A Corpus for Classifying Usages of Japanese Compound Functional Expressions

Masatoshi Tsuchiya† and **Takehito Utsuro**‡ and **Suguru Matsuyoshi**‡
Satoshi Sato†† and **Seiichi Nakagawa**‡‡
†Computer Center / ‡‡Department of Information and Computer Sciences,
Toyohashi University of Technology, Tenpaku-cho, Toyohashi, 441–8580, Japan
‡Graduate School of Informatics, Kyoto University, Sakyo-ku, Kyoto, 606–8501, Japan
††Graduate School of Engineering, Nagoya University,
Furo-cho, Chikusa-ku, Nagoya, 464–8603, JAPAN

Abstract

Aiming at being used as a corpus for training/testing a tool which properly recognizes and interprets Japanese (compound) functional expressions, this paper studies how to develop a corpus of Japanese functional expressions. This paper focuses on 125 major (mostly compound) functional expressions which have non-compositional usages, and reports current status of developing a corpus of those major functional expressions.

1 Introduction

As in the case of other languages, the Japanese language has various types of functional words such as post-positional particles and auxiliary verbs. In addition to those functional words, the Japanese language has much more compound functional expressions which consist of more than one words including both content words and functional words. Those single functional words as well as compound functional expressions are very important for recognizing the syntactic structures of Japanese sentences and for understanding their semantic contents. Recognition and understanding of them are also very important for various kinds of NLP applications such as dialogue systems, machine translation, and question answering. However, recognition and semantic interpretation of compound functional expressions are especially difficult because it often happens that one compound expression may have both a literal (in other words, compositional) *content word* usage and a non-literal (in other words, non-compositional) *functional* usage.

For example, Table 1 shows two example sentences of a compound expression "に (ni) ついて (tsuite)", which consists of a post-positional particle "に (ni)", and a conjugated form "ついて (tsuite)" of a verb "つく (tsuku)". In the sentence (A), the compound expression functions as a case-marking particle and has a non-compositional functional meaning "*about*". On the other hand, in the sentence (B), the expression simply corresponds to a literal concatenation of the usages of the constituents: the post-positional particle "に (ni)" and the verb "ついて (tsuite)", and has a content word meaning "*follow*". Therefore, when considering machine translation of those Japanese sentences into English, it is necessary to precisely judge the usage of the compound expression "に (ni) ついて (tsuite)", as shown in the English translation of the two sentences in Table 1.

There exist widely-used Japanese text processing tools, i.e., pairs of a morphological analysis tool and a subsequent parsing tool, such as JUMAN[1]+ KNP[2] and ChaSen[3]+ CaboCha[4]. However, they process those compound expressions only partially, in that their morphological analysis dictionaries list only limited number of compound expressions. Furthermore, even if certain expressions are listed in a morphological analysis dictionary, those existing tools often fail in resolving the ambiguities of their usages, such as those in Table 1. This is mainly because the framework of those existing tools is not designed so as to resolve such ambiguities of compound (possibly functional) expressions by carefully considering the context of those expressions.

Considering such a situation, it is necessary to

[1] http://www.kc.t.u-tokyo.ac.jp/nl-resource/juman-e.html

[2] http://www.kc.t.u-tokyo.ac.jp/nl-resource/knp-e.html

[3] http://chasen.naist.jp/hiki/ChaSen/

[4] http://chasen.org/~taku/software/cabocha/

Table 1: Translation Selection of a Japanese Compound Expression “に (ni) ついて (tsuite)”

(A)	私 (watashi) (*I*)	は (ha) (*TOP*)	彼 (kare) (*he*)	に (ni) ついて (tsuite) (*about*)		話した (hanashita) (*talked*)
	(I talked about him.)					
(B)	私 (watashi) (*I*)	は (ha) (*TOP*)	彼 (kare) (*he*)	に (ni) (*ACC*)	ついて (tsuite) *follow*	走った (hashitta) (*ran*)
	(I ran following him.)					

