A Morpho-Syntactic Analyzer of Controlled Japanese

Yukiko Sasaki Alam

Department of Digital Media
Hosei University, Tokyo

Proceedings of the GEAF 2007 Workshop

Tracy Holloway King and Emily M. Bender (Editors)

CSLI Studies in Computational Linguistics ONLINE

Ann Copestake (Series Editor)

2007

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

The proposed morpho-syntactic analyzer parses controlled Japanese texts such as articles in newspapers, technical magazines and professional journals and public documents that are transcribed wherever applicable by using *Joyo Kanji* (frequently used Chinese characters). The analyzer parses sentences in controlled Japanese texts into morpho-syntactic units, further dividing them into the content and the functional parts, and assigning a functional role or roles to each unit in the sentences. As the system is not equipped with a dictionary, the parsing algorithm is based on the orthographic characteristics of words and morphemes, and the role assignment to each unit is based on the functional elements located at the end of the unit, which is a feature of a Head-final language like Japanese. The system is a light-weight rule-based morpho-syntactic analyzer that could be a useful tool for natural language processing. As the system identifies syntactic units rather than individual morphemes, together with the functional and/or syntactic roles of the units, it would help a computational system understand the syntactic and functional structures of sentences, and eventually interpret the semantics of the sentences.

1 Introduction

There being no spaces between words in Japanese, a main concern of Japanese morphological and syntactic analyzers has been word segmentation. Word-breaking is a fundamental task in natural language processing for Japanese, and various approaches have been taken. While many morphological analyzers, notably *Juman* (Kurohashi and Nagao, 2003) and *Chasen* (Matsumoto et al., 2000), have concentrated on the segmentation of morphemes (such as prefixes, suffixes, inflections, Case markers, particles and the components of compound words), the current analyzer focuses on the segmentation of phrases and the identification of the functional roles of the phrases in sentences.

The proposed system is intended to parse controlled Japanese texts which are written wherever applicable by using *Joyo Kanji* (frequently used Chinese Characters), the members of which are determined by the Ministry of Education and Science of Japan. Such texts typically include articles in newspapers, technical magazines and journals, and official documents. The analyzer draws on the information of the orthographic types of words and morphemes used in sentences as well as linguistic knowledge of functional morphemes and words.

In the past, other researchers have also developed morphological analyzers exploiting the information of orthographic types of words and morphemes: to name a few, Asahara (2003), Kazama (2001), Kashioka et al (1998), and

Kameda (1996). Unlike such previous studies, however, the main focus of the present analyzer is not on phrase segmentation per se, but on identifying the functional roles of phrases played in the sentences.

Futhermore, unlike Kudo (2002), Sekine (2001), Uchimoto (2000), Kanayama et al (2000), and Haruno et al (1999), all of which are statistically modeled systems, the current analyzer runs on a purely rule-based algorithm. The purpose of the present paper is to demonstrate that a light-weight rule-based analyzer can successfully identify phrases in sentences, and determine the functional roles of the phrases.

The paper first gives an overview of the current system, and then describes the algorithm used in the system. Before concluding, it discusses what are the difficulties faced by the current system, and what areas need further research.

2 Overview of the Morpho-Syntactic Analyzer

The current analyzer runs by referring to the different orthographic types of Japanese words and morphemes. Japanese sentences are transcribed in several orthographic types: *Kanji* (Chinese characters), *Katakana* (phonetic characters for words of foreign origin), *Hiragana* (phonetic characters for words of Japanese origin, inflections, particles, etc.), Arabic numerals, the Roman alphabet, special symbols and punctuation.

The most important feature used by the current analyzer is that most functional morphemes in Japanese are transcribed in *Hiragana*, including all the particles indicating Case markers, verbal inflections, auxiliaries, and suffixes indicating different types of clauses. In addition to their special orthographic feature, unexceptionally these functional elements are located at the end of phrases, [1] thus marking phrase boundaries. The current analyzer is based on these two characteristics, i.e. *Hiragana*-transcribed functional elements and their phrase-final positions.

