Statistical Models for Case Ambiguity Resolution in Korean

Kihwang Lee

Doctor of Philosophy
Institute for Communicating and Collaborative Systems
School of Informatics
University of Edinburgh

2005

Abstract

This thesis deals with the resolution of case ambiguity in Korean. Even though Korean is a case marked language, in which phonetically recognisable case markers (case particles) mark cases explicitly, nominal words without any accompanying case particles are used frequently in naturally occurring texts and speech. When the case particles are not present, it is basically a matter of conjecture to infer the grammatical function of the nominal words. The position of a nominal word itself cannot give much help as Korean is a relatively free word order language. The case ambiguity problem has brought a great controversy in Korean linguistics and has been regarded as an unavoidable obstacle for automatic processing of the Korean language.

The aim of this thesis is to tackle the case ambiguity problem in Korean with statistical methods. To achieve the aim we pursue the following objectives.

First, through an examination of the relevant theoretical work, we precisely define the realm of the case ambiguity problem in Korean. We also clearly identify the case particles that are involved in case ambiguitiy problem.

Second, we clearly specify our knowledge-lean training data construction method. We also attempt to measure the effectiveness of the data collection method by applying the method to two treebanks of Korean.

Third, we suggest two case decision methods for the task of case ambiguity resolution: discrete case decision method and sequential case decision method. In the discrete case decision method, each case ambiguity in a sentence is treated in isolation. For this method we use statistical classifiers based on simple joint probabilistic models that can be easily extended. We incorporate two new features, the list of neighbouring case particles and the distance between the focus nominal and the predicate, which have never been used in previous approaches. In the sequential case decision method, every case decision in a sentence is treated in the context of a series of case decisions that take place in the sentence. This method is similar to other sequential category assignment tasks such as part-of-speech tagging. Thus we adopt the well-known Markov chain tagging model.

Finally, our statistical case ambiguity resolution models are evaluated by comparing the outputs of the system applied to a test set with the multiple human annotations on the test set. $Kappa$ is used to measure the pairwise agreements between the system outputs and human annotations. From the evaluation results, we show the effectiveness of the two new features.

As a conclusion, we present the contributions and the limitations of our approach to the case ambiguity problem. Several possible future directions are also laid out.

Acknowledgements

O give thanks unto the LORD; for he is good: for his mercy endureth for ever. (Psalms 118:2)

During seven years of PhD study here in Edinburgh, I received so much help from many people both academically and personally.

First of all, I am deeply grateful to my principal supervisor, Dr Henry S. Thompson, for his encouragement and support. Dr Thompson truly understood my situation and helped me overcome all the difficulties I encountered taking all possible steps. Most of all, he prayed for me. I would also like to thank my second supervisor, Dr Miles Osborne, who introduced the beauty of machine learning to me. I owe much to my former supervisor, Dr Chris Brew now at Ohio State University in the United States. It was he who kindly responded to my initial contact and encouraged me to come to Edinburgh.

I am grateful to my examiners Prof Anne de Roecke and Dr Jon Oberlander. They read my thesis line by line from top to bottom and gave me invaluable comments and suggestions for improving the thesis. My gratitude also goes to the dissertation draft committee members Prof Mark Steedman and Dr Steven Clark.

I can't thank enough my family and friends who supported and prayed for me. My parents Prof Sangsup Lee and Prof Jungmai Kim financially and morally supported my study and my family's stay in Edinburgh. Their endless loving support and prayer always stood firmly behind me. My parents in law Mr Ho-Young Jung and Mrs I-Soon Kim thought of me as their own son and gave me a warm and loving support. I will never forget the joy of opening the packages they frequently sent to us. I am also thankful to my brother and sister in law Mr Kihwal Lee and Mrs Heejin Kim, and brother in law Mr Jin-Yong Jung who heartily supported me.

I would like to express my gratitude to Prof Kishim Nam who first taught me linguistics and gave me a full support when I decided to study abroad. I am also grateful to Prof Hasoo Kim and Prof Sanggyu Seo for their guidance and support. I would like to express my thanks to Prof Ik-Hwan Lee and Prof Min-Haeng Lee for their concern and support.

I have also received big support and help from many members of the Korean community in Edinburgh. I can't include all the names here, but I should mention three couples Dr Jae-Young Lee and Mrs Soyang Lim, Dr Yongmin Lee and Mrs Miran Yoo, and Mr Byunghwa Lee and Mrs Sun-A Kim. Their support and compaionship were just like oases in the desert for my family.

Lastly, I thank my wife Youngsoo Jung for her sacrifice and tears for me. Without her loving care and prayer, I would not be in the place where I am now. She is my angel, my friend,

and my companion of life. I also thank my boys Hwanho and Seho who gave me a joy of life as a parent.

I also gratefully acknowledge that my study was partially supported by the British Scholarship Scheme 1998 jointly funded by the Foreign and Commonwealth Office, the British Council, the Department of Trade and Industry and the University of Edinburgh. The travel expenses of my visit to the NASSLI 2003 were partially supported by the Small Project Grant funded by Alumni Fund of the University of Edinburgh Development Trust.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Kihwang Lee*)

Table of Contents

List of Abbreviations

NOM	nominative case particle
ACC	accusative case particle
GEN	genitive case particle
LOC	locative case particle
DAT	dative case particle
INST	instrumental case particle
DIR	directional case particle
FUNC	function case particle
COM	comitative case particle
COMP	comparative case particle
QUOT	quotative case particle
VOC	vocative case particle
CONJ	conjunctive particle
TOP	topic auxiliary particle
PL	plural marker
HON	honorification marker
PST	past tense marker
FTR	future tense marker
DCL	declarative marker
INT	interrogative marker
COCON	coordinate conjunctive marker
SUBCON	subordinate conjunctive marker
AUXCON	auxiliary conjunctive marker
NML	nominaliser
ADV	adverbialiser
ADN	adnominaliser
COP	copular

Chapter 1

Introduction

This thesis deals with the use of statistical methods for case ambiguity resolution. More specifically, we propose statistical models that learn case assignment preference from a corpus and apply the models for the task of case ambiguity resolution in Korean. This chapter presents the motivation for the current work and briefly introduces the proposed approach. Finally, it gives an overview of the thesis.