Table 2: Classification of Functional Expressions based on Grammatical Function

Grammatical Function Type		# of major expressions	# of variants	Example
post-positional particle type	subsequent to predicate / modifying predicate	36	67	となると (to-naru-to)
	subsequent to nominal / modifying predicate	45	121	にかけては (ni-kakete-ha)
	subsequent to predicate, nominal / modifying nominal	2	3	という (to-iu)
auxiliary verb type		42	146	ていい (te-ii)
total		125	337	—

develop a tool which properly recognizes and semantically interprets Japanese compound functional expressions in a scale much larger than existing Japanese text processing tools. Especially aiming at being used as a corpus for training/testing such a tool, this paper studies how to develop a corpus of Japanese functional expressions. We focus on 125 major (mostly compound) functional expressions which have non-compositional usages, and report current status of developing a corpus of those major functional expressions.

2 Developing a Corpus of Japanese Functional Expressions

2.1 Japanese Functional Expressions

There exist several collections which list Japanese functional expressions and examine their usages. For example, (Morita and Matsuki, 1989) examines 450 functional expressions and (Group Jamashii, 1998) also lists 965 expressions and their example sentences. Compared with those two collections, *Gendaigo Hukugouji Youreishu* (National Language Research Institute, 2001) (henceforth, denoted as *GHY*) concentrates on 125 major functional expressions which have non-compositional usages, as well as their variants[5] (337 expressions in total), and collects example sentences of those expressions. As a first step of developing a corpus of Japanese functional expressions, we start with those 125 major functional expressions and their variants.

As in Table 2, according to their grammatical functions, those 337 expressions in total are roughly classified into *post-positional particle* type, and *auxiliary verb* type. Functional expressions of post-positional particle type are further classified into three subtypes: i) those subsequent to a predicate and modifying a predicate, which mainly function as conjunctive particles and are used for constructing subordinate clauses, ii) those subsequent to a nominal, and modifying a predicate, which mainly function as case-marking particles, iii) those subsequent to a nominal, and modifying a nominal, which mainly function as adnominal particles and are used for constructing adnominal clauses. For each of those types, Table 2 also shows the number of major expressions as well as that of their variants listed in *GHY*, and an example expression. Furthermore, Table 3 gives example sentences of those example expressions as well as the description of their usages.

2.2 Classification of Functional/Content Usages

The task of developing a corpus of Japanese functional expressions roughly consists of collecting ex-

[5] For each of those 125 major expressions, the differences between it and its variants are summarized as below: i) insertion/deletion/alternation of certain particles, ii) alternation of synonymous words, iii) normal/honorific/conversational forms, iv) base/adnominal/negative forms.

Table 3: Examples of Classifying Functional/Content Usages

	Expression	Example Sentence (English Translation)	Usage
(1)	となると (to-naru-to)	しかしこの病気に効果がないとなると事は重大だ。 (The situation is serious *if* it is not effective against this disease.)	functional (となると (to-naru-to) = *if*)
(2)	となると (to-naru-to)	彼が社長になるための条件の一つとなると考えられている。 (They think that it will *become* a requirement for him to be the president.)	content (∼ となると (to-naru-to) = *that (something) becomes* ∼)
(3)	にかけては (ni-kakete-ha)	お金を儲けることにかけては素晴らしい才能をもっている。 (He has a great talent *for* earning money.)	functional (∼ にかけては (ni-kakete-ha) = *for* ∼)
(4)	にかけては (ni-kakete-ha)	あまり気にかけてはいない。 (I do not *worry* about it.)	content ((∼ を) 気にかけては ((∼)-wo-ki-ni-kakete-ha) = *worry about* ∼)
(5)	という (to-iu)	彼は生きているという知らせを聞いた。 (I heard *that* he is alive.)	functional (∼ という (to-iu) = *that* ∼)
(6)	という (to-iu)	「遊びに来て下さい」という人もいる。 (Somebody *says* “Please visit us.”.)	content (∼ という (to-iu) = *say (that)* ∼)
(7)	ていい (te-ii)	この議論が終ったら休憩していい。 (You *may* have a break after we finish this discussion.)	functional (∼ ていい (te-ii) = *may* ∼)
(8)	ていい (te-ii)	このかばんは大きくていい。 (This bag is *nice because* it is big.)	content (∼ ていい (te-ii) = *nice because* ∼)

ample sentences of those expressions and of judging the usages of those expressions. Most of the 125 major functional expressions we consider in this paper are compound expressions which consist of one or more content words as well as functional words. As we introduced with the examples of Table 1, it is often the case that they have both a compositional *content word* usage as well as a non-compositional *functional* usage. For example, in Table 3, the expression “となると (to-naru-to)” in the sentence (2) has the meaning “ *that (something) becomes* ∼”, which corresponds to a literal concatenation of the usages of the constituents: the post-positional particle “と”, the verb “なる”, and the post-positional particle “と”, and can be regarded as a *content word* usage. On the other hand, in the case of the sentence (1), the expression “となると (to-naru-to)” has a non-compositional functional meaning “*if*”.

