Evaluation of a Japanese CFG Derived from a Syntactically Annotated Corpus with Respect to Dependency Measures

Tomoya Noro[†] **Chimato Koike**[‡] **Taiichi Hashimoto**[†]
Takenobu Tokunaga[†] **Hozumi Tanaka**[#]
[†] Graduate School of Information Science and Engineering
Tokyo Institute of Technology, Tokyo
{noro@tt,taiichi@cl,take@cl}.cs.titech.ac.jp
[‡] Graduate School of Science and Engineering, Tokyo Institute of Technology, Tokyo
chimato@it.ss.titech.ac.jp
[#] School of Computer and Cognitive Sciences, Chukyo University, Nagoya
htanaka@sccs.chukyo-u.ac.jp

Abstract

Parsing is one of the important processes for natural language processing and, in general, a large-scale CFG is used to parse a wide variety of sentences. For many languages, a CFG is derived from a large-scale syntactically annotated corpus, and many parsing algorithms using CFGs have been proposed. However, we could not apply them to Japanese since a Japanese syntactically annotated corpus has not been available as of yet. In order to solve the problem, we have been building a large-scale Japanese syntactically annotated corpus. In this paper, we show the evaluation results of a CFG derived from our corpus and compare it with results of some Japanese dependency analyzers.

1 Introduction

Parsing is one of the important processes for natural language processing and, in general, a large-scale CFG is used to parse a wide variety of sentences. Although it is difficult to build a large-scale CFG manually, a CFG can be derived from a large-scale syntactically annotated corpus. For many languages, large-scale syntactically annotated corpora have been built (e.g. the Penn Treebank (Marcus et al., 1993)), and many parsing algorithms using CFGs have been proposed.

However, such a syntactically annotated corpus has not been built for Japanese as of yet. Dependency analysis is preferred in order to analyze Japanese sentences (dependency relation between Japanese phrasal unit, called *bunsetsu*) (Kurohashi and Nagao, 1998; Uchimoto et al., 2000; Kudo and Matsumoto, 2002), and only a few studies about Japanese CFG have been conducted. Since many efficient parsing algorithms for CFG have been proposed, a Japanese CFG is necessary to apply the algorithms to Japanese.

We have been building a large-scale Japanese syntactically annotated corpus to derive a Japanese CFG for syntactic parsing (Noro et al., 2004a; Noro et al., 2004b). According to the result, a CFG derived from the corpus can parse sentences with high accuracy and coverage. However, as mentioned previously, dependency analysis is usually adopted in Japanese NLP, and it is difficult to compare our result with results of other dependency analysis since we evaluated our CFG with respect to phrase structure based measure. Although we evaluated with respect to dependency measure as a preliminary experiment in order to compare, the scale was quite small (evaluated on only 100 sentences) and the comparison was unfair since we did not use the same evaluation data.

In this paper, we show an evaluation result of a CFG derived from our corpus and compare it with results of other Japanese dependency analyzers. We used the Kyoto corpus (Kurohashi and Nagao, 1997) for evaluation data, and chose KNP (Kurohashi and Nagao, 1998) and CaboCha (Kudo and Matsumoto, 2002) for comparison.

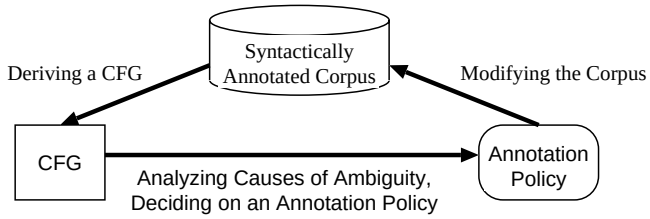


Figure 1: Procedure of building a syntactically annotated corpus

2 Annotation Policy

In this section, we start by introducing our policy for annotating a Japanese syntactically annotated corpus briefly. The details are given in (Noro et al., 2004a; Noro et al., 2004b)

Although a large-scale CFG can be easily derived from a syntactically annotated corpus, such a CFG has a problem that it creates a large-number of parse results during syntactic parsing (i.e. high ambiguity). A syntactically annotated corpus should be built so that the derived CFG would create less ambiguity.