1.1 Motivation

Korean is a case marked language in which case markers (case particles) are used to explicitly mark the type of relationships between nominals and their heads. Consider the following example.[1]

(1) a. Hwanho-*ga* Seho-*ege* uyu-*leul* ju-eoss-da.
Hwanho-NOM Seho-DAT milk-ACC give-PST-DCL
'Hwanho gave milk to Seho.'

b. Hwanho-*ga* uyu-*leul* Seho-*ege* ju-eoss-da.
Hwanho-NOM milk-ACC Seho-DAT give-PST-DCL

c. Seho-*ege* Hwanho-*ga* uyu-*leul* ju-eoss-da.
Seho-DAT Hwanho-NOM milk-ACC give-PST-DCL

d. Uyu-*leul* Seho-*ege* Hwanho-*ga* ju-eoss-da.
Milk-ACC Seho-DAT Hwanho-NOM give-PST-DCL

In (1), we can identify the case particles *-ga*, *-leul*, and *-ege*. These particles are attached to nominals and mark their cases NOMINATIVE, ACCUSATIVE, and DATIVE. Due to the ex-

[1] We follow the Korean Romanisation Standard officially suggested by the Korean Ministry of Culture and Tourism. It is shown in Appendix A.

plicit case marking, we don't have any difficulty in interpreting the scrambled sentences (1b)-(1d).[2] Although the word order SOV is recognised as canonical in Korean, there seems to be no difference in the acceptability of the various word orders in the above examples. In contrast to the nominals with accompanying case particles which can be found in (1), nominals lacking case particles are frequently observed in naturally occurring Korean texts and speeches. In an extreme situation, there can be no case marking at all in a sentence. Sentences (2a)-(2d) are such examples.

(2) a. Hwanho-*neun* Seho-*man* uyu-*do* ju-eoss-da.
Hwanho-TOP Seho-only milk-also give-PST-DCL
'As for Hwanho, he also gave milk only to Seho.'

b. Hwanho-*neun* uyu-*do* Seho-*man* ju-eoss-da.
Hwanho-TOP milk-also Seho-only give-PST-DCL

c. Seho-*man* Hwanho-*neun* uyu-*do* ju-eoss-da.
Seho-only Hwanho-TOP milk-also give-PST-DCL

d. ?Uyu-*do* Seho-*man* Hwanho-*neun* ju-eoss-da.
Milk-also Seho-only Hwanho-TOP give-PST-DCL

In sentences (2a) through (2d), case particles are missing for all the nouns. Instead auxiliary particles are used to add extra semantic/pragmatic contents to the sentences. However, these sentences, except (2d), are perfectly acceptable and the underlying cases are recognised as the same as those in (1a)-(1d).

The effect of missing case particles can be quite severe as illustrated in (3) and (4).

(3) a. Baem-eun hwangsogaeguli-do samki-nda.
Snake-TOP bullfrog-even swallow-DCL
'As for snakes, even bullfrogs swallow them.'
'As for snakes, they can even swallow bullfrogs.'

b. Baem-*i* hwangsogaeguli-*acc* samki-nda.
Snake-NOM bullfrog-ACC swallow-DCL
'Snakes swallow bullfrogs.'

c. Baem-*eul* hwangsogaeguli-*ga* samki-nda.
Snake-ACC bullfrog-NOM swallow-DCL
'Bullfrogs swallow snakes.'

(4) a. Asiana paeob-∅ jeongbu-∅ jeoggeug jungjae-∅ nas-eo
Asiana strike-∅ government-∅ actively mediation-∅ put forward-SUBCON
'Government actively puts forward to mediate the strike of Asiana.'

[2] These are only a subset of all possible word order variations.

b. Asiana paeob-*e* jeongbu-*ga* jeoggeug jungjae-*leul*
Asiana strike-LOC government-NOM actively mediation-ACC
nas-eo
put forward-SUBCON
'Government actively puts forward to mediate the strike of Asiana.'

(3a) is a perfectly acceptable sentence which can be encountered in everyday life. However, it is not trivial to interpret this sentence. That is, it is hard to determine 'who swallows who' in (3a). We know that the predicate *samki-* 'swallow' is a transitive verb which requires NOMINATIVE case-marked and ACCUSATIVE case-marked nominals as its arguments that serve as the SUBJECT and the DIRECT OBJECT of the sentence. If we recover the missing case particles in (3a), we can have two sentences, (3b) and (3c). In other words, (3a) has a case ambiguity which leads to two totally opposite interpretations of the sentence. The preferable interpretation would be (3c) in which nominals *baem* 'snake' and *hwangshogaeguli* 'bullfrog' are marked as ACCUSATIVE and NOMINATIVE respectively. The word order of this preferable interpretation is different from the canonical word order SOV. The reason we get this preferable interpretation is that we know the *hwangsogaeguli* 'bullfrog' is a frog that is big and strong enough to *samki-* 'swallow' even a *baem* 'snake'.

