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

Satoru Ikehara[†1], Masato Tokuhisa[†1], Jin'ichi Murakami[†1],
Masashi Saraki[†2], Takashi Ikeda[†3] and Masahiro Miyazaki[†4]

† 1 Tottori University, Tottori-city, 680-8552 Japan. {ikehara, tokuhisa, murakami}@ike.tottori-u.ac.jp
† 2 Nihon University, Tokyo, 101-0061 Japan. saraki@st.rim.or.jp
† 3 Gifu University, Gifu-city, 501-11 Japan, ikeda@info.gifu-u.ac.jp
† 4 Niigata University, Niigata-ity, 950-2181 Japan, miyazaki@ie.niigata-u.ac.jp

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

A huge investment has been made in the research and development of MT technology in the 1980s, resulting in some noteworthy achievements[(1)]. However, it is a difficult problem to develop MT systems between languages belonging to different language 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.

Since 1990s, corpus based approaches have been expected as one of the methods for solving the problem. *Example-based MT*, for instance, was proposed by Nagao [(2)] and then advanced by Sato [(3, 4)]. However, this method was not enough to reach a level at which it can be put into practical use due to the lack of incomparably large corpus and considerable side-effects caused by the difficulties associated with increasing corpus size.

In 2000s, researches on *Statistical MT* has become very active. The principle of the method was first proposed for the translation between English and French both of which are included in the same language family [(5)]. This method was applied to the translations between Japanese and English[(6)] both of which are different from each other in language family, by adopting HMM model after the success of the researches on speech recognition. However, it is so difficult to prepare statistically significant amount of examples that applications are very limited.

Thus, there are limits to the methods directly relied on existing database. The realization of an organized and codified knowledge base will be expected.

An example of this kind of knowledge base is *Pattern-based MT* [(7)-(10)], which has already been used in many commercial systems combining the *Transfer-method* and *Translation-memory* [(11)] 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 [(12-14)], it is a long way from being actually used.

To address such problem, a *Multi-Level Translation Method* (MLTM) [(15)] 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 [16,17] resulting in a marked improvement in the translation quality of simple sentences [18].

However, the MLTM has two problems [19,20], 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 *Analogical Mapping Method (AM-method)* [21] has recently been proposed in which fundamentals thereof can be established by the *Semantic Typology* [22] and *Analogically Equivalent Thinking* [23] theories. In this method, the *non-linear* sentence structures of a source language are semantically mapped into those of a target language using a *Semantically Classified Sentence Pattern Dictionary (SP-dictionary)* where one or more *semantic patterns* (SPs) for the target are defined corresponding to a pattern of the source.

In order to realize this method, we have started 5 year project to developed a *SP-dictionary* and have compiled 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.

2. Outline 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 [22], 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 [23]. 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 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*

K. Ichikawa formulated the analogical reasoning in scientific discovery [18] 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:

$$C\,(\,A_\alpha \overset{\varepsilon}{=} B_\beta\,) \qquad (1)$$

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

---

*1 Nagao proposed an *Analogical Translation Method* based on the similarities between syntactic structures and word meanings used in corpus writings (2-4). 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.

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:

$$A_\alpha \Rightarrow C(A_\alpha) \Rightarrow \varepsilon \Rightarrow C(B_\beta) \Rightarrow B_\beta \qquad (2)$$

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 multi-word expressions) [*1]. The categories and items are based on the *Semantic Attributes* of the *Valency Patterns* defined in "*A- Japanese-Lexicon*" [16].

(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. Fig. 1 and Fig. 2 show an application example of *AM-method* for Japanese to English MT system.

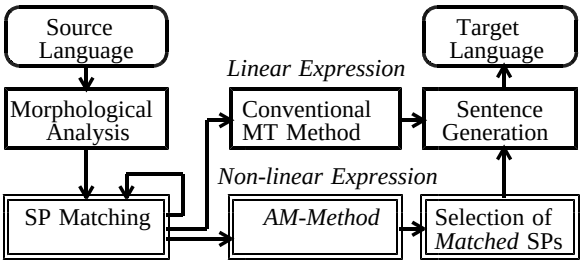

Figure 1. Translation process by *AM-method*

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

An SP is considered as part of the 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 *semantic composition*. SPs are defined from the view point of the *linearity* and *non-linearity* of expressions as will be described in the following.

3.1 Method of Judging *Non-linearity*

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

The development of conventional NLP technologies has been supported by the principle of *Semantic Composition*. There have been many studies and discussions among the adherents of *Compositionality* and *Contextuality* [24-34]. Frege defined this principle as *"The meaning of a complex expression is determined by the meanings of its parts, and the way in which those parts are combined"*.