Based on the discussion above, in the current version of the corpus, we classify the usages of those expressions into two classes: *functional* and *content*. Here, *functional* usages include both non-compositional and compositional functional usages, although most of the functional usages of those 125 major expressions can be regarded as non-compositional. On the other hand, *content* usages include compositional content word usages only.

2.3 Procedure of Corpus Development

The corpus from which we collect example sentences of functional expressions is 1995 Mainichi newspaper text corpus (1,294,794 sentences,

Table 5: Number of Expressions classified by Rate of Functional Usage (for 127 Functional Expressions)

Grammatical Function Type		Rate of Functional Usage			
		100%	100∼90%	90∼50%	50∼0%
post-positional particle type	subsequent to predicate / modifying predicate	2	8	14	9
	subsequent to nominal / modifying predicate	18	9	7	3
	subsequent to predicate, nominal / modifying nominal	0	3	0	0
auxiliary verb type		32	14	6	2
total		52	34	27	14

Table 4: Number of Sentences collected from 1995 Mainichi Newspaper Texts (for 337 Expressions)

	# of expressions
$50 \leq$ # of sentences	187 (55%)
$0 <$ # of sentences < 50	117 (35%)
# of sentences $= 0$	33 (10%)

47,355,330 bytes). For each of the 337 expressions, m (set as 50 in the current version of the corpus) sentences are collected according to the following procedure.

1. The expression is morphologically analyzed by MeCab[6], and its morpheme sequence[7] is obtained.
2. The corpus is morphologically analyzed by MeCab, and m sentences which include the morpheme sequence of the expression are collected.
3. Each sentence is annotated with one of the usages *functional/content* by an annotator.
4. At least one researcher validates the annotation of the usage.
5. (For the collected m sentences, if the rate of functional usage is less than a threshold p, sentences with functional usage are collected by considering the constraints against the immediately preceding morpheme as well as those against immediately subsequent morpheme. Similarly, if the rate of content usage is less than a threshold q, sentences with content usage are collected.)[8]

3 Corpus of Japanese Functional Expressions: Current Status

First, Table 4 classifies the 337 expressions according to the number of sentences collected from the 1995 Mainichi newspaper text corpus. For more than half of the 337 expressions, more than 50 sentences are collected, although about 10% of the 377 expressions do not appear in the whole corpus. We have finished validation by at least one researcher for 127 expressions (out of those 187 expressions with more than 50 sentences), which include the 125 major ones. For each of the grammatical function types of functional expressions listed in Table 2, Table 5 classifies the expressions of that type according to the rate of functional usage. Roughly speaking, functional expressions of the auxiliary verb type tend to have high rate of functional usage. Among those of the post-positional particle types, those which funtion as case-marking particles also tend to have high rate of functional usage. These are partly because of the characteristics of the newspaper text.

4 A Corpus Browsing Tool

We also developed a tool for browsing the corpus of Japanese functional expressions. This tool is for browsing the example sentences of Japanese functional expressions and also for selectively retrieving sentences which satisfy certain constraints. Specific facilities of the tool are described as below.

First, the list of 125 major functional expressions as well as their variants (337 expressions in total) can be browsed. Those functional expressions are

[6] `http://chasen.org/~taku/software/mecab/`

[7] For those expressions whose constituent has conjugation and the conjugated form also has the same usage as the expression with the original form, the morpheme sequence is expanded so that the expanded morpheme sequences include those with conjugated forms.

[8] In the current version of the corpus, this requirement is ignored.

classified according to their grammatical functions as in Table 2.

