Analysis of Relations between Noun and Deverbal Nouns in Japanese Compounds Based on Lexical Conceptual Structure

Koichi Takeuchi*, Kiyoko Uchiyama†, Masaharu Yoshioka‡, Kyo Kageura* & Teruo Koyama*
*National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430, Japan
†Keio University Graduate School of Media and Governance
‡Graduate School of Engineering, Hokkaido University
*{koichi,kyo,t_koyama}@nii.ac.jp
†kiyoko@sfc.keio.ac.jp
‡yoshioka@db-ei.eng.hokudai.ac.jp,
TEL +81-3-4212-2580
FAX +81-3-3556-1916

Summary

In this paper, we describe a principled approach for analysing relations between constituent words of compound nouns, specifically those whose heads are deverbal nouns, based on the classification of deverbal nouns by their lexical conceptual structure (called TLCS in this paper) and the classification of nouns in modifier position vis-a-vis TLCS of head deverbal nouns. As the compound nouns which have deverbal nouns as heads constitute a major part of compound nouns, it is an important starting point for analysing relations of constituents in compound nouns in general. Through the qualitative analysis of the data and the experimental evaluation of 858 compound nouns, we show that the use of TLCS as the theoretical basis is very promising for constructing compound analyser.

Keywords: lexical conceptual structure, deverbal noun, Japanese compound

1 Introduction

In this paper, we describe a new approach for analysing relations between constituent words of Japanese compound nouns. We focus here on the compound nouns whose head is a deverbal noun, as they constitute a major class of compound nouns and thus constitute an important starting point for analysing compound nouns in general. The method we propose is based on the classification of deverbal nouns by their lexical conceptual structure (that we call them TLCS)[1] and the classification of nouns in modifier position vis-a-vis the TLCS of head deverbal nouns.

The existing work on compound noun analyses can roughly be classified into two, i.e. statistical approach and semantic approach. Statistical techniques (Kobayashi, 1995)[2] (Lauer, 1995)[3] are very useful when the training corpus is available, but they are less successful because of lack of grammatical and semantic information. Some researchers proposed original semantic approach (Yokoyama and Sakuma, 1996)(Miyazaki et al., 1993) whose semantic information is well constructed, but it is not clear to what extent their approach can be extended to large scale data. The approach we propose is based on LCS scheme and is expected to

[1] It was named to prevent confusing other type of LCS. 'T' was named after the first character of the first author's surname.

[2] The accuracy of his method of analysing Japanese compound noun is reported about 93%. The result does not include the analysis of relation between constituent words but of chunking pattern in compounds.

[3] The thesis reports the accuracy of analysing English compounds is about 80%.

overcome the shortcomings of the previous approaches.

2 Basic Framework: Compound Noun Analysis and LCS

Recent work in morphology (Kageyama, 1993) (Kageyama, 1996), (Kageyama, 1997) (Kageyama, 1998) (Kageyama, 1999), has shown that the relations between the words in Japanese compounds with deverbal nouns as heads (henceforth deverbal heads) can be divided into two, i.e. (i) the modifier becomes an internal argument (i.e., object) of deverbal heads and (ii) the modifier functions as an adjunct. The two relations are convenient level to grasp rough semantic relations in compounds.

Take, for example, the following two compound nouns:

a. kikai-sousa 'machine-operate' (machine operation)

b. kikai-hon'yaku 'machine-translate' (machine translation)

The modifier 'kikai' is an internal argument of a deverbal head in the former, while it is an adjunct[4] in the latter. We assume that disambiguating these two relations is the essential starting point of analysing compound nouns with deverbal heads.

We assume that the relation can be determined by the combination of the TLCS on the side of deverbal heads and the consistent categorisation of modifier nouns on the basis of their behavior vis-a-vis a few canonical TLCS types taken by the deverbal heads. Figure 1 shows the example of disambiguating relations depending on the TLCS types of deverbal heads.

Since the TLCS has an abstract structure of verb meaning, the semantic structure can be applied to the disambiguation.