A sequence of *Kanji* characters followed by *Hiragana* characters would be a good candidate for a phrase, which consists of a content word followed by a functional element, as illustrated below:

[$_{\text{PP}}$ [$_{\text{NP}}$ content word in *Kanji*] [$_{\text{P}}$ functional element in *Hiragana*]]

It is relatively straightforward to identify such phrases, as demonstrated by the example output of the analyzer in Table 1.

[1] This is because Japanese is a Head-final language where the non-Head content part is followed by the Head functional part at all the morphological and syntactic levels of Japanese including words, phrases and clauses.

TABLE 1 A successful output of the analyzer

研究グループの鈴木宏志教授によると、
kenkyuu-guruupu-no-suzuki-hiroshi-kyouju-niyoruto,
research-group-of-suzuki-hiroshi-professor-according-to,
全国の盲導犬協会から
zenkoku-no-moudouken-kyoukai-kara
entire-country-of-guide-dog-association-from
盲導犬の口腔粘膜や血液の提供を
moudouken-no-koukou-nenmaku-ya-ketsueki-no-teikyou-o
guide-dog-of-oral-membrane-and-blood-of-donation-OBJECT
受け、遺伝子を解析する。
uke, idenshi-o-kaiseki-suru
receiving, gene-OBJECT-analysis-do

‘According to Professor Hiroshi Suzuki in the research group, they will analyze genes by receiving the oral membranes and the blood of guide dogs donated by the Associations of Guide Dogs in the entire country.’

CONTENT WORD	FUNCTION ELEMENT	GRAMMATICAL ROLE
研究グループ	*の*	名詞修飾句 (NOMINAL MODIFIER)
鈴木宏志教授	*によると*	出典 (SOURCE)
全国	*の*	名詞修飾句 (NOMINAL MODIFIER)
盲導犬協会	*から*	始点 (POINT OF DEPARTURE)
盲導犬	*の*	名詞修飾句 (NOMINAL MODIFIER)
口腔粘膜	*や*	列挙接続語 (CONJ – ETC)
血液	*の*	名詞修飾句 (NOMINAL MODIFIER)
提供	*を*	目的語 (OBJECT)
受	*け*	述語−接続形　(PREDICATE - CONJUNCTIVE)
中段 (Break)	、	読点 (COMMA)
遺伝子	*を*	目的語 (OBJECT)
解析	*する*	述語- 現在・未来　(PREDICATE - PRESENT or FUTURE)

Whenever a content word in *Kanji* (and/or *Katakana*) is followed only by a functional element in *Hiragana* (colored in red in Table 1), which is further followed by another content word in *Kanji* (and/*or Katakana*), word and phrase boundaries are clearly distinguished as in:

[$_{PP}$ *KENKYU-GURUUPU* (‘research group’)-*no* (nominal modifier marker)]
[$_{PP}$ *SUZUKI-HIROSHI-KYOUJU* (‘Prof. Hiroshi Suzuki’)-*niyoruto* (‘according to’)] …(omitted) …
[$_{PP}$ *KETSUEKI* (‘blood’)-*no* (nominal modifier marker)]
[$_{PP}$ *TEIKYO* (‘donation’)-*o* (Object marker)]
[$_{VP}$ *UK* (‘receive’)-*e* (verbal conjunctive form)]
[$_{PP}$ *IDENSHI* (‘genes’)-*o* (Object marker)]
[$_{VP}$ *KAISEKI* (‘analysis’)-*suru* (‘do’)].

The words in uppercase are written in *Kanji* or *Katakana*, while those in lowercase are in *Hiragana*.

As long as a content word is transcribed all in *Kanji* and/or *Katakana*, it is relatively straightforward to identify phrases, but unfortunately a content word can be transcribed by a mixture of *Kanji* and *Hiragana* characters, followed by *Hiragana*-written functional elements as in:

[$_{PP}$ [$_{NP}$ content word both in *Kanji and Hiragana*] [$_{P}$ functional element in *Hiragana*]],

or a content word can be transcribed all in Hiragana as in:

[$_{VP}$ [$_{V}$ content verb stem *in Hiragana*] [$_{INFL}$ verbal inflection in *Hiragana*]] [$_{CONJ}$ clause-final suffix in *Hiragana*].