Sentences like (4a) which are frequently used as headings of news articles are similarly ambiguous. The average Korean adult speaker will be able to interpret this sentence as (4b). We cannot successfully interpret (4a) solely by linguistic knowledge without the help of the real world knowledge.

Although humans can successfully process sentences like (2)-(4) without much difficulty in most situations, for an automatic natural language processing system, coping with such sentences are not an easy task at all. For effective syntactic and semantic analyses, the case ambiguity problem must be dealt with.

Case-related phenomena including the case ambiguity problem briefly introduced above have been in the centre of Korean linguistics for a long time. Linguistic efforts tried to adapt the concept of case from inflected languages to the Korean language, which has a distinctive postpositional element called *josa* 'particle'. Some of them also attempted to describe the case assignment mechanism in Korean within established linguistic frameworks such as GB theory. Although it is true that the pure linguistic approaches provided valuable information and unveiled many secrets regarding case-related phenomena, their findings are still insufficient to deal with the diverse situations that can be observed in a naturally occurring text. This arises because most pure linguistic studies are based on small sets of data.[3]

[3]Nam (1993) and Nam (1997) are two prominent exceptions. Nam (1993) described the usages of two adverbial particles *-e* and *-eulo/-lo* based on the corpus evidence. Nam (1997) approached the identification and

In a sense, we cannot expect too much from the theoretical linguistic work since there is a big possibility that many issues related to case-related phenomena will be considered as extra-linguistic issues in pure linguistics. For example, information regarding word order preference or distribution of particles such as which particle is most frequently used with which predicate in which position is hard to find in theoretical linguistic work. It is, of course, still uncertain how this kind of information will help us to understand case-related phenomena and implement practical language processing systems. We strongly believe, however, that such information is beneficial to theoretical linguistic work as well as studies aiming at practical applications such as the current study.

We do not under-estimate the importance of theoretical work. We extensively use the relevant information provided by pure linguistic work but our primary focus is on automatic case ambiguity resolution. To achieve the goal of establishing statistical models for case ambiguity resolution and implement working system, we use a large-scale corpus of Korean to collect data that can train our models. We also perform multiple human annotation on our test set which we believe to be never tried before. We hope that our work will promote data intensive linguistic work on case-related issues in Korean.

There have been several efforts to attack the case ambiguity problem in Korean. The previous approaches are divided into two groups: knowledge-based approaches and statistical approaches.

Knowledge-based approaches need language resources such as subcategorisation dictionaries and thesauruses. However, large scale subcategorisation dictionaries and thesauruses in Korean suitable for real-world tasks are not available at the moment. Constructing these resources requires a huge amount of time and effort. The previous knowledge-based approaches all used experimental small scale language resources for their experiments and demonstrated the usefulness of the language resources.

In statistical approaches, natural language corpora were used for training the statistical models. The training material was collected from the unambiguous examples occurring in corpora typically using partial parsers. Some approaches tried to improve the performances of their models by incorporating experimental thesauruses and achieved high disambiguation accuracies over 86%.

The current work, which is an extension of the previous statistical approaches, is motivated by the following issues:

First, the case ambiguity resolution task should be defined precisely. In the previous approaches, the underlying case ambiguity problem in Korean was not fully explored. Target

---

classification of particles in a quantitative perspective using corpus data.

case particles should also be carefully selected reflecting the reality of case ambiguity. In most previous work, only a small set of case particles consisting of two or three case particles were used as target classes.

Second, as already mentioned, using unannotated material with partial parsing technique for training has been accepted as a standard procedure in statistical approaches to the case ambiguity resolution in Korean. This method is justified only because fully annotated material is not available. The adequacy of using a knowledge-lean data collection method has never been confirmed. The limitations of the shallow data collection method have also not been pointed out,

Third, any clue that could be useful and readily available in training data should be used as a feature for statistical models for the maximal use of the training data. Previous work used only a minimal set of features and recent efforts concentrated on utilising external resources instead of using more features in the training data.

Fourth, the statistical models should be easily extendable to incorporate more features. Previously proposed models were not explicitly probabilistic even though they used statistical information gathered from corpora. These models are also not easily expandable to incorporate more features.

Lastly, the evaluation of a case ambiguity resolution system should be performed on an independent test set of a reasonable size. Using an alternative evaluation measure other than the usual simple agreement measure should also be considered. Previous approaches evaluated their systems on relatively small test sets. Some test sets were constructed only for a limited number of predicates and even contained sentences from the training material.

1.2 Proposed Approach

The aim of this thesis is to tackle the case ambiguity problem in Korean with statistical methods while pursuing the following objectives.

First, through an examination of the relevant theoretical work, we precisely identify the causes for the case ambiguity problem in Korean. We also set the target case particles reflecting the reality of the problem.

Second, we clearly specify our choice of training data construction method. Our method does not depend on any high-level language processing tools other than a standard part-of-speech tagger and simple heuristic rules reflecting the structural characteristics of the Korean language. We also attempt to measure the effectiveness of the data collection method by applying the method to two treebanks of Korean consisting of 25,258 syntactically analysed sentences in total.

Third, we suggest two case decision methods for the task of case ambiguity resolution: discrete case decision method and sequential case decision method. In the discrete case decision method, each case ambiguity in a sentence is treated in isolation. For this method we use statistical classifiers based on simple joint probabilistic models that can be easily extended. We incorporate new features which have never been used before into these classifiers. In the sequential case decision method, every case decision in a sentence is treated in the context of a series of case decisions that take place in the sentence. This method is similar to other sequential category assignment tasks such as part-of-speech tagging. Thus we adopt the well-known Markov chain tagging model.

