Automatic Acquisition of Lexical-Functional Grammar Resources from a Japanese Dependency Corpus*

Masanori Oya* and Josef van Genabith*

*National Centre for Language Technology and School of Computing,
Dublin City University, Dublin, Ireland
{moya, josef}@computing.dcu.ie

Abstract. This paper describes a method for automatic acquisition of wide-coverage treebank-based deep linguistic resources for Japanese, as part of a project on treebank-based induction of multilingual resources in the framework of Lexical-Functional Grammar (LFG). We automatically annotate LFG f-structure functional equations (i.e. labelled dependencies) to the Kyoto Text Corpus version 4.0 (KTC4) (Kurohashi and Nagao 1997) and the output of of Kurohashi-Nagao Parser (KNP) (Kurohashi and Nagao 1998), a dependency parser for Japanese. The original KTC4 and KNP provide unlabelled dependencies. Our method also includes zero pronoun identification. The performance of the f-structure annotation algorithm with zero-pronoun identification for KTC4 is evaluated against a manually-corrected Gold Standard of 500 sentences randomly chosen from KTC4 and results in a pred-only dependency f-score of 94.72%. The parsing experiments on KNP output yield a pred-only dependency f-score of 82.08%.

Keywords: Lexical-Functional Grammar, Japanese, automatic linguistic resource acquisition, zero-pronoun identification

1. Introduction

We present a method to automatically annotate Lexical-Functional Grammar (LFG)-style functional structure equations (labelled dependencies) on the unlabelled Kyoto University Text Corpus version 4 (KTC4) (Kurohashi and Nagao 1997), to acquire more abstract and (somewhat) less language-dependent LFG f-structure representations for Japanese sentences. We apply the algorithm to enrich the output of a Japanese dependency parser (Kurohashi-Nagao Parser, KNP) (Kurohashi and Nagao 1998), to construct f-structure representations for KNP output; the enriched parser output is available for further cross-linguistic research or applications such as machine translation.

Our annotation method is based on the assumption that non-configurational, relatively free word-order languages, of which Japanese is one example, do not require phrase structure trees as an indispensable level of linguistic representation. Rather, the rich morphological information on each unit in a sentence, along with the unlabelled dependency between syntactic units in KTC4 and KNP output, provides us with as much information as what can be deduced from phrase-structure trees in other configurational, fixed word-order languages.

Our method provides zero pronoun identification as a preliminary process for long distance dependency (LDD) resolution, based on the morphology of verbs and on the probability of subcategorization frames, associated with particular verbs.

This paper has the following structure: in Section 2 we summarize the background of this research, including LFG and related work. In Section 3, we describe in detail our method of automatic annotation of f-structure functional equations on KTC4, and show how we approach the problem of zero-pronoun identification and present results of our f-structure annotation and

* We gratefully acknowledge support from Science Foundation Ireland grant 04/IN/I527 for the research reported in this paper.

parsing experiments. We discuss the overall results and their implications in Section 4, and conclude in Section 5.

2. Background

2.1 Lexical-Functional Grammar

Lexical-Functional Grammar (LFG) (Bresnan 2001; Dalrymple 2001) is a syntactic theory in which there are two levels of representation: c-structures are phrase-structure trees, and f-structures are attribute-value matrices encoding abstract grammatical relations such as subject, object, oblique or adjunct, mapped from the c-structure through functional equations annotated to c-structure nodes. Figure 1 is the c-structure for the sentence "Taro went to Seoul", and Figure 2 is the f-structure for the same sentence:

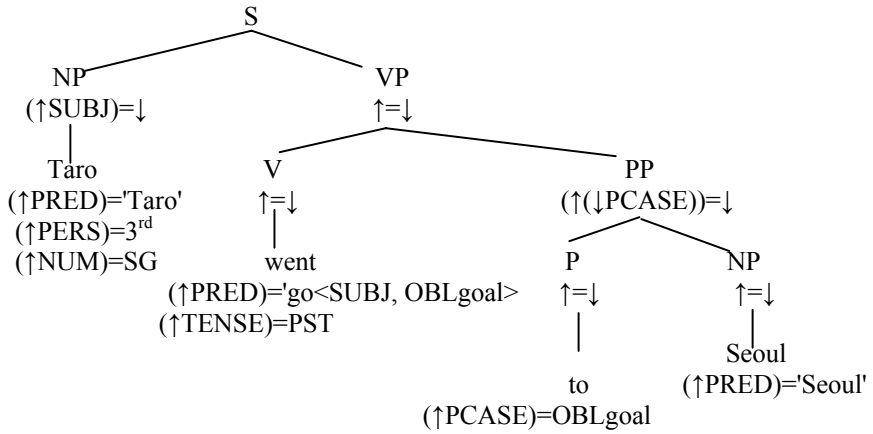

Figure 1: The c-structure for "Taro went to Seoul".