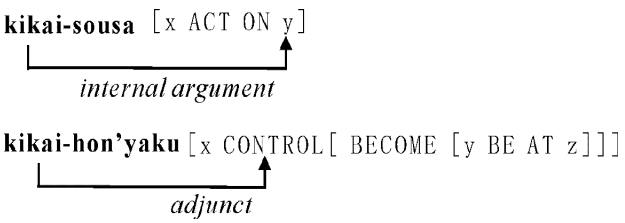

Figure 1: Disambiguating relations depending on TLCS types of deverbal heads.

The actual assignment of TLCS to many deverbal nouns as well as the method of categorising modifiers are original contributions of the present research to the lexical-semantic theory. From the point of view of application, we will see that our method determine the relations of 858 compounds with 99.3%. Though they are 'internal' data, the result show that our approach is very promising for compound noun analysis.

3 Lexical Conceptual Structure

LCS aims at expressing lexical meaning of verb and clarifying mechanism of behavior of verb in sentence. In order to explain various kinds of linguistic phenomena, many kinds of LCSes and the predicate set are proposed(Jackendoff, 1990) (Kageyama, 1993) (Pustejovsky, 1995). Their work is well investigated but it is unclear that what kind of LCS scheme and predicate set should be needed to analyse or disambiguate linguistic phenomena to some extent.

In this paper we set original predicates and basic structure types from the view point of building practical analyser of Japanese compound nouns. Focusing on the disambiguation of relations between modifier noun and deverbals, we think it is possible to clarify what kind of structure are effective to the analyses. We call them TLCS and try to construct them referring to the previous work(Kageyama, 1996) with limiting the variety of predicates.

Assuming that we can define proper TLCS for all the verbs and deverbal nouns, we established 12 TLCSes as listed in following Table 1,

[4] In English deverbal compounds, there are grammatical rules (Roeper and Siegel, 1978) (Selkirk, 1982) that apply well to phenomena of relations between modifiers and deverbal heads. The deverbal compounds in the above case **b.** doesn't exist in English. For example, 'past-making' correspond to the case **a.** exists, but 'fast-making' to the case **b.** doesn't exist(Kageyama(ed.), 2001). In these compounds, the inflection of verb gives a good hint to analyse the structure. However, it is the same problem of analysing Japanese compounds to analyse nominal English compounds like 'machine operation' and 'machine translation'.

Table 1: List of TLCS types

1[x ACT ON y]	enzan (calculate), sousa (operate)
2[x CONTROL[BECOME [y BE AT z]]]	shori (process), hon'yaku (translate)
3[x CONTROL[BECOME [y NOT BE AT z]]]	shahei (shield), yokushi (deter)
4[x CONTROL [y MOVE TO z]]	densou (transmit), dempan (propagate)
5[x=y CONTROL[BECOME [y BE AT z]]]	kaifuku (recover), shuuryou (close)
6[BECOME[y BE AT z]]	houwa (become saturated)
	bumpu (be distributed)
7[y MOVE TO z]	idou (move), sen'i (transmit)
8[x CONTROL[y BE AT z]]	iji (maintain), hogo (protect)
9[x CONTROL[BECOME[x BE WITH y]]]	ninshiki (recognize), yosoku (predict)
10[y BE AT z]	sonzai (exist), ichi (locate)
11[x ACT]	kaigi (hold a meeting), gyouretsu (queue)
12[x CONTROL[BECOME [[NAME]y BE AT z]]]	shomei (sign)

using the data of technical terms described in section 5. In the table, we also show the examples of deverbal nouns which are assigned to each TLCS.

In TLCSes, words written in capital letters (as 'CONTROL', 'ACT' and etc.) mean semantic predicates. 'x' denotes an external argument, while 'y' and 'z' denote internal arguments (see (Kageyama, 1996)(Levin and Hovav, 1995)). TLCS **1** $\sim$ **4**, **8** and **9** represent different types of transitive verbs. TLCS **11** and **12** are intransitive verbs. TLCS **5** represents an ergative verb and, **6**, **7** and **10** are unaccusative verbs. In the next section, we explain all kinds of components that appear in the 12 TLCSes.