Finally, our statistical case ambiguity resolution models are evaluated by comparing the outputs of the system applied on a test set with the multiple human annotations on the test set. $Kappa$ is used to measure the pairwise agreements between the system outputs and human annotations. From the evaluation, the limitations of the unannotated training material and the shallow data collection method will be revealed. This will lead us to some of the possible future directions.

1.3 Overview of the Thesis

This thesis is organised as follows:

Chapter 2 surveys the theoretical background and related work. After clarifying the concept of case in general, we look into the usage of case particles and study the theoretical

work on case marking and assignment in Korean. Next, the case ambiguity problem in Korean is clearly identified and the conditions of the ambiguity are explored. Previous studies related to the current work are also presented.

Chapter 3 focuses on methodological issues concerning the training data collection method and statistical modelling for case ambiguity resolution in Korean. The corpora used for the data collection and evaluation are introduced and the proposed statistical models for our task are described. The data collection strategy for the current work and the evaluation method are also presented.

Chapter 4 describes the training data construction process and various experimental setups including the test set preparation and the performance bounds. The evaluation result for the knowledge-lean data collection method and the analysis of the human annotation results for the test set are also presented.

Chapter 5 contains the experimental results for our approach to the statistical case ambiguity resolution in Korean. Evaluation results for the discrete and the sequential case decision models are presented. We also discuss the roles of the features used in the statistical models and compare the two case decision models.

Finally, Chapter 6 concludes this thesis by summarising the results and the contribution of the thesis and suggesting the possible future directions.

Chapter 2

Background and Related Work

In this chapter, we survey the theoretical background and related work. After clarifying the concept of case in Section 2.1, we sketch the usage of case particles and study the theoretical work on case marking and assignment in Korean in Section 2.2. In Section 2.3, the case ambiguity problem is identified and the conditions of the ambiguity are explored. We turn to the related work in Section 2.4. Finally, Section 2.5 summarises this chapter.

2.1 Case

When words are put together to form a bigger linguistic unit, each word receives its own status and role in the unit being closely related to each other. It is usual that one particular word gets a special status of *head* while other words each get statuses of *dependents* among the words in a linguistic construction.[1]

Case is a system which marks the type of relationships that dependent nouns bear toward their heads. The head of a noun can be a preposition, postposition or another noun as well as a verb. Traditionally, case refers to inflectional marking system. However, it is also used to describe other marking systems such as postpositions (Blake, 1994). Typical examples of cases are NOMINATIVE, ACCUSATIVE, GENITIVE, DATIVE, LOCATIVE, and INSTRUMENTAL.

In some languages such as English, phonetically realised case markers do not exist. Instead, head-dependent relationships are realised by word order. For such languages, the notion of *abstract cases* can be applied. While introducing the abstract case, Chomsky (1981) distinguished *structural case* and *inherent case*. Structural cases are assigned to noun phrases according to their positions in structural configurations. For example, in

[1] A head is defined as "a constituent of an endocentric construction that, if standing alone, could perform the syntactic function of the whole construction." (Loos et al., 1997; Crystal, 2002)

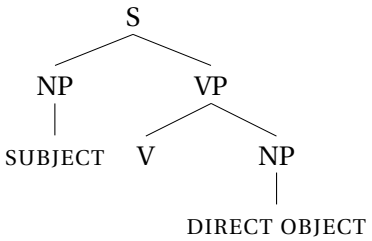


Figure 2.1: *A simplified phrase structure of an English sentence*

English, NOMINATIVE case is assigned to a noun phrase when it is in the subject position. Similarly, a noun phrase in the direct object position gets assigned an ACCUSATIVE case. Figure 2.1 is a simplified phrase structure of a transitive sentence in which SUBJECT and OBJECT positions are identified structurally.

Inherent case is mostly analogous to the traditional *oblique case*.[2] That is, an inherent case is assigned in the context of a lexical relationship of a dependent and a head rather than in a structural configuration. Inherent case assignment is often an idiosyncratic property of the assigning head. For instance, in English, prepositions assign inherent cases to noun phrases that are dependents of them and the actual cases are determined by the individual case-assigning prepositions.

Cases and *grammatical relations* should not be confused although they are closely related.[3] Grammatical relations are what cases express and refer to purely syntactic relations such as SUBJECT, DIRECT OBJECT and INDIRECT OBJECT (Blake, 1994; Woolford, 1999). It is not necessary that grammatical relations have one-to-one correspondence with cases. In Korean, NOMINATIVE and ACCUSATIVE cases are mostly associated with SUBJECT and DIRECT OBJECT respectively. In other languages, however, other pairings of cases and grammatical relations are also observed.

Semantic roles are also distinct from cases and grammatical relations (Higginbotham, 1999).[4] Semantic roles are semantic relations between a head and dependent nouns and refer to relations such as AGENT, PATIENT, and THEME. Again, there are not fixed mappings between cases and semantic roles. Nevertheless, in languages with rich case systems, the cases will give some information about the semantic roles. For instance, in Korean, AGENT role is mostly associated with NOMINATIVE case but not with ACCUSATIVE case.

As mentioned in Chapter 1, case provides vital clues for effective analyses of syntactic structure and semantic content of a sentence in Korean and other languages such as Japanese

[2] In ancient Greek, non-nominative cases are collectively classified as oblique cases (Blake, 1994).

[3] Grammatical relation is sometimes called *grammatical role* or *grammatical function*.

[4] Semantic roles are also called *thematic roles* or θ*-roles*.