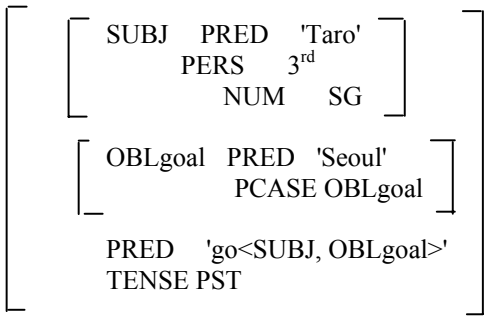

Figure 2: The f-structure for "Taro went to Seoul".

C-structures capture language-specific properties, such as word order and the hierarchical grouping of phrases, while f-structures are more abstract and somewhat more language-independent representations of surface grammatical relations (labelled dependencies). LFG is used in various fields of NLP research, such as Machine Translation (Owczarzak et al. 2007) or Question Answering (Judge et al. 2006).

2.2 Automatic Induction of LFG Resources

Treebank-based automatic acquisition of deep linguistic resources has been one of the important topics in the field of NLP (Hockenmeier et al., 2002; Cahill et al. 2002; Miyao et al. 2003). It is expected to overcome the shortcomings of manual production of linguistic resources: manual development is time-consuming, expensive and limited in terms of coverage. Ideally, automatic methods are expected to be able to induce linguistic resources that are deep, including not only syntactic properties of given sentences but also semantic properties such as predicate-argument structures and long-distance dependencies (LDDs). Several methods to

achieve this goal have been developed to date, based on different grammatical formalisms like Combinatory Categorial Grammar (CCG) (Steedman, 2000), Head-Driven Phrase Structure Grammar (HPSG) (Pollard and Sag, 1994), and LFG. For example, Hockenmaier and Steedman (2002) presented an algorithm to translate the Penn-II Treebank into a CCG-style Treebank. Miyao and Tsujii (2005) developed probabilistic models for parsing with HPSG grammars acquired from the Penn-II treebank. Cahill et al. (2002, 2003, 2004) developed a method for automatic annotation of LFG f-structure on the Penn-II Treebank. The approach of Cahill et al. (2002, 2003, 2004) is as follows: first, LFG functional equations are automatically annotated on the phrase-structure trees in the English treebank. The equations specify the constraints on the f-structure mapping from the c-structure. The equations are collected and sent to a constraint solver to generate f-structures for these sentences. Long-distance dependencies (LDD) are resolved on f-structures using LDD path frequencies acquired from the f-structure annotated treebank and automatically acquired subcategorization frames (O'Donovan et al. 2004). This method has been applied to several languages other than English, including Chinese and German (Burke et al. 2004; Cahill et al. 2003).

3. Acquisition of LFG Resources from a Japanese Text Corpus

A wide-coverage LFG grammar for Japanese (Masuichi et al.2003) has been manually developed in the ParGram project (Butt et al. 2002) along with grammars for a number of other languages. To the best of our knowledge our research is the first method for the automatic treebank-based acquisition of deep Japanese LFG resources, focusing on morphological information and on unlabelled dependency relationships among the syntactic units in a sentence, as provided by an existent wide-coverage Japanese corpus.

We use KTC4 as the corpus from which wide-coverage LFG resources are acquired. The method we develop implements the idea that the part-of-speech tags on each morpheme and the unlabelled dependency tags on each syntactic unit in KTC4 provide us with enough information for constructing what Cahill et al. (2003, 2004) call "proto" f-structures for the texts in the corpus, without employing context-free grammar syntactic trees. This idea is inspired by the difference in the type of syntactic information encoded in the English Penn treebank (Marcus et al. 2004) and that in the Japanese text corpus. This difference reflects language-particular properties of Japanese. Japanese is a non-configurational language which has relatively free-word order and where grammatical functions of syntactic phrases are shown not by the word order (as in English), but by the morphology of each syntactic phrase, such as case particles for specifying the grammatical function of an NP (e.g., the case particle "-wo" specifies that the noun phrase with this particle is an OBJ of the verb on which this noun phrase is dependent), or verbal inflections for specifying tense or modal information, and sometimes for the distinction between relative clauses and sentential modifiers. According to this morphological information and unlabelled dependency links as represented in KTC4, f-structure functional equations are automatically annotated on each syntactic unit of the sentences in KTC4; these equations are sent to a constraint solver to construct the f-structures for these sentences.