We have been building such Japanese corpus by using the following method (Figure 1):

1. Derive a CFG from an existing corpus.
2. Analyze major causes of ambiguity.
3. Determine a policy for modifying the corpus.
4. Modify the corpus according to the policy and derive a CFG from it again.
5. Repeat steps (2) - (4) until most problems are solved.

We focused on two major causes of ambiguity:

Lack of Syntactic Information: Some syntactic information which is important for syntactic parsing might be lost during the CFG derivation since CFG rules generally represent only structures of subtree with the depth of 1 (relation between a parent node and some child nodes).

Need for Semantic Information: Not only syntactic information but also semantic information is necessary for disambiguation in some cases.

To avoid the first cause, we considered which syntactic information is necessary for syntactic parsing and added the information to each intermediate node in the structure. On the other hand, we considered ambiguity due to the second cause better be left to the subsequent semantic processing since it is difficult to reduce such ambiguity without recourse to semantic information during syntactic parsing. This can be achieved by representing the ambiguous cases as the same structure. We assume that syntactic analysis based on a large-scale CFG is followed by semantic analysis, and the second cause of ambiguity is supposed to be disambiguated in the subsequent semantic processing.

The main aspects of our policy are as follows:

Verb Conjugation: Information about verb conjugation is added to each intermediate node related to the verb (cf. "SPLIT-VP" in (Klein and Manning, 2003) and "Verb Form" in (Schiehlen, 2004)).

Compound Noun Structure: Structure ambiguity of compound noun is represented as the same structure regardless of the meaning or word-formation as Shirai et al. described in (Shirai et al., 1995).

Adnominal and Adverbial Phrase Attachment: Structure ambiguity of adnominal phrase attachment is represented as the same structure regardless of the meaning while structure ambiguity of adverbial phrase attachment is distinguished by meaning. In case of a phrase like "*watashi no chichi no hon* (my father's book)", the structure is same whether the adnominal phrase "*watashi no* (my)" attaches to the noun "*chichi* (father)" or the noun "*hon* (book)". On the other hand, in case of a sentence

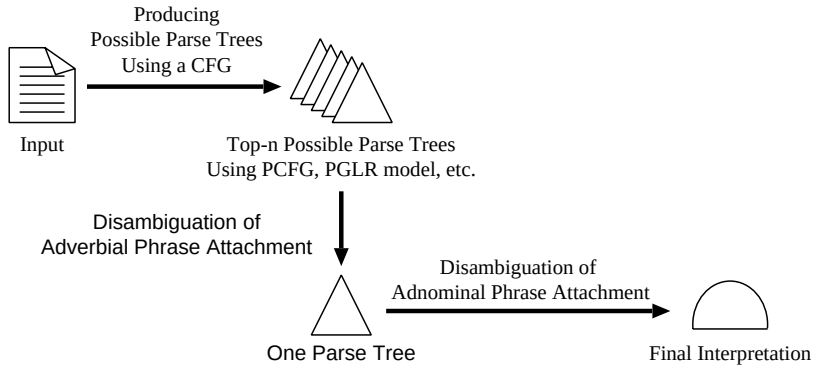


Figure 2: Procedure in the subsequent processing

$$\text{Segmentation Accuracy} = \frac{\text{\# sentences segmented into } \textit{bunsetsu} \text{ correctly}}{\text{\# all sentences}}$$

$$\text{Dependency Accuracy} = \frac{\text{\# correct dependency relations}}{\text{\# all dependency relations}}$$

$$\text{Sentence Accuracy} = \frac{\text{\# sentences determined all relations correctly}}{\text{\# sentences segmented in } \textit{bunsetsu} \text{ correctly}}$$

Figure 3: Dependency measures

like "*kare ga umi wo egaita e wo katta*", we distinguish the structure according to whether the adverbial phrase "*kare ga* (he)" attaches to the verb "*egaita* (paint)" (it means "I bought a picture of a sea painted by him") or the verb "*katta* (buy)" (it means "he bought a picture of a sea").