The most typical example based on the principle is the *Transfer method* for conventional MT systems. In this method, assuming that the meanings of parts are given by lexicon and the way of combination is given by syntax, the parts are separately converted and combined to generate the target language expression.

However, this method has reached its limits. Especially in the translation between the languages from different families, original meanings are lost during the translation process and high quality translation cannot be obtained.

In this research, we assume that the meaning of the whole expression cannot be determined by the parts, but the meanings of the parts can be deter-

*1 As long as languages have grown in each comunity, it is impossible to build such a concept system that includes all concepts from every languages. Considering that the translation is semantic approximation to the very end, it is better to prepare LSC system for each pair of languages in order to define as many concepts as possible. The translation between expressions that have not a comon concept are impossible at all and not the subject of MT.

mined by the meaning of whole expression. Therefore we propose a pattern based method to determine the meaning of the whole expression in advance.

Linguistic expression is a means of representing speaker's conceptual cognition. A speaker first selects the most suitable expression structure (or frame) out of those which may occur in their mind to represent their thoughts. Then, careful not to lose the total meaning, the speaker selects parts for each component to complete the sentence. 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 that cannot be replaced by any other components.

Then, we discriminate between the former as a *"Linear Components"* and the latter as "Non-linear *Components"*. Specifically, the *linearity* and *non-linearity* of a component and an entire expression are defined as follows:

[Definition 1]: *Linearity* of expression 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*)
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 when linguistic expressions are *linear expressions.*

Our definitions are compatible with Frege's explanation. Frege explained the feature of compositionality of logical expressions as 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 [34].

Linear components correspond to *compositional components* since they are replaceable with another equivalent components without changing the meaning, but whether *decomposable components* or not cannot be determined without checking it's inner structure. In contrast to this, *non-linear components* cannot be replaced with other components without changing the entire meaning so they are not *compositional components*.

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 methods based on *semantic composition*.

(2) Definition of Meaning for 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*.

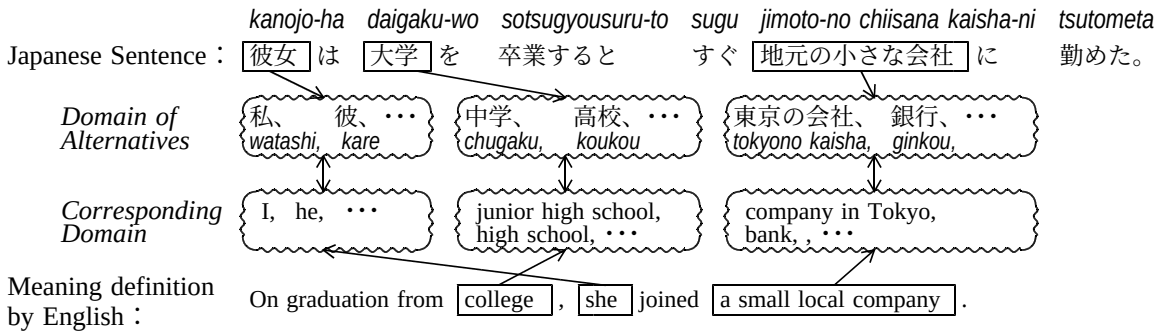

Fig. 2 Example of *linear components*

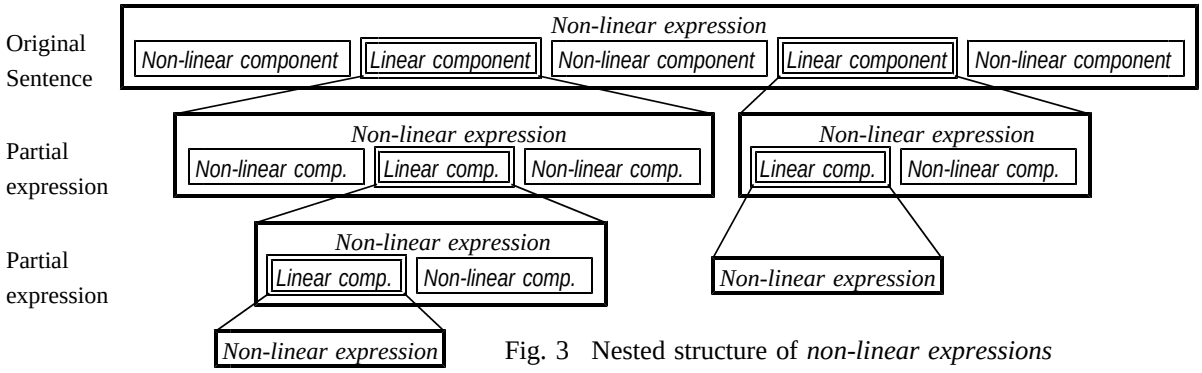

Fig. 3 Nested structure of *non-linear expressions*

(3) Characteristics of *linear components*

Fig. 2 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 Fig. 2.