3.1 Automatic Annotation of f-Structure Functional Equations to KTC4 Representations

This section describes how the method developed in this research augments KTC4 unlabelled dependency representations with the information necessary to construct "proto" f-structures, through f-structure functional equations which are resolved by a constraint solver.

KTC4 encodes morphological and syntactic information by tags in the format displayed in Figure 3, for the example sentence "*Taro ga souru ni itta* (Taro went to Seoul)". The parenthesized lines provide glosses in English, which are not contained in KTC4:

```
#S-ID:950101001-001
* 0 2D
太郎  たろう * 名詞 人名    * *      (Taro   Noun   Person**)
が    が      *  助詞 格助詞 * *     (ga     particle case **)
```

```
* 1 2D
ソウル そうる * 名詞 地名 * *        (souru "Seoul" * Noun Place**)
に * 助詞 格助詞 * *        (ni particle Case**)
* 2 -1D
行った いった 行く 動詞 * 子音動詞 過去形 (itta "went" Verb * ConsonantStem pst)
EOS
```

Figure 3: KTC4 annotation for the sentence "Taro ga souru ni itta (Taro went to Seoul)"

The first line in Figure 3 is the sentence ID. Lines which start with a star are the first lines of syntactic units. The representations also specify the unit ID number and the target unit ID number of the unit on which this unit is dependent, and the character after the target unit ID specifies the type of dependency: D denotes a direct dependency, P a coordinate relation and A an apposition. Note that apart from this, dependencies are unlabelled. If the unit does not have any target unit, then it is the root unit of the sentence, and this is indicated by "-1D".

The f-structure functional annotation algorithm assumes that each one of the syntactic units in the KTC4 representation corresponds to one sub-f-structure and that they combine with each other according to the unlabelled dependency relation provided in the KTC4 representation, to constitute one f-structure for the sentence as a whole. In other words, what is projected from one node in a phrase-structure tree of a configurational language, such as English, is projected from one syntactic unit of Japanese. The labels in the dependencies in the f-structure representation (i.e. the LFG grammatical functions) are captured from the morphological particle information in the KTC4 representation. For example, the first syntactic unit (indexed 0) in Figure 3 contains the case-particle "-ga" which is a subject marker for a noun, and this unit depends on the last syntactic unit (indexed 2), meaning that the information in the first unit provides the value of a SUBJ attribute in the f-structure associated with the head verb (indexed 2). The second syntactic unit (indexed 1) contains the case-particle "-ni" which signals that the syntactic unit functions as an oblique argument of the predicate. The last syntactic unit (indexed 2) has a morpheme whose part of speech is verb. Since its inflection form is the past form, the tense value is past. As it does not have any morpheme which specifies the statement type and style, by default this sentence is a declarative statement in plain style. The dependency relation tag (-1D) specifies that it does not have any target unit on which it is dependent, hence this unit is the root unit of the sentence. From these pieces of information, the f-structure annotation algorithm automatically annotates each syntactic unit with appropriate equations for its grammatical function, for its predicate value, and for some other lexical values such as tense.

For the example sentence in Figure 3 above, the first unit is annotated as the subject of the sentence, the second unit is the oblique-case marked argument, and the last unit is the main predicate of the f-structure of the whole sentence. The output of the annotation algorithm is shown in the Figure 4, and the f-structure generated from these functional equations by the constraint solver is shown in Figure 5:

```
#S-ID:950101001-001
* 0 2D
太郎 たろう * 名詞 人名 * *        (Taro Noun Person **)
が が * 助詞 格助詞 * *        (ga particle Case **)
F0:pred ='Taro',
F0:case='ga',
F2:subj=F0,
* 1 2D
ソウル そうる * 名詞 地名 * *        (souru "Seoul" * Noun Place**)
に * 助詞 格助詞 * *        (ni particle Case**)
F1:pred='Seoul',
F1:case='ni',
F2:obl=F1,
* 2 -1D
行った いった 行く 動詞 * 子音動詞 過去形 (itta 'went' iku Verb * ConsonantStem pst)
```

F2:pred='iku',
F2:tns='pst',
F2:stmt='decl',
F2:style='plain'.
EOS

Figure 4: KTC4 annotation for the sentence “Taro ga souru ni itta (Taro went to Seoul)” with functional equasions.