Both undesirable cases are exemplified by the last phrase in Table 2 below.

TABLE 2 An unsuccessful output of the phrase analyzer

電力業界では、九州、四国が
denryoku-gyoukai-dewa, Kyuushuu, Shikoku-ga
electricity-industry-in-TOPIC, Kyushu-Shikoku-SUBJECT
０６年度採用を横ばいにとどめるが、…。
06-nendo-saiyou-o-yokobai-ni-todom-eru-ga, ….
06-fiscal-year-employment-OBJECT-the same level-in-keep-but,

‘In the electricity industry, Kyushu and Shikoku keep the employment in the 06 fiscal year in the same level, …’

CONTENT WORD	FUNCTION ELEMENT	GRAMMATICAL ROLE
電力業界	*では*	話題 (TOPIC)
中段 (Break)	、	読点　(COMMA)
九州	(省略)	次の内容要素と同じ (SAME AS NEXT CONTENT ELEMENT)
中段 (Break)	、	読点　(COMMA)
四国	*が*	主語 (SUBJECT)
０６年度採用	*を*	目的語 (OBJECT)
横	*ばいにとどめるが*	逆接接続語節 (CLAUSE-BUT)

The last row of Table 2 contains a content word in a mixture of *Kanji* and *Hiragana*, and the analyzer fails to recognize the end of the content word, leaving out part of the content word and placing it in the box for the functional element as: [[$_{\text{NP}}$ *YOKO*] [*bainitodomeruga*]]. The proper analysis would be:

[$_{\text{PP}}$ [$_{\text{NP}}$ *YOKObai* (‘same level’)] [$_{\text{P}}$ *ni* (postposition indicating state)]]
[$_{\text{VP}}$ [*todom* (‘keep’)]+[*eru* (non-past verbal inflection)]]
[$_{\text{CONJ}}$ [(preceding clause] [*ga* (suffix meaning ‘but’)]].

The failure is due to the content noun words that often consist of a mixture of *Kanji* and *Hiragana* as well as due to the fact that the content verb stem *todom* ‘remain’ was transcribed not in the regularly expected *Kanji* but exceptionally in *Hiragana*.