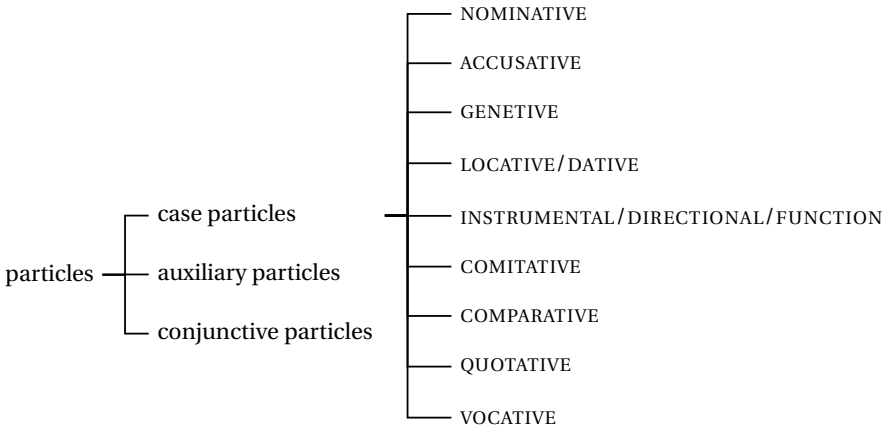


Figure 2.2: *Classification of case particles in Korean*

with rich case marking systems.

2.2 Case Marking in Korean

This section presents an inventory of Korean case markers (case particles) and their usages and briefly surveys some of the theoretical work on case marking and assignment mechanisms in Korean.

2.2.1 Case Particles

Korean is typologically classified as an *agglutinative* language. A typical characteristic of Korean as an agglutinative language is the conjugation of predicates such as verbs, adjectives, and the copula. The stems of Korean predicates cannot be used independently and require endings to function in sentences.

Another distinctive feature of Korean is the existence of postpositional elements called *particles*. There are three types of particles in Korean: *case particles*, *auxiliary particles*, and *conjunctive particles*. Case particles are attached to noun phrases and mark their cases. Auxiliary and conjunctive particles are not related to case marking. Auxiliary particles add semantic/pragmatic meanings such as emphasis and focus. Conjunctive particles conjoin multiple noun phrases.

Figure 2.2 illustrates the classification of the Korean case particles.[5]

[5]The description of case particles in this section is based on Nam and Koh (1993), Sohn (1999), and Lee and Ramsey (2000).

2.2.1.1 Nominative Case Particle

(5a) is a sentence showing a NOMINATIVE case marking by the particle *-i/-ga*.[6] Particles *-kkeseo* and *-eseo* in (5b) and (5c) are two other NOMINATIVE case particles. The particle *-kkeseo* can be used when the preceding noun is an esteemed and honoured person. The particle *-eseo* is used with an impersonal and collective noun.[7]

(5) a. Bi-*ga* naeli-nda.
rain-NOM fall-DCL
'It rains.'

b. Seonsaeng-nim-*kkeseo* o-si-eoss-da.
teacher-HON-NOM come-HON-PST-DCL
'The teacher came.'

c. Gyohoe-*eseo* guhopum-eul bunjaeng jiyeog-e bonae-eoss-da.
church-NOM relief supplies-ACC troubled areas-LOC send-PST-DCL
'Church sent the relief supplies to the troubled areas.'

A noun phrase marked as NOMINATIVE case usually functions as the SUBJECT of a sentence. It can also function as the OBJECT of a transitive adjective, the complement of the negation copula *ani-* 'not be' and the verb *doe-* 'become' as depicted in (6).

(6) a. Hwanho-neun Seho-*ga* joh-a?
Hwanho-TOP Seho-NOM like-INT?
'Hwanho, do you like Seho?' (OBJECT)

b. Seho-neun malsseongjaengi-*ga* ani-da.
Seho-TOP trouble maker-NOM not be-DCL
'Seho is not a trouble maker.' (COMPLEMENT)

c. Hwanho-ga chodeunghagsaeng-*i* doe-eoss-da.
Hwanho-NOM primary school student-NOM become-PST-DCL
'Hwanho became a primary school boy.' (COMPLEMENT)

Several studies suggested that the NOMINATIVE case particle *-i/-ga* has a modal semantic content such as 'exclusive reference' (Nam, 1972), 'exclusive opposition' (Im, 1972), and 'specific predication' and 'selective specification' (Shin, 1975).

[6] Particles *-i* and *-ga* are phonologically conditioned variants. Particle *-i* is used after a consonant while *-ga* is used after a vowel. Other particles are shown in the same manner.

[7] Regarding particle *-eseo* as a NOMINATIVE case particle can be a controversial issue in Korean linguistics since *-eseo* is typically used as a LOCATIVE case particle (See Section 2.2.1.4). The standard Korean grammar considers *-eseo* as a NOMINATIVE case particle from the fact that *-eseo* is perfectly interchangeable with *-i/-ga* in sentences like (5c) while preserving the meaning of the sentence.

2.2.1.2 Accusative Case Particle

The case particle which marks ACCUSATIVE case is *-eul/-leul*. An ACCUSATIVE case marked noun phrase functions not only as the DIRECT OBJECT of a transitive verb but also as the purpose of an action, and the duration or distance of an action as shown in the following examples.

(7) a. Seho-ga uyu-*leul* manhi masi-eoss-da.
Seho-NOM milk-ACC much drink-PST-DCL
'Seho drank much milk.' (DIRECT OBJECT)

b. Hwanho-ga oneul sopung-*eul* ga-nda.
Hwanho-NOM today picnic-ACC go-DCL
'Hwanho goes on a picnic today.' (purpose of an action)

c. Seho-ga han sigan-*eul* geol-eoss-da.
Seho-NOM one hour-ACC walk-PST-DCL
'Seho walked for one hour.' (duration of an action)

There were some claims that the ACCUSATIVE case particle *-eul/-leul* has a semantic content like the NOMINATIVE case particle *-i/-ga*. The meaning of the particle *-eul/-leul* suggested in Im (1979) and confirmed by Hong (1986) and Chung (1988) is 'wholeness'.