3.1 TLCS Components

Each TLCS in Table 1 consists of the combination of semantic components that form basic meaning. After we show and explain the semantic components as follows, we explain how we set the 12 TLCSes.

- 'y BE AT z' means that 'y' exists at 'z' as a state or a place.
- 'BECOME' represents changing into the state of a predicate that connects at the right side.
- 'y MOVE TO z' means that 'y' changes the place to 'z'. This component is subcategory of '[BECOME[y BE AT z]]'.
- 'x CONTROL' means 'x' controls a predicate that connects its right side.
- 'x=y' represents an internal argument 'y' can change the state by itself.
- 'x ACT ON y' means the continuous action of 'x' to 'y' without changing state of 'y'.
- 'x ACT' represents the continuous action of 'x'.
- 'x BE WITH y' means that 'x' owns 'y'.
- 'NOT' only represents negation.
- '[NAME]y' represents argument 'y' fills with something and the verb of this TLCS cannot take an internal argument. The predicate 'NAME' itself is less important.

The verb whose TLCS contains 'ACT ON' or 'CONTROL' is transitive. The verb whose TLCS contains 'BECOME' or 'MOVE TO' has achievement nature.

3.2 Setting TLCS scheme

Analysing the data of technical terms (which have about 220 kinds of deverbals), we set 12 TLCSes based on semantic components.

Since our expectation of TLCSes is that the structures would support to disambiguate the relations whether the modifier is an internal argument or an adjunct for deverbal head, we carefully categorise verbs depending on their arguments type and semantic relations between the arguments. The verbs are categorised into

transitive and intransitive verbs by the argument types, and semantic relations between the arguments are characterised and decomposed using TLCS components.

Table 2 shows the typology of TLCSes. The 'both' in Table 2 denotes that TLCS **5** have both transitive and intransitive nature,[5] that is known as transitivity alternation.

Table 2 also shows that our categorisation method is more detailed than traditional category. We think proposed method must be key of analysing deverbal compounds.

Table 2: Typology of TLCS

argument type	TLCS num.	key compo.
transitive	1,	ACT ON
	2,3,4,8,9	CONTROL
intransitive	6,7	BECOME
	10	BE AT
	11	ACT
	12	[NAME]y
both	5	x=y

3.3 Assigning TLCS to verb

We assign TLCS to verb using typology of TLCS and the characteristic of TLCS components.

For example, the verb 'shori' (process) whose TLCS is **2** in Table 1. If we can find that 'shori' (process) is transitive, its TLCS would contain 'ACT ON' or 'CONTROL' from Table 2. Moreover, if we find it has an achievement nature, its TLCS would contain 'CONTROL' and 'BECOME'. Finally, we investigate whether the object of the verb has the meaning of changing state, its TLCS would be defined as **2**. In order to investigate the character of deverbal noun, we consult a dictionary and text corpus.[6]

[5] For example, transitive case is *kaigi-o-shuuryou-suru* 'meeting-ACC-end-do' [end a meeting], and intransitive case is *kaigi-ga-shuuryou-suru* 'meeting-NOM-end-do' [a meeting ends]. 'ga' and 'o' designate case markers of nominative (NOM) and accusative (ACC), respectively. This phenomenon also appears in English, which is called ergative verb in the article (Kageyama, 1996).

[6] We use Nikkei newspaper article (1992 - 1998).

4 Categorisation of Modifier Nouns

Assuming that modifier nouns have lexical character depending on the predicates in TLCS, we establish the following five categories or features on the basis of the nouns behavior vis-a-vis a few TLCS types and accusativity of them. This categorisation schema is built using technical term and newspaper[6] data, with some linguistic rationales.