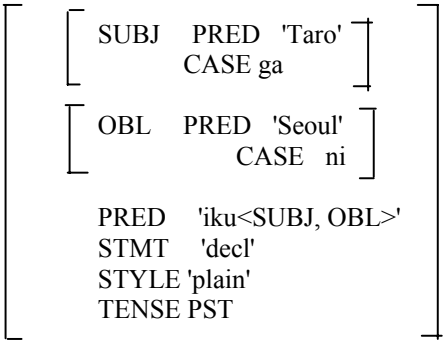

Figure 5: The f-structure for the sentence “Taro ga souru ni itta (Taro went to Seoul)”

The advantage of this method is that the annotation algorithm can be applied not only to the tagged sentences in KTC4, but also to raw texts using JUMAN (a Japanese morphological analyzer), and the KNP parser. This is because KTC4 has been developed along with the development of the KNP parsing system (Kurohashi and Nagao 1998). Using the method on JUMAN-KNP output, we can annotate KNP parser output with LFG f-structure functional equations. The f-structures from parser output can be employed in various applications such as Machine Translation or Question Answering.

Table 1 evaluates the f-structures generated by the method against a 500-sentence gold standard. Details are described in Section 3.3. Overall weighted precision for all features is 95.66%, recall 86.02% and f-score 90.59%:

Table 1: Evaluation results for all features without zero pronoun identification:

Feature	Precision	Recall	F-score		Feature	Precision	Recall	F-score
adj	644/668 = 96	644/662 = 97	97		obj_p	1005/1053 = 95	1005/1033 = 97	96
adjform	207/208 = 100	207/209 = 99	99		obl	321/346 = 93	321/577 = 56	70
asp	120/121 = 99	120/121 = 99	99		padj	1080/1135 = 95	1080/1118 = 97	96
case	990/1065 = 93	990/1088 = 91	92		pfrm	96/96 = 100	96/96 = 100	100
caus	8/8 = 100	8/8 = 100	100		progform	120/121 = 99	120/121 = 99	99
cj	350/361 = 97	350/356 = 98	98		ptrav	394/395 = 100	394/398 = 99	99
comp	300/308 = 97	300/325 = 92	95		prtcj	0/0 = 0	0/29 = 0	0
coord_form	125/162 = 77	125/158 = 79	78		prtcnj	1/1 = 100	1/2 = 50	67
copulaform	105/109 = 96	105/109 = 96	96		prtcs	124/126 = 98	124/129 = 96	97
exrl	0/0 = 0	0/32 = 0	0		rel	281/337 = 83	281/287 = 98	90
mod	66/67 = 99	66/70 = 94	96		sadj	236/246 = 96	236/270 = 87	91
nadv	83/83 = 100	83/83 = 100	100		stmt	1380/1477 = 93	1380/1421 = 97	95
neg	145/151 = 96	145/151 = 96	96		style	1398/1472 = 95	1398/1418 = 99	97
negform	140/151 = 93	140/151 = 93	93		subj	283/288 = 98	283/1418 = 20	33
nform	32/33 = 97	32/32 = 100	98		sufvform	67/67 = 100	67/67 = 100	100
noda	28/34 = 82	28/34 = 82	82		tns	767/770 = 100	767/875 = 88	93
nsa	908/910 = 100	908/915 = 99	100		topic	330/345 = 96	330/350 = 94	95

num	27/30 = 90	27/27 = 100	95	vform	502/507 = 99	502/508 = 99	99
obj	410/414 = 99	410/558 = 73	84	voice	91/96 = 95	91/96 = 95	95

3.2 Zero-Pronoun Identification

The language-particular properties of Japanese mentioned above allow us to induce LFG resources from the KTC4 corpus or KNP parser output. However, zero-pronouns cause a major problem for the method. Along with ordinary pronouns, Japanese has zero pronouns, which have no morphological or phonological realization but, for all intents and purposes, function as pronouns in other languages. Since they are used quite often both in spoken and written Japanese, identification of them is one of the issues in Japanese NLP (Kawahara et al. 2004a, 2004b, among others). Moreover, zero pronoun identification is required to resolve LDDs which is one of the important research topics in automatic induction of deep linguistic resources.