2.2.1.3 Genitive Case Particle

The case particle *-ui* marks the GENITIVE case. This particle links two noun phrases. The possible semantic relationships between the noun phrases linked by the GENITIVE particle are extremely diverse and impossible to give a simple definition. Some representative usages of the GENITIVE case particle and their semantic interpretations are given in (8).

(8) a. Jeo chaegsang-i neo-*ui* chaegsang-i-da.
that desk-NOM you-GEN desk-COP-DCL
'That table is your table.' (possession)

b. Cameron-eun Hwanho-*ui* chingu-i-da.
Cameron-TOP Hwanho-GEN friend-COP-DCL
'Cameron is Hwanho's friend.' (relationship)

c. Gim seonsaeng-*ui* jean-i badadeulyeoji-eoss-da.
Kim teacher-GEN suggestion-NOM be accepted-PST-DCL
'Mr Kim's suggestion was accepted.' (creator, originator)

2.2.1.4 Locative and Dative Case Particles

The particles *-e*, *-ege*, *-kke*, and *-eseo* are LOCATIVE case particles. These particles express a wide variety of meanings. The meanings are determined by the contexts. (9a)-(9d) are typical examples of the uses of the LOCATIVE case particles.

(9) a. Jib-*e* jangnangam gicha-ga manh-da.
Home-LOC toy train-NOM many-DCL
'There are many train toys at home.' (static location)

b. Hwanho-ga bagmulgwan-*e* ga-ass-da.
Hwanho-NOM museum-LOC go-PST-DCL
'Hwanho went to a museum.' (destination)

c. Beoseu-ga yeol si-*e* tteona-nda.
Bus-NOM ten hour-LOC leave-DCL
'The bus leaves at ten o'clock.' (point of time)

d. Seho-ga bulkkochnoli soli-*e* jam-eul kkae-eoss-da.
Seho-NOM fireworks sound-LOC sleep-ACC wake up-PST-DCL
'Seho was waken up by the sound of fireworks.' (cause)

The particles *-e*, *-ege* and *-kke* are often treated as DATIVE case markers. These particles are only used with animate nouns while *-e* is used with inanimates. Particle *-kke* is an honorific form.

(10) a. Hwanho-ga hwabun-*e* mul-eul ju-eoss-da.
Hwanho-NOM flower pot-LOC water-ACC give-PST-DCL
'Hwanho gave water to the flower pot.'

b. Hwanho-ga Seho-*ege* mul-eul ju-eoss-da.
Hwanho-NOM Seho-LOC water-ACC give-PST-DCL
'Hwanho gave water to Seho.'

c. Hwanho-ga halabeoji-*kke* mul-eul deuli-eoss-da.
Hwanho-NOM grandfather-LOC water-ACC give-PST-DCL
'Hwanho gave water to grandfather.'

The Particles *-eseo* and *-egeseo* belong to another group of LOCATIVE case particles. These particles are used to indicate a source or an origination of an activity and a dynamic location, i.e, a location of an activity.

(11) a. Halmeoni-kkeseo hangug-*eseo* o-si-eoss-da.
grandmother-NOM Korea-LOC come-HON-PST-DCL
'Grandmother came from Korea.' (source)

b. Seho-ga chimdae-*eseo* ttwi-nda.
Seho-NOM bed-LOC jump-DCL
'Seho is jumping on the bed.' (dynamic location)

2.2.1.5 Instrumental, Directional and Function Case Particles

Case particle *-eulo/-lo* marks INSTRUMENTAL, DIRECTIONAL, and FUNCTION cases. (12) are typical examples of the usage of *-eulo/-lo*.

(12) a. Seho-ga gawi-*lo* jongi-leul jaleu-ass-da.
Seho-NOM scissors-INST paper-ACC cut out-PST-DCL
'Seho cut out the paper with scissors.' (INSTRUMENTAL)

b. Halabeoji-kkeseo jihasil-*lo* naelyeoga-si-eoss-da.
Grandfather-NOM basement-DIR go down-HON-PST-DCL
'Grandfather went down to the basement.' (DIRECTIONAL)

c. Samchon-i haggyo wiwonhoe wiwon-*eulo* bongsaha-nda.
Uncle-NOM school board member-FUNC serve-DCL
'Uncle serves as a member of the school board.'

The INSTRUMENTAL case is highly polysemous. Examples (13a)-(13e) show the usages of *-eulo/-lo* with senses of 'means', 'material', 'constituency', 'cause/reason', and 'manner'.

(13) a. Hwanho-ga beoseu-*lo* jib-e o-ass-da.
Hwanho-NOM bus-INST home-LOC come-PST-DCL
'Hwanho came home by bus.' (means)

b. Seho-ga chalheulg-*eulo* jeobsi-eul mandeul-eoss-da.
Seho-NOM clay-INST plate-ACC make-PST-DCL.
'Seho made a plate with clay.' (material)

c. Keompyuteo siseutem-eun hadeuweeo-wa sopeuteuweeo-*lo*
Computer system-TOP hardware-CONJ software-INST
guseongdoe-nda.
consist of-DCL
'A computer system consists of hardware and software.' (constituency)

d. Jeungjo halabeoji-kkeso am-*eulo* dolaga-si-eoss-da.
Great grandfather-NOM cancer-INST die-HON-PST-DCL
'Great grandfather died of a cancer.' (cause/reason)

e. Kim seonsaeng-eun byeongwon-eso maeil yeolsim-eulo ilha-nda.
Kim teacher-TOP hospital-LOC everyday enthusiasm-INST work-DCL
'Mr Kim enthusiastically works at the hospital everyday.' (manner)

2.2.1.6 Comitative Case Particle

The particle used to mark COMITATIVE case is *-gwa/-wa*. This particle is typically used with reciprocal verbs such as *gyeolhonha-* 'marry', *dalm-* 'resemble', *manna-* 'meet', and *ssau* 'fight'.