Conjunctive Structure: Conjunctive structure is not specified during syntactic parsing, instead their analysis is left for the subsequent processing (contrary to (Kurohashi and Nagao, 1994)).

We have decided to deal with adnominal phrase attachment and adverbial phrase attachment separately in our policy since we believe that a different algorithm should be used to disambiguate them. In the subsequent processing, we assume that adverbial phrase attachment would be disambiguated by choosing one parse tree among the results at first, and adnominal phrase attachment would be disambiguated by choosing one interpretation among all of interpretations which the parse tree represents (Figure 2).

We used the EDR corpus (EDR, 1994) for developing our annotation policy, and annotated 8,911 sentences in the corpus and 20,190 sentences in the RWC corpus (Hashida et al., 1998). In the following evaluation, we used the latter one.

3 Experimental Setup

As mentioned previously, in general, analyzing dependency relations between *bunsetsu* is preferred in Japanese, which makes it difficult to compare the result by the CFG with the result by dependency analysis. In order to compare with other dependency analysis, we evaluated our derived CFG with respect to dependency measures shown in Figure 3. Note that sentences which are not segmented into *bunsetsu* correctly are dropped from the evaluation data when we evaluate dependency accuracy and sentence accuracy.

A CFG is derived from all sentences in our corpus, with which we parsed 6,931 sentences (POS sequences) in the Kyoto corpus [1] by MSLR parser (Shirai et al., 2000). The Kyoto corpus has an-

[1] On average, 8.89 *bunsetsu* in a sentence.

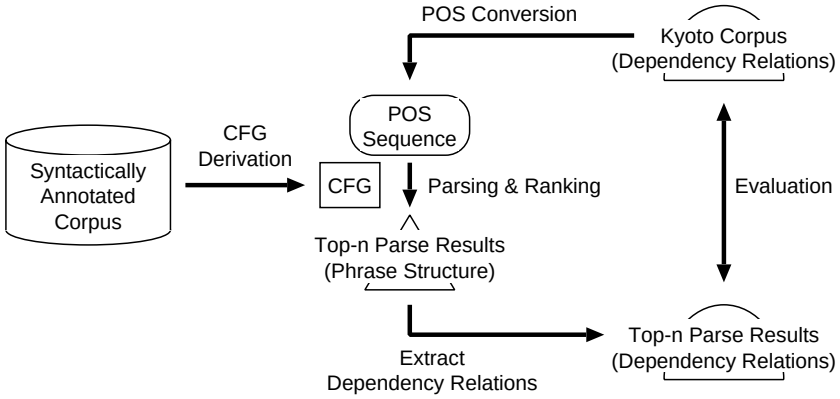


Figure 4: Evaluation with respect to dependency measure

notation in terms of dependency relations among *bunsetsu*, and it is usually used for evaluation of dependency analysis. The parser is trained according to probabilistic generalized LR (PGLR) model (Inui et al., 2000) (all sentences are used for training), and parse results are ranked by the model.

The experiment was carried out as follows (Figure 4):

1. Convert POS tags automatically to the RWC tag set.
2. Parse the POS sequence using a CFG derived from our corpus.
3. Rank the parse results by PGLR model and pick up the top-n parse results.
4. Extract dependency relations among *bunsetsu* for each result.
5. Choose the result which is closest to the gold-standard and evaluate it.

Since the tag set of the Kyoto corpus is different from that of the RWC corpus, a POS conversion in step (1) is necessary. It is a rule-based conversion, and the accuracy is about 80%. It seems that the low conversion accuracy would damage the evaluation result. We will discuss this issue in the next section.