4.1 Categorisation

- **Categorised by accusativity of modifiers**
  An accusativity of modifier concerns all verbs, i.e., the modifier without accusativity means that the modifier does not appear in sentences as an object of any verb(Uchiyama et al., 1999). Therefore the modifier is adjunct of the relation in compounds. We categorise it as a negative category denoted by '-ACC', if it does, '+ACC'. For example, 'kimitsu' (secrecy) and 'kioku' (memory) are '+ACC', and 'sougo' (mutual-ity) and 'kinou' (inductiv-e/ity) are '-ACC'.
- **Categorised by specific components in TLCS**
  We assume that the specific components are 'ON', 'CONTROL', 'x=y' and 'BECOME [y BE AT z]' that are part of TLCS types of **1**, **2**, **5** and **6**, respectively.
  In order to categorise modifiers, we checking whether they appear in sentences as an object of the verb whose TLCS has each specific component.
  If the modifier does not appear as an object in each case, the modifier is categorised as a negative category denoted by '-', if it does, '+'. In Table 3, the categories of 'ON', 'CONTROL', 'x=y' and 'BECOME [y BE AT z]' are denoted as 'ON', 'EC', 'IC' and 'UA'. The examples of the modifier nouns that are categorised as negative or positive category for each component are as follows:
  ON 'koshou' (fault) and 'seinou' (performance) are '+ON', and 'heikou' (parallel) and 'rensa' (chain) are '-ON'.

EC 'imi'(semantic) and 'kairo' (circuit) are '+EC',and 'kikai' (machine) and 'densou' (transmission) are '-EC'.

IC 'fuka' (load) and 'jisoku' (flux) are '+IC', and 'kakusan' (diffusion) and 'senkei' (linearly) are '-IC'.

UA 'jiki' (magnetic) and 'joutai' (state) are '+UA', and 'junjo' (order) and 'heikou' (parallel) are '-UA'.

4.2 Categorisation of Combination of Modifier Nouns and TLCS of Deverbal Heads

These categories can be used for disambiguating the relations in compound nouns with deverbal heads with various TLCS types. The range of application of the above categorisations can be summarised in Table 3. The number in TLCS column corresponds to the list number in Table 1.

Table 3: Categorisation of combination of modifier nouns and TLCS of deverbal heads.

role	modifier category	TLCS
adjunct	-ACC	any
	-ON	1
	-EC	2,3,4
	-IC	5
	-UA	6,7
	any	10,11,12
role	modifier category	TLCS
int. argument	+ACC	8, 9
	+ON	1
	+EC	2,3,4
	+IC	5
	+UA	6,7

The verbs whose TLCS types are **10**, **11**, and **12** do not take an internal argument. The verbs whose TLCSs are **8** and **9** take an internal argument when the modifier is '+ACC'. The other categories of TLCS are defined from the meaning of categorisation.

With these rules and categories of nouns, we can analyse the relations between the words in compounds with deverbal heads. For example, when the modifier 'kikai' (machine) is categorised as '-EC' but '+ON', the modifier in *kikai-hon'yaku* (machine-translation) is analysed as adjunct (that means 'translation by a machine'), and the modifier in *kikai-sousa* (machine-operation) are analysed as internal argument (that means 'operation of a machine'), correctly.

5 Experimental Evaluation

5.1 Experiments and Results

We applied the categories to the analysis of 858 two-constituent compound noun terms with deverbal heads, which are taken from dictionaries of technical terms (Aiso, 1993) and (Committee on editing dictionary, 2001). These compounds were assumed to be segmented. We used technical terms as they are highly productive and the processing of compound terms are strongly needed. The procedure of analysis as follows:

Step 1 If the TLCS of the deverbal head is the type of **10**, **11**, **12** in Table 1, then declare the relation as adjunct and terminate. If not, go to next.

Step 2 If a modifier is the category '-ACC', then declare the relation as adjunct and terminate. If not, go to next.

Step 3 If the combination of the modifier category and the TLCS of the deverbal head applies to the case of '-ON', '-EC', '-IC' or '-UA' in Table 3, then declare the relation as adjunct and terminate. If not, go to next.