3 Algorithm of the Morpho-Syntactic Analyzer

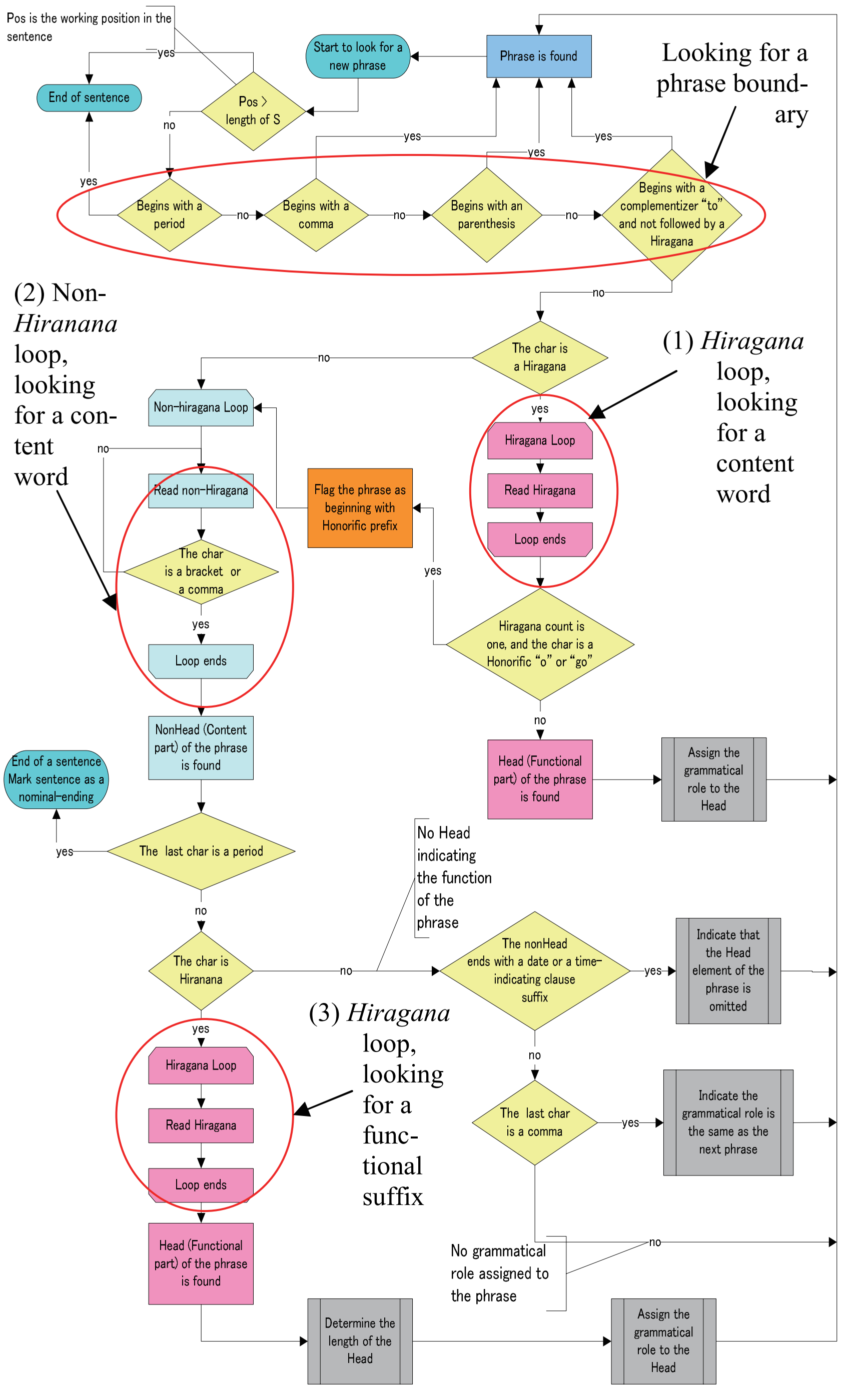
As the above two examples illustrate, the success of the present analyzer in detecting phrases depends upon whether phrases are (a) typical ones consisting of a content word in *Kanji* (and/or *Katakana*) followed by a functional element in *Hiragana*, or whether they are complex ones, for instance, (b) consisting of a complex functional element in *Hiragana* or whether they are atypical ones (c) containing a content word transcribed in *Hiragana*. The current analyzer attempts to handle (a) and (c).

The algorithm of the analyzer, illustrated in Figure 1, begins to look for a new phrase by checking special characters and suffixes including a period, a comma, a parenthesis, and a complementizer. It then checks for an atypical case of a phrase, i.e., whether the phrase begins with a *Hiragana* or a *Hiragana* sequence (the loop marked (1) in Figure 1). When it finds only one *Hiragana* followed by a non-*Hiragana* sequence, it asks whether the *Hiragana* is equal to an Honorific prefix or not. If it is, it flags the phrase as prefixed with an honorific, and goes on to process the non-*Hiragana* sequence that follows. On the other hand, when it finds more than one *Hiragana* that precedes a non-*Hiragana*, the *Hiragana* chunk is treated as a phrase and sent to the procedure to identify the grammatical role, primarily by analyzing the final portion that is expected to comprise a functional morpheme or morphemes.

When a phrase begins with a non-*Hiragana* character, the analyzer keeps reading it (the loop marked (2) in Figure 1) until it hits a comma, a bracket, a period or a *Hiragana*, and assigns the non-*Hiragana* sequence as the content part of the phrase. The algorithm then checks whether the non-Hiragana content part ends with a period. If it does, the phrase is determined to be the final nominal phrase of the sentence with the functional element omitted.

