, S. 1997. *Machine Translation based on Analogy*. Kyoritsu Publisher, (in Japanese).

Shibata, T. and Yamada, S. 2002. *Ruigo Daijiten (A large thesaurus)*. Koudansha Publisher.

Szabó, Z. G. 2005. Compositionality. In *The Stanford Encyclopedia of Philosophy*, (http://plato.stanford.edu/archives/spr2005/entries/ compositionality/).

Takeda, K. 1996a. Pattern-based Context Free Grammars for Machine Translation. In *Proc. of ACL*, pages 144–151.

Takeda, K. 1996b. Pattern-based Machine Translation. In *Proc. of COLING*, volume 2, pages 1155–1158.

Watanabe, H. and Takeda, K. 1998. A Pattern-based machine translation system extended by example based processing. In *Proc. of COLING*, pages 1369–1373.