Step 4 Declare the relation as internal argument and terminate.

Since there is no restriction about an adjunct relation, this procedure can only check a possibility of an internal argument for a deverbal head.

According to the manual evaluation, 99.3% (852 words) of the results were correct. Table 4 shows the details how the rules are applied to disambiguating the relations between constituent words in compounds. The most effective category of disambiguation is '-ACC' used in step 2. Though the range of application in our framework is still limited, the performance is very promising, which in turn shows that the use of TLCS and noun categorisation on the ba-

Table 4: Statistics of effective rules applied to analysis

process	applied rules	frequency
Step 1	in case of **10**,**11**,**12**	47
Step 2	-ACC	234
Step 3	-ON	68
	-EC	147
	-IC	21
	-UA	45
Step 4	internal argument	290
	total	852

sis of TLCS is highly promising.

5.2 Diagnosis and Discussions

We found a tiny number of modifier nouns deviate our assumptions (in the 858 terms, we found six). Take, for instance, the following cases:

e1. jiki -shahei
magnet shield
magnet shield
e2. jiki -kioku
magnet memorise
magnet memory

Here the first line denotes a Japanese compound, the second means word-for-word translation and the third shows translation of the compound. A mark '-' designate that forming one word by connecting with before morpheme.

In the former case, the modifier is an internal argument, while in the latter it is an adjunct, but the TLCS structure of both deverbal heads is the same causation type according to our definition.

The next error case is the same as above. The

e3. insatsu -yokusei
print inhibit
print inhibit
e4. insatsu -haisen
print wire
printed wiring

modifier is an internal argument in the case **e3**, while it is an adjunct in case **e4**.

The TLCS type of deverbal heads in **e1** $\sim$ **e4** is causation type but their types have a little different point with something in common. It is the existence of the predicate 'NOT'. The deverbals both 'shahei' (shield) in **e1** and 'yokusei' (inhibit) in **e3** are categorised into TLCS **3** in Table1, while the deverbals 'kioku' (memorise) in **e3** and 'haisen' (wire) in **e4** are categorised into TLCS **2**. Therefore the above two sets of error cases show that TLCS **3** having the predicate 'NOT' can take a noun as an internal argument which does not become an internal argument to causative verb as TLCS **2**.

At the moment, these types of TLCS are regarded as the same categorisation of combination rule in Table3 since we do not have enough evidence about this. Further examination is needed to deal with them.

The following two error cases are related to the problem of categorisation about the modifier 'suuchi' (numerical value).

e5. suuchi -kaiseki
numerical value analyse
numerical analysis
e6. suuchi -senkou
numerical value punch
digit punch

e7. suuchi -keisan
numerical value compute
numerical computation
e8. suuchi -seigyo
numerical value control
numerical control

In the both examples, the modifier is an internal argument in the former case while it is an adjunct in the latter. The deverbal heads in **e5** and **e6** are categorised into TLCS **2**, and the deverbals in **e7** and **e8** are TLCS **1**. The former error case is concerned with the categorisation of **EC** at modifier 'suuchi' (numerical value), while the latter is **ON**.

The next error case **e9** and **e10** is related to the categorisation of the modifier 'ronri' (logic).

In the case **e9**, the modifier is an internal argument while in the case **e10**, it is an adjunct. The TLCS of the deverbals 'sekkei' (design) and 'meirei' (order) is causation type, that is, TLCS **2**. However there is room for discussion whether the compound at **e10** should be regarded as deverbal noun. The last error case also has the same problem.

e9. ronri -sekkei
logic design
logic design
e10. ronri -meirei
logic order
logic instruction

e11. nyuuryoku -meirei
input order
input instruction
e12. nyuuryoku -doushutsu
input resolve
input resolution

The category of both deverbal heads in **e11** and **e12** is causation type TLCS **2**. The former is an internal argument relation while the latter is an adjunct one. However there is room for discussion whether the compounds **e10** and **e11** are deverbal compounds or nominal noun compounds. From the fact that the nominal noun 'instruction' is used as a translation of compound in **e10** and **e11**, there is some possibility that they are nominal noun. We would like to investigate these issue in the future work.