Second, among those 337 expressions, for 127 expressions for which we have finished validation by a researcher, example sentences annotated with functional/content usage distinction are browsed. The tool also has a facility of showing the result of morphological analysis of each sentence, where the sentence is segmented into the sequence of morphemes annotated with their parts-of-speech. Figure 1 (a) shows the snapshot of the browser for the expression "～となると (to-naru-to)", whose usages are described in the sentences (1) and (2) of Table 3 of the paper. In this snapshot, for simplicity, parts-of-speech of the morphemes are annotated only to the morpheme immediately preceding the expression, and the part-of-speech tags are translated into English.

Third, example sentences can be selectively retrieved by specifying constraints on the parts-of-speech of the morphemes around the expression. Figure 1 (b) shows the snapshot of the browser specifying the parts-of-speech of the immediately preceding morpheme as noun or adverb. In this case, the retrieved example sentences are those with *content* usage, which is described in the sentence (2) of Table 3 of the paper. Similarly, Figure 1 (c) shows the snapshot of the browser specifying the parts-of-speech of the immediately preceding morpheme as verb or adjective or auxiliary verb. In this case, the retrieved example sentences are those with *functional* usage, which is described in the sentence (1) of Table 3 of the paper. This facility of selective retrieval with constraints is very useful for discovering effective features when developing a tool of recognizing and judging the usage of those functional expressions.

5 Concluding Remarks

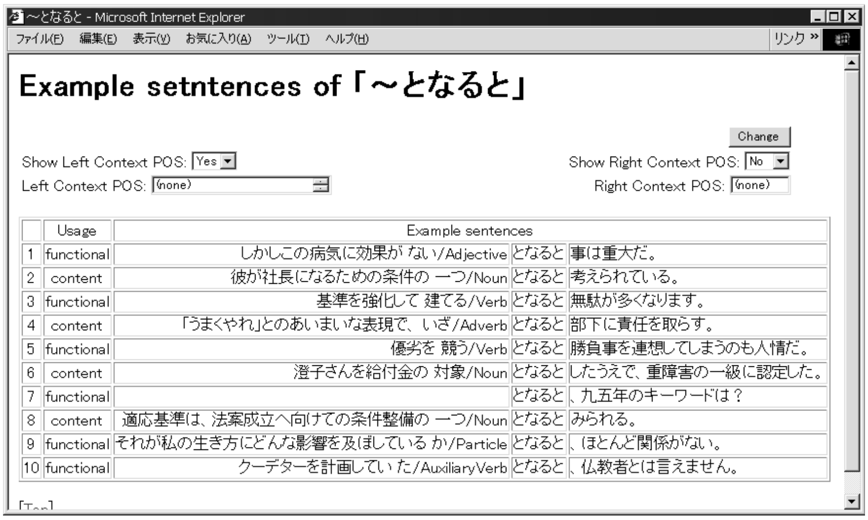
Aiming at being used as a corpus for training/testing a tool which properly recognizes and semantically interprets Japanese (compound) functional expressions, this paper studied how to develop a corpus of Japanese functional expressions. We focused on 125 major (mostly compound) functional expressions which have non-compositional usages, and reported current status of developing a corpus of those major functional expressions.

References

Group Jamashii, editor. 1998. *Nihongo Bunkei Jiten*. Kuroshio Publisher. (in Japanese).

Y. Morita and M. Matsuki. 1989. *Nihongo Hyougen Bunkei*, volume 5 of *NAFL Sensho*. ALC. (in Japanese).

National Language Research Institute. 2001. *Gendaigo Hukugouji Youreishu*. (in Japanese).


～となると - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
リンク »

Example setntences of「～となると」

Change

Show Left Context POS: Yes

Show Right Context POS: No

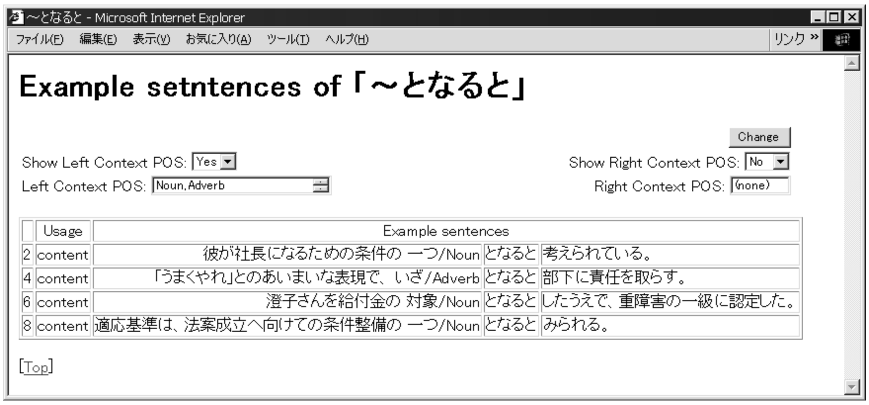
Left Context POS: (none)