Cahill et al. (2004) present a method to automatically obtain approximations of LDD resolution for LFG resources acquired from a treebank. It uses verb subcategorization frames and LDD paths between coindexed materials (e.g., wh-phrase and its gap), both of which are extracted from the f-structures automatically generated for the Penn-II treebank.

Since KTC4 does not annotate zero pronouns on all the texts (only about 5,000 sentences are annotated with zero pronouns), LDD resolution based on KTC4 necessarily is divided into two steps: the first is zero-pronoun identification and the second is their resolution. The method we have presented in Section 3.1 does not detect the presence of zero pronouns. Hence, we have devised an additional method, making use of morphological and syntactic information in the corpus, in order to identify zero pronouns. If the method is able to identify zero pronouns in the KTC4 corpus, then it can also be applied to the output of KNP, which also does not identify zero pronouns. In this paper we concentrate on zero pronoun identification. Zero pronoun and LDD resolution will be addressed in future research.

3.3 Experiment 1: Zero Pronoun Identification in KTC4

The quality of the f-structures automatically acquired from KTC4 is evaluated against gold-standard f-structures which are manually created for a set of 500 test sentences randomly chosen from the first half of KTC4 (Table 1). 200 sentences randomly chosen from the second half of KTC4 are used as a development set. For both the development and test sets, f-structure functional equations are annotated automatically by the method without zero-pronoun identification and then their f-structures are manually corrected. The zero pronouns in the 500 Gold Standard f-structures are added manually, based on the context in which each of them appeared in the original text, verbal morphology, and A Japanese Lexicon (Ikehara et al. 1999), a hand-coded Japanese case-frame dictionary. Table 2 shows the numbers of the core arguments in the Gold Standard f-structures and the numbers of zero pronouns of each core argument (SUBJ, OBJ, OBL).

Table 2: The numbers of the core arguments and the numbers of zero pronouns of each core grammatical function in the Gold standard f-structures:

Grammatical functions	token numbers	token numbers of pro
SUBJ	1411	1121 (approx. 79% of all SUBJ)
OBJ	536	122 (approx. 22% of all OBJ)
OBL	568	199 (approx. 35% of all OBL)

We have developed five methods for zero pronoun identification. The **first method** is the Null method, which ignores zero pronouns and nothing is added to the f-structure annotation output.

The **second method** is the Simplistic method, which simply adds zero pronouns SUBJ-pro, OBJ-pro and OBL-pro whenever full NPs with the particle "-ga", "-wo" or "-ni" are missing for local verbs, regardless of the case frame of the verb.

The **third method** is the Morphological method, which uses a list of verbs whose morphology specifies their transitivity. The list is automatically constructed from KTC4 (except for the Gold Standard sentences), based on the morphology of the verbs. For some Japanese verbs, morphological information of the verb indicates unambiguously whether it is a transitive or intransitive verb. Verbs which end with "su", whether su is the verb-ending morpheme or part of the verb-ending morpheme, are all unambiguously transitives. For those verbs which are unambiguously transitives, if they appear in KTC4 without an object NP, then an object zero pronoun OBJ-pro is assumed to be present. F-structure equations are added automatically which specify that the verb takes an object whose predicate value is "pro".

The **fourth method** is the probabilistic method, which uses a list of verbs with high transitivity rate (the rate that each verb appears with an OBJ NP dependent on it). The problem of the morphology-based method 3 is its low coverage. The total number of verb types in KTC4 is 3506, and the total number of verb tokens is 95383. The number of morphologically unambiguously transitive verb types which have their intransitive counterpart is 1286, and their token number is 33911. As for the other 2220 verb types (61472 verb tokens), their morphology does not tell us their valency.

KTC4 does not annotate the text with tags which specify the valency of verbs. Therefore, an approach must be developed to determine the valency of verbs which are not unambiguously marked by their morphology, and one of the possible approaches to achieve this task is to look at the syntactic environment in which the verb appears; e.g., we can estimate the probability that a verb whose morphology does not specify its valency is used transitively in the corpus. The phrase "used transitively in the corpus" means that the verb takes a noun phrase which is dependent on the verb and the noun phrase has the case particle "-wo". If the probability that a verb lemma is used as a transitive verb is higher than a certain threshold, and if it appears without object in a given sentence, then the appropriate f-structure equations are automatically added to the f-structure for the sentence. The list of verbs and their transitivity rate is automatically acquired from the second half of KTC4, which does not contain Gold Standard sentences. The threshold is 0.3 in this experiment, i.e., the list includes verbs whose transitivity rates are above 0.3.