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 Fig. 3.

Thus, the *linear components* are expressions which can be separated again into *linear* and *non-linear components.* Finally, all the expressions are represented by the combination of one or more *non-linear components* and zero or more *non-linear expressions*. Here, it is very important to notice that *non-linear expressions* are *"meaning units"* and SPs are defined for them..

In the broad sense of the meaning of expression structures, our linguistic model has traits in common with the concept of "*schema"* in *Cognitive Linguistics* advocated by Laugacker[35] and shows similarity in the *Farne Semantics* or *Construction Grammar* proposed by Fillmore, Atkins and Goldberg [36].

However, major concern of these researches is a semantic relationship between *linear components*. In contrast, our method has focused attention on the importance of *non-linear components*. The importance of the information presented by patterns was also pointed out for the analysis of Multiword Expressions[37, 38].

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[*1]. 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 2 shows the constituents of SPs.

SPs are defined by using *Literal, Variable, Functions* and *Symbols*. *Literals* are used to represent *non-linear components. Variables* are used for *linear component*. There are 3 kinds of variables: *word variables, phrase variables* and *clause variables.* These are used to define *Word Level, Phrase*

*1 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 embedded structure by combining the finite *non-linear expressions* as shown in the last section of this paper.

Level and *Clause Level* SPs and domains are semantically defined using semantic attributes.

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.

Table 2. Elements for defining SPs

Linear components
Non-linear components

Group	Type of Usage		
Literals	Japanese Character, English Character	○	
Variables (15 types)	To represent 3 level SPs: ①Word variable (9), ②Phrase Variable (5), ③Clause variable(1) Constrain by Semantic Attribute		○
Functions (107 + α types)	①Word form, ②Tense, Aspect, Modality ③Transformation of part of speech ④Sentence generation, ⑤Others	○	○
Symbols (7 types)	①Synonymous word or expression ②Permutable word order, ③Arbitrary components, ④Erased components, ⑤Others	○	○

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. *Semantic Pattern* 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 %)[18] for simple sentences by the realization of "Goi-Taikei: *A-Japanese Lexicon*"[16]. 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. 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 [39] 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 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* [16] and *Ruigo Daijiten*[40] 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 Fig. 4.

It was necessary to have 13.6 person-years of analysts for the development of the *SP-dictionary*.

Table 3. Generalization Levels of SPs

Level	Processes of Generalization
word-level	**(1) Replacement of *linear words* by variables,** (2) Marking of optional, (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* *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／V4(R9100).*teiru*のだから／N5 (N1453).*dantei*。 *ha* *ni* *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* *motozuite* *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).*tearu*ので、N2 (G2005) に当たっては／VP3 (R3901).*gimu* *node* *niatattewa*
Example	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　注意しなくてはならない。 *sorewa* *kiwamete yuudokudearunode* *siyouniatattewa* *juunibunni* *chuuisinakerebanaranai*
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. Gnnnn: Semantic Attribute Number defined by *A-Japanese-Lexicon*[16].
Rnnnn: Semantic Attribute Number defined by *Ruigo Daijiten*[40].

Fig. 4　Examples of Generated SPs

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.

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* SPs and *clause-level* SPs to the example sentences are 99.5%, 81.3% and 10.1%.

Table 4. The Number of Different SPs

SP type	Type of SP			
	word-l.	*phrase-l.*	*clause-l.*	Total
Complex S.	59,658	52308	5,938	107,905
Compound S.	49,897	36,016	3,996	89,909
Mixed Type	12,174	10,025	1,551	23,750
Total	121,729	88,349	11,485	221,563

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 c*ompositional semantics* cannot deliver the expected results of high quality translations as shown in the example

5.2 The Ratio of *Linear Components*

(1) Frequency of Variables

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

Table 5. Ratio of *Linear Components*

Component Type	Frequency	Replacements by Variable	*Ratio of Linearity*
Full Word	763,968	472,521	62 %
Phrase	463,636	102,000	22 %
Clause	267,601	11,486	4.3 %

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*.