Right Context POS: (none)

	Usage	Example sentences		
1	functional	しかしこの病気に効果が ない/Adjective	となると	事は重大だ。
2	content	彼が社長になるための条件の 一つ/Noun	となると	考えられている。
3	functional	基準を強化して 建てる/Verb	となると	無駄が多くなります。
4	content	「うまくやれ」とのあいまいな表現で、いざ/Adverb	となると	部下に責任を取らす。
5	functional	優劣を 競う/Verb	となると	勝負事を連想してしまうのも人情だ。
6	content	澄子さんを給付金の 対象/Noun	となると	したうえで、重障害の一級に認定した。
7	functional		となると	、九五年のキーワードは？
8	content	適応基準は、法案成立へ向けての条件整備の 一つ/Noun	となると	みられる。
9	functional	それが私の生き方にどんな影響を及ぼしている か/Particle	となると	、ほとんど関係がない。
10	functional	クーデターを計画してい た/AuxiliaryVerb	となると	、仏教者とは言えません。

[Top]

(a) Without Constraints on the POS of the Left Morpheme


～となると - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
リンク »

Example setntences of「～となると」

Change

Show Left Context POS: Yes

Show Right Context POS: No

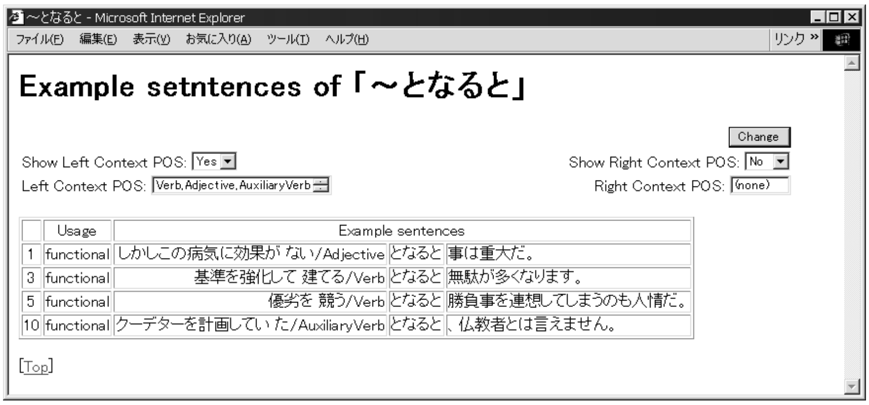
Left Context POS: Noun,Adverb

Right Context POS: (none)

	Usage	Example sentences		
2	content	彼が社長になるための条件の 一つ/Noun	となると	考えられている。
4	content	「うまくやれ」とのあいまいな表現で、いざ/Adverb	となると	部下に責任を取らす。
6	content	澄子さんを給付金の 対象/Noun	となると	したうえで、重障害の一級に認定した。
8	content	適応基準は、法案成立へ向けての条件整備の 一つ/Noun	となると	みられる。

[Top]

(b) With the Constraints on the POS of the Left Morpheme as “Noun,Adverb”.


～となると - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
リンク »

Example setntences of「～となると」

Change

Show Left Context POS: Yes

Show Right Context POS: No

Left Context POS: Verb,Adjective,AuxiliaryVerb

Right Context POS: (none)

	Usage	Example sentences		
1	functional	しかしこの病気に効果が ない/Adjective	となると	事は重大だ。
3	functional	基準を強化して 建てる/Verb	となると	無駄が多くなります。
5	functional	優劣を 競う/Verb	となると	勝負事を連想してしまうのも人情だ。
10	functional	クーデターを計画してい た/AuxiliaryVerb	となると	、仏教者とは言えません。

[Top]

(c) With the Constraints on the POS of the Left Morpheme as “Verb,Adjective,AuxiliaryVerb”.

Figure 1: Example Sentences of a Functional Expression “～となると (to-naru-to)”