The **fifth method** is a combination of methods 3 and 4: add to the list of method 3 those verbs whose morphology does not specify their transitivity but that have a high transitivity rate.

In all methods, 500 f-structures generated by different zero-pronoun identification methods are converted into dependency triples of a grammatical function, a predicate and its argument: for example, a dependency triple "subj(go, Taro)" which is obtained from a sentence "Taro went to Seoul" means that the subject of the verb "go" is "Taro". The triples are compared with the dependency triples of the Gold Standard f-structures, and the precision, recall and f-score for each grammatical function are calculated using the software of Crouch et al. (2002).

Table 3 shows the evaluation results of the five methods. The figures in the parentheses are recall, precision and f-score of zero pronouns only. "Pred-only" means the result includes the precision, recall and f-score of dependency triples of the predicates, arguments and adjuncts in the 500 test sentences, but not atomic features such as tense, mood, aspect features.

In all methods except for Method 1, SUBJ-pro is added simplistically; since every verb subcategorises for a subject, hence if a clause lacks a subject NP, then pro-SUBJ is added into the clause. However, this result does not yield 100% accuracy because of the wrong annotation of functional equations, especially those on nominal predicates functioning as sentential adjuncts, hence more cleaning up operations are required.

From all the results, method 5 performs best for OBJ zero pronoun identification. The results of zero pronoun identification for OBL are lower than that for OBJ, because of the ambiguity of "ni" marked NPs. This particle can be used as the OBL case marker, or as a postposition which functions as a temporal or a locative adverbial.

Table 3: Results of Experiment 1

		Precision	Recall	F-score
Method 1 (null)	Pred-only	95.71	75.18	84.22
	SUBJ	97.56(0)	19.84(0)	32.97(0)
	OBJ	98.79(0)	73.16(0)	84.06(0)
	OBL	93.31(0)	58.05(0)	71.57(0)
Method 2 (simplistic)	Pred-only	78.22	95.51	86.01
	SUBJ	98.64(98.91)	97.38(97.60)	98.00(98.25)
	OBJ	39.47(14.13)	97.67(95.80)	56.22(24.62)
	OBL	39.25(19.10)	89.94(88.46)	54.65(31.41)
Method 3 (morphological)	Pred-only	95.75	92.69	94.2
	OBJ	92.83(71.55)	88.01(58.04)	90.35(64.09)
	OBL	92.48(88.05)	68.28(28.36)	78.55(42.90)
Method 4 (probabilistic)	Pred-only	95.76	92.92	94.32
	OBJ	97.97(87.09)	77.99(18.88)	86.94(31.03)
	OBL	82.17(63.92)	82.32(67.30)	82.24(65.56)
Method 5 (combination)	Pred-only	95.08	94.37	94.72
	OBJ	93.26(76.29)	91.59(72.02)	92.41(74.09)
	OBL	84.46(68.65)	81.97(66.34)	83.19(67.47)

Results for Method 5 for all features (rather than pred-only) are as follows: precision is 95.61%, recall 94.68% and f-score 95.15%. Compared to Table 1, this shows a marked increase due to the effect of zero-pronoun identification.

3.4 Experiment 2: Zero Pronoun Identification in KNP Parser Output

Experiment 2 explores how the methods in Experiment 1 can identify zero pronouns in raw texts, using KNP, a Japanese dependency parser. We stripped off the dependency and other tags in the 500 Gold Standard sentences and parsed them with KNP. The parser output is automatically annotated with f-structure functional equations, and zero pronouns are identified using the same methods as in Experiment 1. The output f-structures are converted into triples and compared to the Gold Standard triples. Table 4 shows the results of each method. The general tendency of recall, precision and f-scores of SUBJ are the same as Experiment 1:

Table 4: Results of Experiment 2

		Precision	Recall	F-score
Method 1 (null)	Pred-only	83.57	79.06	72.37
	SUBJ	79.93(0)	16.59(0)	27.47(0)
	OBJ	89.63(0)	66.54(0)	76.37(0)
	OBL	85.38(0)	51.64(0)	64.35(0)
Method 2 (simplistic)	Pred-only	67.96	82.77	74.64
	SUBJ	89.60(92.16)	88.84(90.93)	89.21(92.04)
	OBJ	35.88(12.88)	88.90(87.41)	51.12(22.45)
	OBL	34.68(16.63)	79.89(78.36)	48.36(27.43)
Method 3 (morphological)	Pred-only	83.28	80.74	81.99
	OBJ	85.26(65.95)	77.63(43.35)	81.26(52.31)
	OBL	84.96(82.85)	61.69(27.88)	71.47(41.72)
Method 4 (probabilistic)	Pred-only	83.13	80.62	81.86
	OBJ	89.31(84.00)	70.30(14.68)	78.67(24.99)
	OBL	72.82(53.33)	72.44(57.69)	72.62(55.42)
Method 5 (combination)	Pred-only	82.91	81.27	82.08
	OBJ	85.82(68.91)	77.99(44.75)	81.71(54.23)
	OBL	72.82(56.33)	72.44(57.69)	72.62(57.00)

4. Discussion

Method 5 yields the best pred-only f-score for both KTC4 and KNP parser output. The experiments show that the morphology-based approach and the probability-based approach

improve the f-scores of the annotation algorithm in terms of the pred-only f-scores of the sentence as a whole.

However, these two approaches do not yet identify zero pronouns as precisely as expected, and the improvement remains moderate; for example, the f-score of zero-pronoun OBJ in Method 5 in parsing is only slightly above 82%.

5. Conclusion

This paper presents a method for automatically acquiring LFG resources from the KTC4 Japanese text corpus and KNP parser output along with a basic zero-pronoun identification method. The performance of the f-structure annotation algorithm for KTC4 is evaluated against a manually-corrected Gold Standard of 500 sentences randomly chosen from KTC4 and the evaluation results in a pred-only dependency f-score of 94.72%. The parsing experiments on KNP output yields a pred-only dependency f-score of 82.08%. The results show that LFG resources automatically acquired from a Japanese text corpus can be improved through zero-pronoun identification.

References

Bresnan, J. 2001. *Lexical-Functional Syntax*. Blackwell Publishers, Oxford.

Burke, M., A. Cahill, M. McCarthy, J. van Genabith, and A. Way. 2002. Evaluating Automatic F-Structure Annotation for the Penn-II Treebank. *Proceedings of TLT 2002, Treebanks and Linguistic Theories*, pp. 42-60.

Burke, M., O. Lam, A. Cahill, R. Chan, R. O'Donovan, A. Bodomo, J. van Genabith and A. Way. 2004. Treebank-based Acquisition of a Chinese Lexical-Functional Grammar. *Proceedings of the 18th Pacific Asia Conference on Language, Information and Computation (PACLIC-18)*, 161-172.

Burke, M., A. Cahill, M. McCarthy, R. O'Donovan, J. van Genabith and A. Way. 2004. Evaluating Automatic F-Structure Annotation for the Penn-II Treebank. *Journal of Language and Computation; Special Issue on "Treebanks and Linguistic Theories"*, eds., E. Hinrichs and K.Simov, Kluwer Academic Press, 523-547.

Butt, M., H. Dyvik, T. H. King, H. Masuichi, and C. Roher. 2002. The Parallel Grammar Project. *Proceedings of COLING-2002 Workshop on Grammar Engineering and Evaluation*, pp. 1-7.

Cahill, A., M. McCarthy, J. van Genabith, and A. Way. 2002. Automatic Annotation of the Penn-II Treebank with LFG F-Structure Information. *Proceedings of Third International Conference on Language Resources and Evaluation, Las Palmas, Spain, June 5th, 2002,* pp. 8-15

Cahill, A., M. Forst, M. McCarthy, R. O' Donovan, C. Rohrer, J. van Genabith and A. Way. 2003. Treebank-Based Multilingual Unification-Grammar Development. *Proceedings of the Workshop on Ideas and Strategies for Multilingual Grammar Development*, at the 15th European Summer School in Logic Language and Information, pp.17-24.

Cahill, A., M. Burke, R. O'Donovan, J. van Genabith and A. Way. 2004. Long-Distance Dependency Resolution in Automatically Acquired Wide-Coverage PCFG-Based LFG Approximations. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pp. 320-327.