(14) a. Hwanho-ga Seho-*wa* dalm-ass-da.
Hwanho-NOM Seho-COM resemble-PST-DCL
'Hwanho and Seho resemble each other.'

b. Abeoji-kkeso seonsaeng-nim-*gwa* manna-si-eoss-da.
Father-NOM teacher-HON-COM meet-HON-PST-DCL
'Father met the teacher.'

The particle *-gwa/-wa* can also be used to connect two noun phrases. This connective use should be distinguished from the COMITATIVE case marking. Consider the following examples.

(15) a. Hwanho-ga George-*wa* datu-eoss-da.
Hwanho-NOM George-COM quarrel-PST-DCL
'Hwanho quarrelled with George.'

b. Hwanho-*wa* George-ga datu-eoss-da.
Hwanho-CONJ George-NOM quarrel-PST-DCL
'Hwanho and George quarrelled with each other.'

(16) a. Hwanho-ga Seho-*wa* bidio-leul bo-nda.
Hwanho-NOM Seho-COM video-ACC watch-DCL.
'Hwanho is watching a video with Seho.'

b. Hwanho-*wa* Seho-ga bidio-leul bo-nda.
Hwanho-CONJ Seho-NOM video-ACC watch-DCL
'Hwanho and Seho are watching a video.'

(15a) and (16a) are instances of *-gwa/-wa* being used as COMITATIVE case particles, and (15b) and (16b) are instances of *-gwa/-wa* being used as CONJUNCTIVE particles. In (15a) and (15b), where a reciprocal verb *datu-* 'quarrel' is used, the different interpretations for two usages of *-wa* are clearly recognised. In (16a) and (16b), where the verb *bo-* 'watch' is not a reciprocal verb, the difference between the semantic contents of the two sentences is not evident. In (16a), *Hwanho* 'Hwanho' is watching a video with *Seho* 'Seho' intentionally or necessarily. On the other hand, in (16b), *Hwanho* 'Hwanho' and *Seho* 'Seho' are just watching a video together. It does not need to be necessary or intentional.[8]

[8] If there are pauses between *Hwanho-wa* and *George-ga*, and *Hwanho-wa* and *Seho-ga -wa* can be recognised as a COMITATIVE case particle.

In an informal situation, *-hago* or *-lang*/*-ilang* can be used instead of *-gwa*/*-wa*.

(17) a. Hwanho-ga eomma-*hago* gongwon-e ga-ass-da.
Hwanho-NOM Mum-COM park-LOC go-PST-DCL
'Hwanho went to the park with mum.'

b. Seho-ga hyeong-*ilang* nolae-leul buleu-nda.
Seho-NOM brother-COM song-ACC sing-DCL
'Seho sings a song with his brother.'

2.2.1.7 Comparative Case Particles

There are no comparative or superlative affixes in Korean. Comparison is expressed by COMPARATIVE case particles *-boda* '(rather) than, (more/less) than', *-mankeum* 'as much/many as, equal to', *-cheoleom* 'like, the same as', and *gathi* 'like, the same as'.

(18) a. Seho-ga Hwanho-*boda* iljjig ileona-ass-da.
Seho-NOM Hwanho-than early wake up-PST-DCL
'Seho woke up earlier than Hwanho.'

b. Seho-ga Hwanho-*mankeum* sagwa-eul meog-eoss-da.
Seho-NOM Hwanho-as many as apple-ACC eat-PST-DCL
'Seho ate as many apples as Hwanho.'

c. Hwanho-ga eoleun-*cheoleom* mal-eul ha-nda.
Hwanho-NOM adult-like speech-ACC do-DCL
'Hwanho speaks like an adult.'

d. Seho-ga aegi-*gathi* gu-nda.
Seho-NOM baby-like behave-DCL
'Seho behaves like a baby.'

2.2.1.8 Quotative Case Particles

Embedded quotative clauses are recognised by the QUOTATIVE case particles *-lago* and *-go*. The former is used for a direct quotation and the latter is used for an indirect quotation as in (19).

(19) a. Halabeoji-kkeseo "Nalssi-ga cham joh-da."-lago
Grandfather-NOM "The weather-NOM very good-DCL."-QUOT
malsseumha-si-eoss-da.
speak-HON-PST-DCL
'Grandfather said "The weather is very good'.

b. Halabeoji-kkeseo nalssi-ga cham joh-da-go
Grandfather-NOM weather-NOM very good-DCL-QUOT
malsseumha-si-eoss-da.
speak-HON-PST-DCL
'Grandfather said that the weather was very good.'

2.2.1.9 Vocative Case Particle

The VOCATIVE case particle *-a/-ya* is attached to a personal name to express that the person is being called typically in informal speech. The particles *-yeo/-iyeo* is a variant which is used only in restricted domains such as poetry and the Bible.