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 *SP-dictionary*

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×P2 represents semantic coverage of the *SP-dictionary*.

6.2 Saturation of *Matched Pattern Ratio*

The relationship between the *Matched Pattern Ratio* (P0) and the number of SPs were evaluated as shown in Fig. 5.

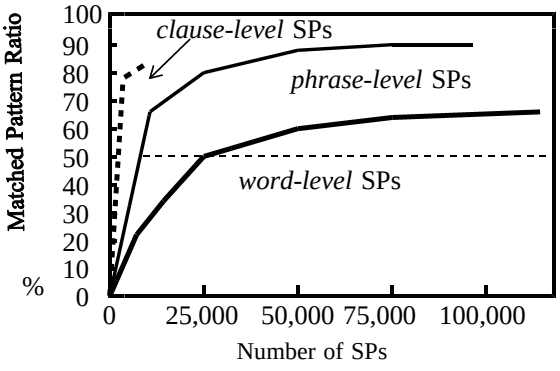

Fig. 5 Relation between No. of SPs and P0

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.

6.3 *Matched Pattern Ratio* and *Precision*

(1) Evaluation Results

The evaluation results of P0, P1 and P2 are shown in Table 6.

Table 6. Evaluation Results for Precision

Level of SPs	Matched Patten Ratio (P0)	Precision (P1)	*Accumurative Precision*(P2)	*Semantic Coverage* (P0xP2)
Word	66.0 %	30.5 %	69.0 %	45.5 %
Phrase	80.4 %	24.4 %	66.2 %	59.5 %
Clause	70.2 %	13.8 %	52.2 %	44.1 %

It is found that P0 of *word-level* SPs is highest. Atthough the number of *clause-level* SPs is only 1/10 compared to that of *word-level* SPs, P0 is higher than that of *word-level* SPs. This means that generality of *clause-level* SP is more than 10 times higher than that of *word-level* SP.

Compared to P1, P2 is a few times higher. This means that the matched SPs contain many incorrct candidates.

After all, *Semantic Coverage* of *phrase-level* SPs is the highest and most promising.

(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 Fig. 6.

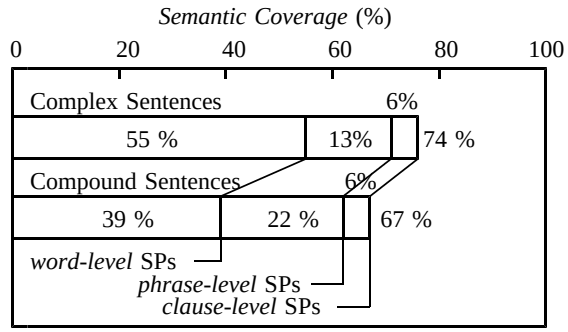

Fig. 6 *Semantic Coverage* of *SP-dictionary*

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, the 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 %.

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* exist 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

(1) M. Nagao: *Natural Language Processing*, Iwanami Shoten, Tokyo, 1996.

(2) M. Nagao: A Framework of a Mechanical Translation between Japanese and English by Analogy Principle, In A. Eithorn and R. Barnerji (Eds.), Artificial and Human Intelligence, North-Holland, pp. 173-180, 1984.

(3) Satoshi Sato: An example based translation and system, Proceedings of COLING-92, pp. 1259-1263, 1992.

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

(5) P. F. Brown, C. John, S. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercar and P. S. Roossin: A Statistical Approach to Machine Translation, Computational Linguistics, Vol. 16, No. 2, pp. 79-85, 1990.

(6) T. Watanabe and E. Sumita: Bidirectional Decoding for Statistical Machine Translation, Proceedings of COLING-02, pp. 1075-1085, 2002.

(7) K. Takeda: Pattern-based Context Free Grammars for Machine Translation, 34th Annual Meeting of the Association for Computational Linguistics, pp. 144-151, 1996

(8) K. Takeda: Pattern-based Machine Translation, Proceedings of the 16th COLING, Vol. 2, pp. 1155-1158, 1996.

(9) H. Watanabe and K. Takeda: A Pattern-based machine translation system extended by example based processing, 17th COLING, pp. 1369-1373, 1998

(10) H. Uchino, S. Shirai, A. Yokoo, Y. Ooyama, and K. Furuse: A Japanese-English Machine Translation System for Market Flash Reports, Journal of IECIE, Vol. J84-DII No. 6, pp. 1167-1174, 2001.