Crouch, R., R. M. Kaplan, T. H. King, and S. Riezler. 2002. A Comparison of Evaluation Metrics for a Broad-Coverage Stochastic Parser. *Proceedings of the "Beyond PARSEVAL" Workshop at the 3rd International Conference on Language Resources and Evaluation (LREC'02), Las Palmas, Spain.*

Dalrymple, M. 2001. *Lexical-Functional Grammar*. Academic Press, London.

Hockenmaier, J. and M. Seedman. 2002. Acquiring Compact Lexicalized Grammars from a Cleaner Treebank. *Proceedings of Third International Conference on Language Resources and Evaluation, Las Palmas, Spain, June 5th, 2002,* pp. 1974–1981.

Ikehara, S, M. Miyazaki, S. Shirai, A. Yokoo, H. Nakaiwa, K. Ogura, Y. Ooyama, and Y. Hayashi. 1999. *Nihongo Goi Taikei*. “A Japanese Lexicon". Iwanami Shoten, Tokyo.

Judge, J., A. Cahill and J. van Genabith. 2006. QuestionBank: Creating a Corpus of Parse Annotated Questions. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pp. 597-504.

Kawahara, D. and S. Kurohashi. 2002. Fertilization of Case Frame Dictionary for Robust Japanese Case Analysis. *Proceedings of the 19th International Conference on Computational Linguistics*, pp. 425-431.

Kawahara, D. and S. Kurohashi. 2004a. Zero Pronoun Resolution Based on Automatically Constructed Case Frames and Structural Preference of Antecedents. *Proceedings of the 1st International Joint Conference on Natural Language Processing (IJCNLP-04),* pp.334-341,

Kawahara, D. and S. Kurohashi. 2004b. Improving Japanese Zero Pronoun Resolution by Global Word Sense Disambiguation. *Proceedings of the 20th International Conference on Computational Linguistics (COLING2004)*, pp.343-349.

Kawahara, D. and S. Kurohashi. 2005. Gradual Fertilization of Case Frames. *Journal of Natural Language Processing,* vol.12, no.2, 109-131.

Kudoh, T. and Y. Matsumoto. 2002. Japanese Dependency Analysis using Cascaded Chunking. *Proceedings of the 6th Conference on Natural Language Learning 2002 (COLING 2002 Post-Conference Workshops)*, pp. 63-69.

Kurohashi, S. and M. Nagao. 1997. Kyoto daigaku text corpus project. *Proceedings of the Third Conference of Natural Language Processing*, pp.115-118.

Kurohashi, S. and M. Nagao. 1998. Building a Japanese Parsed Corpus while Improving the Parsing System. *Proceedings of the 1st International Conference on Language Resources and* Evaluation, pp. 719-724.

Marcus M., G. Kim, M. A. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz, and B. Schasberger. 1994. The Penn Treebank: Annotating Predicate Argument Structure. *Proceedings of the ARPA Workshop on Human Language Technology*. Princton, NJ., pp. 110-115.

Masuichi, H., and T. Okuma. 2003. Japanese Parser on the Basis of the Lexical-Functional Grammar Formalism and its Evaluation. *Journal of Natural Language Processing* vol. 10, pp. 79-109.

Miyao, Y. and J. Tsujii. 2005. Probabilistic Disambiguation Models for Wide-Coverage HPSG Parsing. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Ann Arbor, Michigan.* pp. 83-90.

O'Donovan, R. 2006. *Automatic Extraction of Large-Scale Multilingual Lexical Resources*. Ph.D. thesis, Dublin City University.

O'Donovan, R., M. Burke, A. Cahill, J. van Genabith, and A. Way. 2004. Large-Scale Induction and Evaluation of Lexical Resources from the Penn-II Treebank. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, July 21-26, Barcelona, Spain, pp. 368-375.

Owczarzak, K., J. van Genabith, and A. Way. 2007. Labelled Dependencies in Machine Translation Evaluation. *Proceedings of ACL 2007 Workshop on Statistical Machine* Translation, pp. 104-111.

Pollard, C., and I. A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press.

Steedman, Mark. 2000. *The Syntactic Process*. The MIT Press, Cambridge Mass.

Yoshioka, T., H. Yoshimura, H. Masuichi and T. Okuma. 2003. A Proposal for Experience Knowledge Recycle System. *Proceedings of the 17th Annual Conference of the Japanese Society for Artificial Intelligence, 2003*.