In the 4th step of the experimental procedure, we determine boundaries of *bunsetsu* and dependency relations among the *bunsetsu* in a sentence with the CFG rules included in the phrase structure of the sentence. Some CFG rules in our CFG indicate positions of *bunsetsu* boundaries. For example, a CFG rule "NP $\rightarrow$ AdnP NP" ("NP" and "AdnP" stand for a noun phrase and an adnominal phrase respectively) indicates that there is a boundary of *bunsetsu* between the two phrases in the right-hand side of the CFG rule (i.e. between the noun phrase and the adnominal phrase), and that a *bunsetsu* including the head word of the adnominal phrase depends on a *bunsetsu* including the head word of the noun phrase. An example of "*Nihon teien no nagame ga subarashii* (The view of the Japanese garden is wonderful)" is shown in Figure 5.

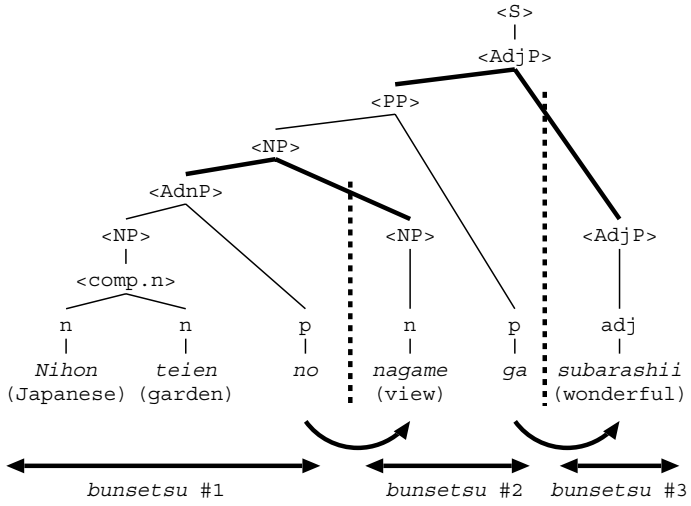
Structure ambiguity of adnominal phrase attachment needs to be disambiguated in extracting dependency relations in step (4) since it is represented as the same structure according to our policy [2]. We disambiguate adnominal phrase attachment based on one of the following assumptions:

NEAREST: Every ambiguous adnominal phrase attaches to the nearest noun among the nouns which the phrase could attach to.

BEST: Choose the best noun among the nouns which could be attached to (assume that disambiguation of adnominal phrase attachment was done correctly) [3].

[2] Since dependency relations are not categorized in the Kyoto corpus, it is difficult to know how many relations representing adnominal phrase attachment are included in the evaluation data. On the other hand, among the top parse results ranked by PGLR model (i.e in case of $n = 1$ in section 4), about 34.1% of all dependency relations represent adnominal phrase attachment, and about 23.4% of them (i.e. about 8.0% of all relations) remain ambiguous.

[3] We choose the best noun automatically by referring to



Bunsetsu No.	Word Sequence	*Bunsetsu* Which is Depended on
1	*nihon teien no*	2
2	*nagame ga*	3
3	*subarashii*	—

Figure 5: Extracting Dependency Relations from a Pharse Structure

"NEAREST" is a quite simple way for disambiguation, and it would be the baseline model. On the other hand, since we assume that structure ambiguity of adnominal phrase attachment is supposed to be disambiguated in the subsequent semantic processing, "BEST" would be the upper bound and we could not overcome the accuracy even if the disambiguation was done perfectly in the subsequent processing.

To take two noun phrases "*watashi no chichi no hon* (my father's book)" and "*watashi no kagaku no hon* (my book on science)" as examples (the correct answer is that the adnominal phrase "*watashi no* (my)" attaches to the noun "*chichi* (father)" in the former case, and attaches to the noun "*hon* (book)" in the latter case), "NEAREST" attaches to the adnominal phrase "*watashi no*" to the nouns "*chichi*" and "*kagaku* (science)" regardless of their meanings. "BEST" attaches the adnominal phrase to the noun "*chichi*" in the former case, and attaches to the noun "*hon*" in the latter case.