(11) M. Nagao, S. Kurohashi, Sato, S. Ikehara, and H. Nakano: *Science of Natural Language* Vol. 9: *Linguistic Information Processing*. Iwanami Publisher, 1998.

(12) F. Almuallin, Y. Akiba, T. Yamazaki, A.Yokoo and S. Kaneda: Two Methods for Leaning ALT-J/E Translation Rules from Examples and a Semantic Hierarchy, COLING94, pp. 58-63, 1994

(13) H. A. Guvenir and I. Cicekli: Leaning Translation Templates from Examples. Information Systems Vol. 23, No. 6, pp. 2353-363, 1988

(14) M. Kitamura and Y. Matsumoto: Automatic Extraction of Word Sequence Correspondence in Parallel Corpora, 4th Annual Workshop on Very Large Corpora, pp. 79-87, 1996

(15) S. Ikehara, M. Miyazaki, and S. Shirai, Y. Hayashi: Recognitions and Multi-level Machine Translation Method based on It. Journal of IPSJ. Vol.28, No.12, pp.1269-1279, 1987.

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

(17) S. Shirai, S.Ikehara, A. Yokoo and H. Inoue: The Quantity of Valency Pattern Pairs required for Japanese to English MT and Their Compilation, Proc. of NLPRS'95, Vol. 1, pp. 443-448, 1995

(18) M. Kanadechi, S. Ikehara, and J. Murakami: Evaluation of English Word Translations for Japanese Verbs using Valency Patterns. The 63-th Annual Conference of IPSJ, 2-267-268, 2001.

(19) S. Ikehara. 2001. Challenge to the Fundamental Problems on Natural Language Processing, Journal of the Japanese Society of Artificial Intelligence, Vol. 16, No. 3, pp. 422-430.

(20) S. Ikehara: Meaning Comprehension Using Semantic Patterns in a Large Scale Knowledge-Base, Proceedings of the PACLING'01, pp. 26-35, 2001.

(21) S. Ikehara: Toward the Realization of Ultimate MT Method = MT Method based on Analogical Thinking ＝, AAMT Journal, No. 32, pp. 1-7, 2002.

(22) J. Arita: *Lecture of Germany Vol. 2*. Nanundo Publisher, pp.48-56, 1987.

(23) K. Ichikawa: *Methodology for Creative Research*, Sanwa Shobo, 1960.

(24) J. Allen: *Natural Language Understanding*, The Benjamin/Cummings Publishing Company, 1995.

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

(26) B.Carpenter: *Type-Logical Semantics*, MIT Press, 1997.

(27) M. B. Platts: *Ways of Meaning - An Introduction to a Philosophy of Language-* (2nd Eds.): MIT Press, 1997.

(28) Peter Ludlow: Semantics, *Tense, and Time -An Essay in the Metaphysics of Natural Language-*, MIT Press, 1999.

(29) Mary Dalrymple (Eds.): *Semantics and Syntax in Lexical Functional Grammar*, MIT Press, 1999.

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

(31) A. Cruse: *Meaning in Language - An Introduction to Semantics and Pragmatics-* (2nd Eds.), Oxford University Press, 2004.

(32) B. H. Partee: *Compositionality in Formal Semantics*, Blackwell Publishing, 2004.

(33) See the bibliography listed in Stanford Encyclopedia of Philosophy:
" http://plato.stanford.edu/entries/compositionality/"

(34) J. Allwood, L. Anderson and O. Dahl: *Logic in Linguistics*. Cambridge University Press, 1977

(35) R. W. Langacker: FOUNDATION OF COGNITIVE GRAMMAR, Stanford University Press, 1987.

(36) C. Fillmore, P. Kay, L. Michaelis and I. Sag: *Construction Grammar*, Stanford Univ Center for the Study, 2005.3/15

(37) T. Baldwin and F. Bond, Multiword Expressions: Problems for Japanese NLP, 8th Annual Meeting of the Association for Natural Language Processing, pp. 379--382, 2002

(38) I. Sag, T. Baldwin, F. Bond, A. Copestake and D. Flickinger; Multiword Expressions: A Pain in the Neck for NLP, in *Computational Linguistics and Intelligent Text Processing:* Third International Conference: CICLing-2002", Ed.: Alexander Gelbuk, Springer-Verlag, 2002

(39) Morphological Analysis Program for Japanese sentences developJed for Japanese to English machine translation system ALT-J/E. see "ALT-JAWS: Japanese Automatic Word Separator", 2002.
"http://www.kecl.ntt.co.jp/icl/mtg/resources/altjaws.html"

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