On the other hand, when the non-*Hiragana* content part is followed by a *Hiragana*, it is likely to embody a typical phrase structure, and the following *Hiragana* sequence is sent to the procedures so as to find out first (i) how much of the *Hiragana* sequence represents a functional element or elements, and then (ii) what is the functional role or the final functional role if there is more than one element.

When the non-*Hiragana* content part is not followed by *Hiragana*, the algorithm checks for two possible instances. First, when it finds the content part to be an expression of a date, time or a clause ending with a suffix denoting time, it marks the phrase as the one whose functional element is omitted. Second, when it finds the character in question to be a comma, it indicates that the phrase is without the functional part, and that the functional role is the same as that of the following phrase, because the comma is treated the same as a conjunction.


Pos is the working position in the sentence
yes
End of sentence
Start to look for a new phrase
Phrase is found
Pos > length of S
no
yes
yes
yes
yes
Begins with a period
no
Begins with a comma
no
Begins with an parenthesis
no
Begins with a complementizer "to" and not followed by a Hiragana
no
The char is a Hiragana
no
yes
Non-hiragana Loop
Hiragana Loop
no
Read non-Hiragana
Flag the phrase as beginning with Honorific prefix
Read Hiragana
Loop ends
The char is a bracket or a comma
yes
yes
Loop ends
Hiragana count is one, and the char is a Honorific "o" or "go"
no
NonHead (Content part) of the phrase is found
End of a sentence Mark sentence as a nominal-ending
Head (Functional part) of the phrase is found
Assign the grammatical role to the Head
yes
The last char is a period
No Head indicating the function of the phrase
no
The char is Hiranana
no
The nonHead ends with a date or a time-indicating clause suffix
yes
Indicate that the Head element of the phrase is omitted
yes
no
Hiragana Loop
Read Hiragana
The last char is a comma
yes
Indicate the grammatical role is the same as the next phrase
Loop ends
no
Head (Functional part) of the phrase is found
No grammatical role assigned to the phrase
Determine the length of the Head
Assign the grammatical role to the Head

Looking for a phrase boundary

(1) *Hiragana* loop, looking for a content word

(2) Non-*Hiranana* loop, looking for a content word

(3) *Hiragana* loop, looking for a functional suffix

FIGURE 1 Algorithm of the morpho-syntactic analyzer

4 Architecture of the Present System

The current system is constructed on an object-oriented design, comprising four Java programming language classes (programs): *MorphAlgorithm*, *Phrase*, *CharIdentifier* and *Grammar*. The *MorphAlgorithm* is the main program that runs on the algorithm introduced in the previous section and charted in Figure 1. The *Phrase* simulates a phrase (a syntactic unit), thus housing access methods to the Head and Complement.[2] The *CharIdentifier* provides the *MorphAlgorithm* with several methods that identify characters. The *Grammar* is instanced by the *MorphAlgorithm* to find out the functional role of the Head of a phrase. The grammatical roles identified are listed in the following tables.

TABLE 3 Case markers/Particles denoting thematic relations

Case/ Particles	Pronunciation	Functional role(s)
が	ga	Subject marker
を	o	Object marker
は	wa	Topic marker
で	de	Place/Instrument/conjunctive
へ	e	Goal
から	kara	Point of departure
まで	made	‘up to/till’
より	yori	Point of departure (formal or archaic)
として	toshite	‘as’ (Representative)
による	niyoru	Means
について	nitsuite	‘concerning’
によると	niyoruto	‘according to’

TABLE 4 Particles denoting conjunction

Particles	Pro-nunciation	Functional role(s)
も	mo	‘too’/conjunction for nouns
や	ya	conjunction for nouns (inclusive)
と	to	conjunction for nouns (exclusive)
か	ka	‘or’/Question particle
および	oyobi	conjunction for nouns (formal)

[2] It is based on the linguistic assumption that a phrase consists of a *Head* component and a *Complement* component.