Although structure ambiguity of compound noun is also represented as the same structure regardless of the meaning or word-formation, we have nothing to do with the structure ambiguity since a *bunsetsu* is a larger unit than a compound noun. Furthermore, since dependency relations are not categorized, we do not have to care about whether two *bunsetsu* have conjunctive relation with each other or not.

In order to compare our result with that of other dependency analyzers, we used two well-known Japanese dependency analyzers, KNP and CaboCha, and analyzed dependency structure of the sentences in the same evaluation data set. In both cases, POS tagged sentences are used as the input. Since CaboCha uses the same tagset as the RWC corpus, we converted POS tags in the same way as step (1) in our experimental procedure. On the other hand, since KNP uses the tagset adopted by the Kyoto corpus, POS tags do not have to be converted in case of analyzing by KNP.

4 Results

Table 1 shows the results when $n = 1$, which means the top parse result of each sentence is used for evaluation. In this case, "NEAREST" means only PGLR model was used for disambiguation without any other information (e.g. lexical infor-

the Kyoto corpus. If the noun which is attached to in the Kyoto corpus is not in the candidates, we choose the nearest noun (i.e. "NEAREST").

Table 1: Segmentation, dependency, and sentence accuracy ($n = 1$)

	Segmentation	Dependency	Sentence
NEAREST	65.68%	87.88%	50.47%
BEST	65.68%	90.27%	57.73%
KNP	96.90%	91.32%	60.07%
CaboCha	84.88%	92.88%	64.48%

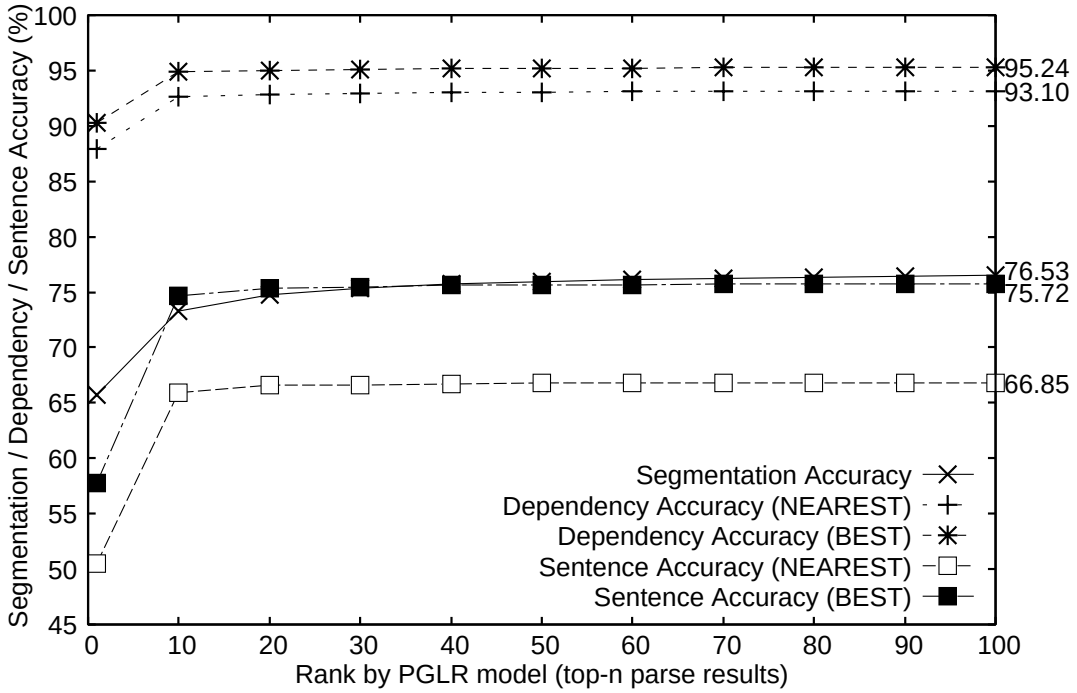


Figure 6: Segmentation, dependency, and sentence accuracy ($n = 1 \ldots 100$)

mation, semantic information, etc.) On the other hand, "BEST" means only disambiguation of adnominal phrase attachment was done in the subsequent processing. Results by KNP and CaboCha are shown in the same table for comparison.