Comparing Japanese deverbal compounds with their translation of English compounds in **e1** $\sim$ **e10**, the analysis of English deverbal compound must be a little easer than that of Japanese one since the function of constituent words in English compound often appears as inflection. When we focus on the example **e4**, there is no inflection in constituent words 'insatsu' (print) and 'haisen' (wire) in Japanese compound, while inflection shows the function of constituent words as 'printed wiring' in English compound. Therefore we think the method using semantic structure level is necessary for Japanese compound analysis.

6 Conclusions

In this paper, we proposed the original lexical conceptual structure that we call TLCS and the categorisation of the modifier nouns, in order to establish a principled approach for compound noun analysis. The proposed method of the use of TLCS and categorisations of nouns vis-a-vis TLCS, which is based on the current development of the lexical-semantic theory, was proved to be promising, according to the experimental results.

Acknowledgment

We express our sincere thanks to Prof. Kageyama, for giving us some important suggestions.We also thank Nihonkeizashinbunsha for allowing us to use their newspaper articles (1992 - 1998).

References

H. Aiso. 1993. *Dictionary of Technical Terms of Information Processing (Compact edition).* Ohmusha. (in Japanese).

R. Jackendoff. 1990. *Semantic Structures.* MIT Press.

T. Kageyama. 1993. *Grammar and Word Formation.* Hitsujishobo. (In Japanese).

T. Kageyama. 1996. *Verb Semantics.* Kurosio Publishers. (In Japanese).

T. Kageyama. 1997. Denominal verbs and relative salience in lexical conceptual structure. In T. Kageyama, editor, *Verb Semantics and Syntactic Structure*, pages 45–96. Kurosio Publishers.

T. Kageyama. 1998. Grammar and morphology. In Y. Ohtsu et al., editor, *Science of Linguistics 3,Word and Lexicon*, pages 2–51. Iwanami Shoten. (In Japanese).

T. Kageyama. 1999. *Morphology and Semantics.* Kurosio Publishers. (In Japanese).

T. Kageyama(ed.). 2001. *Verb Semantics and Syntax structure.* Taishuu Shoten. (In Japanese).

M. Kobayashi. 1995. Research for analysis of Japanese compound nouns using corpus technique. Technical Report 96TR-0002, Department of Computer Science, Tokyo Institute of Technology(in Japanese).

M. Lauer. 1995. *Designing Statistical Language Learners: Experiments on Noun Compounds.* Ph.D. thesis, Department of Computing Macquarie University.

B. Levin and M. R. Hovav. 1995. *Unaccusastivity.* MIT Press.

M. Miyazaki, S. Ikehara, and A. Yokoo. 1993. Combined word retrieval for bilingual dictionary based on the analysis of compound words(in Japanese). *Transaction of Information Society of Japan*, 34(4):743–752.

Committee on editing dictionary of technical terms of computer. 2001. *English-Japanese Dictionary of Technical Terms of Computer.* Nichigai Associates Co. (in Japanese).

J. Pustejovsky. 1995. *The Generative Lexicon.* MIT Press.

T. Roeper and M. Siegel. 1978. A lexical transformation for verbal compounds. *Linguistic Inquiry*, 9:199–260.

E. Selkirk. 1982. *The syntax of words.* MIT Press.

K. Uchiyama, K. Takeuchi, M. Yoshioka, K. Kageura, and T. Koyama. 1999. A grammatical framework for analysing Japanese nominal compounds with special reference to specialised terms. In *Proc. AILA'99*, pages 159–159. *Proc. 12th World Congress of Applied Linguistics AILA'99.*

S. Yokoyama and K. Sakuma. 1996. Analysis of generation of Japanese compounds using semantic tags(in Japanese). *Measurement of Languages*, 20(7):304–314.