TABLE 5 Particles that form modifiers of clauses

Particles	Pronunciation	Functional role(s)
ので	node	'because'
ため	tame	'because'
ために	tameni	'because'
けど	kedo	'although' (informal)
けれど	keredo	'although'
のに	noni	'even though'
ても	temo	'even though'
とき	toki	'when'
れば	reba	'if'
あいだ	aida	'while'

TABLE 6 Particles that form modifiers of verb phrases

Particles	Pronunciation	Functional role(s)
ものの	monono	'even though' (formal)
ながら	nagara	'while'
したまま	shitamama	'while doing'

TABLE 7 Particles denoting approximation or comparison

Particles	Pronunciation	Functional role(s)
ほど	hodo	'or so' (a little formal)
くらい	kurai	'or so'

In addition, the *Grammar* is able to identify the past and non-past affirmative and negative inflections of verbs and adjectives, and the conjunctive forms.

5 Discussion

Accuracy rates could be very high (a) when a text is written primarily in controlled Japanese (i.e., when the text is transcribed wherever applicable by using *Joyo Kanji*), (b) when the content words are followed by single functional elements, (c) when the content words are transcribed exclusively in *Kanji* and/or *Katakana*, and (d) when the text does not contain a long word in Hiragana such as a long adverb or conjunction. Table 1 shows such a sentence, and the accuracy rate is 100%. Accuracy rates become lower when the above conditions are not satisfied.

When the content words of a text are followed by a long sequence in *Hiragana* (counter to (b) above), the sequence is likely to comprise:

(i) more than one compound verbal suffix, or

(ii) a sequence of compound particles such as a Case marker followed by other particles.

It would not be very difficult to parse compound verbal suffixes consisting of long *Hiragana* sequence because of the following two facts: verbal and adjectival inflections in Japanese exhibit systematic paradigms, and such suffixes as causative, passive, aspectual and modal auxiliaries are aligned in rigid and thus predictable orders. To deal with a long predicate comprising more than one verbal suffix, a morphological analyzer is being prepared. Because this kind of a long predicate verb or adjective phrase occurs at least once in a sentence (that requires a predicate), and twice or more when the sentence contains a subordinate clause or clauses, significant improvement is expected, once the morphological analyzer for treating the complex verb phrase is incorporated into the current system.

Compound particles (for instance, consisting of a Case marker followed by a focus particle) also have a fairly rigid order, and it would be possible to analyze them in the system, once the orders are identified and implemented. However, it would be necessary to conduct a comprehensive linguistic study in this area for the successful identification of each functional element in sequence. At present a sequence of particles is treated as one chunk, the functional role of which is identified by the final particle.

When the content words in a sentence are transcribed in a mixture of *Kanji* (or *Katakana*) and *Hiragana* (counter to (c) above), the current system is unable to deal with such content words, because it does not have a dictionary. It would be interesting to investigate how frequently such words are transcribed in a mixture of *Kanji* (and/or *Katakana*) and *Hiragana*. Most adverbs and conjunctions are transcribed in *Hiragana*, even though there are some such as *OMOigakezu* ('by chance') and *sorenimoKAKAwarazu* ('in spite of that') that are transcribed in a mixture of *Kanji* and *Hiragana*. As a result, it is not so problematic to parse words in the two categories. Problems are caused mainly by nouns and compound verbs. However, nouns derived from verbs and adjectives are written in a mixture of the two characters: for instance, the noun *KAri* ('loan') derived from the verb *KAriru* ('borrow') and the noun *TANOshisa* ('pleasure') derived from the adjective *TANOshii* ('pleasant'). Such derivations are predictable, so it would be possible to prepare a morphological analyzer to handle them. Further research on derivations would be needed to improve the current system.

Quasi-compound verbs such as KAkeKOmu ('run into'), TAmeKOmu ('save up'), HIkiNObasu ('stretch out') and HIkiHANAsu ('separate') are problematic. They take the form of compound verbs, but they do not seem to be semantically compound verbs, because the original meanings of the following suffix verb or the preceding prefix verb are no longer independent but incorporated into the meanings of the main stem verbs. Therefore it is appropriate to handle such compound verbs as single verbs. As the current system aims at analyzing sentences into phrases, it is undesirable to treat them as separate verbs. This problem cannot be solved without a dictionary that lists quasi-compound verbs or a morphological engine that deals with such verbs.