As seen from Table 1, accuracy is still lower than KNP and CaboCha even if disambiguation of adnominal phrase attachment was done correctly in the subsequent processing. However, in this case, we do not use any information but PGLR model for disambiguation of any relations except adnominal phrase attachment (i.e. adverbial phrase attachment).

Next, assuming that disambiguation of other relations, we carried out another evaluation changing n from 1 to 100. The result is shown in Figure 6. Dependency accuracy could achieve about 95.24% for "BEST", which exceeds the dependency accuracy by KNP and CaboCha, if choosing the best result among top-100 parse results ranked by PGLR model would be done correctly in the subsequent processing [4]. From the results, we can conclude the accuracy will increase as soon as lexical and semantic information is incorporated in the subsequent processing [5].

However, segmentation accuracy is still significantly lower. The main reasons are as follows:

POS Conversion Error: As mentioned previously, we converted POS tags automatically since the POS system of the Kyoto corpus is

[4] Even if only top-10 parse results are considered, our CFG have a possibility to outperform KNP and CaboCha

[5] In some studies, it is said that lexical information has little impact on accuracy (Bikel, 2004). However, we think some lexical information is useful for disambiguation, and it is necessary to consider what kind of lexical information could improve the accuracy.

different from that of the RWC corpus. However, accuracy of the conversion is not high (about 80%). Since we used only POS information and did not use any word information for parsing, the result can be easily affected by the conversion error. Segmentation accuracy by CaboCha is also a little lower than accuracy by KNP. Since POS tags were converted in the same way, we think the reason is same. However, the difference between the accuracy by KNP and CaboCha is smaller since CaboCha uses not only POS information but also word information.

Difference in Segmentation Policy: There is difference in *bunsetsu* segmentation policy between the Kyoto corpus and our corpus. For example:

1. ***3 gatsu 31 nichi gogo 9 ji 43 fun goro,*** *jishin ga atta*
   (An earthquake occurred **at around 9:43 p.m., March 1st**.)
2. ***gezan suru no wo*** *miokutta*
   (We gave up **going down the mountain**.)

In the former case, the underlined part is segmented into 5 *bunsetsu* ("*3 gatsu*", "*31 nichi*", "*gogo*", "*9 ji*", and "*43 fun goro,*") in the Kyoto corpus, while it is not segmented in our corpus. On the other hand, in the latter case, the underlined part is segmented into 2 *bunsetsu* ("*gezan suru*" and "*no wo*") in our corpus, while it is not segmented in the Kyoto corpus. By correction of these two types of error, segmentation accuracy improved by 4.35% (76.53% $\rightarrow$ 80.88%) and dependency accuracy improved by 0.61% (95.24% $\rightarrow$ 95.85%).

5 Conclusion

We have been building a large-scale Japanese syntactically annotated corpus. In this paper, we evaluated a CFG derived from the corpus with respect to dependency measure. We assume that parse results created by our CFG is supposed to be re-analyzed in the subsequent processing using semantic information, and the result shows that parsing accuracy will increase when semantic information is incorporated.

We also compared our result with other dependency analyzers, KNP and CaboCha. Although dependency accuracy of our CFG cannot reach those of KNP and CaboCha if only PGLR model is used for disambiguation, it would exceed if disambiguation in the subsequent processing was done correctly.

As future work, since we assume that the parse results created by our CFG are re-analyzed in the subsequent processing, we need to integrate the subsequent processing into the current framework. Collins proposed a method for reranking the output from an initial statistical parser (Collins, 2000). However, it is not enough for us since we represent some ambiguous cases as the same structure (we need to consider the ambiguity included in each parse result). Our policy has been considered with several types of ambiguity: structure of compound noun, adnominal phrase attachment, adverbial phrase attachment and conjunctive structure. We are planning to provide each method individually and integrate them into a single process.

Although we attempt to re-analyze after parsing, it seems that some problem should be solved before parsing. For example, ellipsis often occurs in Japanese. It is difficult to deal with ellipsis (especially, postpositions and verbs) in a CFG framework, resulting in higher ambiguity. It would be helpful if the positions where some words are omitted in a sentence were detected and marked in advance.

References

Daniel M. Bikel. 2004. A distributional analysis of a lexicalized statistical parsing model. In *2004 Conference on Empirical Methods in Natural Language Processing*, pages 182–189.

Michael Collins. 2000. Disriminative reranking for natural language parsing. In *17th International Conference on Machine Learning*, pages 175–182.

EDR, 1994. *EDR Electronic Dictionary User's Manual*, 2.1 edition. In Japanese.

Koichi Hashida, Hitoshi Isahara, Takenobu Tokunaga, Minako Hashimoto, Shiho Ogino, and Wakako Kashino. 1998. The RWC text databases. In

The First International Conference on Language Resource and Evaluation, pages 457–461.

Kentaro Inui, Virach Sornlertamvanich, Hozumi Tanaka, and Takenobu Tokunaga. 2000. Probabilistic GLR parsing. In Harry Bunt and Anton Nijholt, editors, *Advances in Probabilistic and Other Parsing Technologies*, pages 85–104. Kluwer Academic Publishers.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430.

Taku Kudo and Yuji Matsumoto. 2002. Japanese dependency analysis using cascaded chunking. In *CONLL 2002*.

Sadao Kurohashi and Makoto Nagao. 1994. A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures. *Computational Linguistic*, 20(4):507–534.

Sadao Kurohashi and Makoto Nagao. 1997. Kyoto university text corpus project. In *the 3rd Conference for Natural Language Processing*, pages 115–118. In Japanese.

Sadao Kurohashi and Makoto Nagao. 1998. Building a Japanese parsed corpus while improving the parsing system. In *the first International Conference on Language Resources and Evaluation*, pages 719–724.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Tomoya Noro, Taiichi Hashimoto, Takenobu Tokunaga, and Hozumi Tanaka. 2004a. Building a large-scale japanese CFG for syntactic parsing. In *The 4th Workshop on Asian Language Processing*, pages 71–78.

Tomoya Noro, Taiichi Hashimoto, Takenobu Tokunaga, and Hozumi Tanaka. 2004b. A large-scale japanese CFG derived from a syntactically annotated corpus and its evaluation. In *The 3rd Workshop on Treebanks and Linguistic Theories*, pages 115–126.

Michcael Schiehlen. 2004. Annotation strategies for probabilistic parsing in German. In *the 20th International Conference on Computational Linguistics*, pages 390–396.

Kiyoaki Shirai, Takenobu Tokunaga, and Hozumi Tanaka. 1995. Automatic extraction of Japanese grammar from a bracketed corpus. In *Natural Language Processing Pacific Rim Symposium*, pages 211–216.

Kiyoaki Shirai, Masahiro Ueki, Taiichi Hashimoto, Takenobu Tokunaga, and Hozumi Tanaka. 2000. MSLR parser – tools for natural language analysis. *Journal of Natural Language Processing*, 7(5):93–112. In Japanese.

Kiyotaka Uchimoto, Masaki Murata, Satoshi Sekine, and Hitoshi Isahara. 2000. Dependency model using posterior context. In *6th International Workshop on Parsing Technologies*.