The current system is not equipped with a dictionary, and does not contain an exhaustive list of adverbs and conjunctions. At present it identifies twenty-two adverbs and thirteen conjunctions. Since the numbers of adverbs and conjunctions are relatively definitive and not large, a future task would be to see how much improvement can be achieved, once an exhaustive list of words in these categories is incorporated into the system.

Finally, the system is unable to handle elements in parentheses, which are often semantically related to the preceding elements in various manners. Parenthetical elements could be explanations of the preceding abbreviations or vice versa. There are no formal clues to the understanding of the relations between the two elements. This area remains to be explored.

6 Conclusion

The current morpho-syntactic analyzer, without a dictionary, aims at parsing into phrases texts written in *Joyo Kanji* (frequently used Chinese characters). The phrases are divided into content and functional sections and functional roles are assigned. The results suggest that this light-weight phrase analyzer could be a useful tool for natural language processing, while awaiting further study and additional modules of implementation for better results. In machine translation, once the functional roles of phrases are identified, it will not be necessary to further break up phrases into morphemes, thus saving time and avoiding unnecessary parsing. Text understanding would be improved when the phrases of sentences are understood.

References

Asahara, Masayuki. 2003. *Corpus-based Japanese morphological analysis*. Ph.D. Thesis: Nara Institute of Science and Technology.

Fuchi, Takeshi and Shinichiro Takagi. 1998. Japanese Morphological Analyzer using Word C0-occurrence. In *Proceedings of the COLING*, pp. 409-413.

Haruno, Masahiko, Satoshi Shirai, and Yoshifumi Ooyama. 1999. Using Decision Trees to Construct a Practical Parser. *Machine Learning*, 34:131-149.

Kameda, Masayuki. 1996. A Portable & Quick Japanese Parser: QJP. In *Proceedings of the COLING*, pp. 616-621.

Kanayama, Hiroshi, Kentaro Torisawa, Yutaka Mitsuishi and Jun'ichi Tsujii. 2000. A Hybrid Japanese Parser with Hand-crafted Grammar and Statistics. In *Proceedings of the COLING*, pp. 411-417.

Kashioka, Hideki, Yasuhiro Kawata and Yumiko Kinjo. 1998. Use of Mutual Information Based Character Clusters in Dictionary-less Morphological Analysis of Japanese. In *Proceedings of the COLING*, pp. 658-662.

Kazama, Jun'ichi. 2001. *Adaptive Morphological Analysis with a Small Tagged Corpus*. Master Thesis: University of Tokyo.

Kurohashi, Sadao and Makoto Nagao. 2003. Building a Japanese Parsed corpus — while improving the parsing system. In Anne Abeille (ed.), *Treebank Building Using Parsed Corpora*, pp. 249-260. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Matsumoto, Yuji, Akira Kitauchi, TatsuoYamashita, Yoshitaka Hirano, Hiroshi Matsuda, Kazuma Takaoka and Masayuki Asahara. 2001. *Morphological Analysis System ChaSen version 2.2.4 Manual*. Nara, Japan: Nara Institute of Science and Technology.

Sekine, Satoshi. 2001. A Fast Japanese Sentence Analyzer. In *Proceedings of the First International Workshop on MultiMedia Annotation*.

Suzuki, Hisami, Chris Brockett, and Gary Kacmarcik. 2000. Using a Broad-Coverage Parser for Word-Breaking in Japanese. In *Proceedings of the COLING*, pp. 822-828.

Uchimoto, Kiyotaka, Masaki Murata, Satoshi Sekine and Hitoshi Isahara. 2000. Dependency model using posterior context. In *Proceedings of the Sixth International Workshop on Parsing Technologies*, pp. 321-322.