(20) a. Hwanho-*ya*, ije ja-l sigan-i-da.
Hwanho-VOC, now sleep-ADN time-COP-DCL
'Hwanho, it is time to go to bed.'

b. Nim-*ieyo*, dangsin-eun baegbeon-ina danlyeonha-n
My-love-VOC, you-TOP hundred-times-as many as temper-ADN
geumgyeol-i-bnida.
gold-COP-DCL
'My love, you are a piece of gold purified as many as hundred times.'

2.2.2 Theories of Case Marking and Assignment in Korean

2.2.2.1 Traditional Approaches

In traditional Korean grammars, case was defined as 'the status of a word in a sentence as a constituent of the sentence' (Choi, 1937/1983) or 'the status (function) which a noun phrase, that is led by a verb takes in a sentence as a constituent of the sentence' (Heo, 1983). In short, case was understood as the function of a noun phrase as a constituent of a sentence. Accordingly a case particle was defined as 'particle which grants a function as a sentential constituent to a noun phrase.' There were no separately established case assignment mechanisms in traditional descriptive grammar frameworks.

The standard school grammar (Nam and Koh, 1993) extended the traditional grammars and incorporated a number of new concepts from modern syntactic theories. A noteworthy newly introduced concept related to case is the *jalisu* 'arity'.[9] *Jalisu* 'arity' is an idiosyncratic property of a predicate that specifies the number of its arguments and their cases.

[9] *Jalisu* 'arity' is similar to *valency* and *subcategorisation frame*. Valency refers to the capacity of a verb to take a specific number and type of arguments (Loos et al., 1997).

Thus, the role of a case particle is to mark the case of a noun specified in *jalisu* 'arity' of a predicate. Some predicates have more than one *jalisu* 'arity' as shown in (21) and (22).

(21) a. Bakwi-ga jal do-nda.
Wheel-NOM well turn-DCL
'The wheel turns well.'

b. Dal-i jigu dule-leul do-nda.
Moon-NOM earth around-ACC go around-DCL
'The moon goes around the earth.'

(22) Cha-ga meomchu-eoss-da.
Car-NOM stop-PST-DCL
'The car stopped.'

a. Unjeonsa-ga cha-leul meomchu-eoss-da.
Driver-NOM car-ACC stop-PST-DCL
'The driver stopped the car.'

2.2.2.2 Case Grammar

Since Korean has a rich case marking system, Case Grammar (Fillmore, 1968, 1969) was rigorously applied to the description of Korean from the early stage (e.g., Park 1970; Yang 1972; Kim 1973; Sung 1974). These works all adopted the following rewrite rules for Korean following the standard work of the Case Grammar.

(23) a. $S \rightarrow P + M$

b. $P \rightarrow C_1 \ldots C_n + V$

c. $C \rightarrow NP + K$

where M = Model, P = Proposition, C = Case, K = Case Marker

The rewrite rules in (23) specifies a semantic structure of a sentence rather than a syntactic structure. Thus, the cases (C) are *semantic cases* distinguished from the *surface cases*. Sung (1974) identified 10 cases and their markers as shown in Table 2.1.

In a Case Grammar approach, the particles *-i/-ga* NOMINATIVE and *-eul/-leul* ACCUSATIVE are treated as SUBJECT and OBJECT markers. These markers are introduced to the surface structure by transformations. Consider the following example.

(24) a. $[[[\text{Seho-ege}]_{\text{Loc}}\ [\text{moja-}\varnothing]_{\text{Obj}}\ [\text{iss}]_{\text{V}}]_{\text{Prop}}\ [\text{da}]_{\text{Mod}}]_{\text{Sentence}}$

b. $[[\text{Moja-}\varnothing]_{\text{Obj}}\ [[\text{Seho-ege}]_{\text{Loc}}\ [\text{iss}]_{\text{V}}]_{\text{Prop}}\ [\text{da}]_{\text{Mod}}]_{\text{Sentence}}$

c. [[Moja-∅-SM] [Seho-ege] [iss] [da]]

case	case marker
AGENT	*-ege*
DATIVE	*-ege*
INSTRUMENTAL	*-eulo/-lo*
OBJECT	∅
COMITATIVE	*-gwa/-wa*
SOURCE	*-eseo*
GOAL	*-e, -eulo/-lo*
LOCATIVE	*-e, -eseo*
TIME	*-e*
PATH	*-eulo/-lo*

Table 2.1: *Semantic cases and their markers in Korean*

d. Moja-ga Seho-ege iss-da.
Hat-NOM Seho-DAT exist-DCL
'Seho has a hat.'

If we apply the subjectivisation transformation to the OBJECT case in (24a), the OBJECT case escapes from the Proposition and attaches itself to the Sentence directly (24b). Then the subject marker (SM) is adjoined to the OBJECT case. Finally, the OBJECT case marker is deleted to form the surface sentence (24d).

2.2.2.3 Government and Binding Theory

Following Chomsky (1981, 1986), in which case assignment procedure is explained in the context of syntactic configuration of *government*, a number of case assignment mechanisms were proposed (e.g., Kang 1986; Im 1987; Kim 1990, 1994; Yoo 1995). These studies all treat the NOMINATIVE case particle *-i/-ga* and the ACCUSATIVE particle *-eul/-leul* as structural case markers that do not have any lexical meaning. These structural case markers are distinguished from the inherent case markers like *-e, -eulo/-lo*, and *-gwa/-wa* that have concrete lexical meanings. (25) is the 'Case Assignment Principle in Korean' proposed in Kim (1994).

(25) The Case Assignment Principle in Korean
   a. Government of tense element of INFL: NOMINATIVE *-i/-ga*
   b. Government of verb
      i. [+state] verb: NOMINATIVE *-i/-ga*
      ii. [-state] verb: ACCUSATIVE *-eul/-leul*

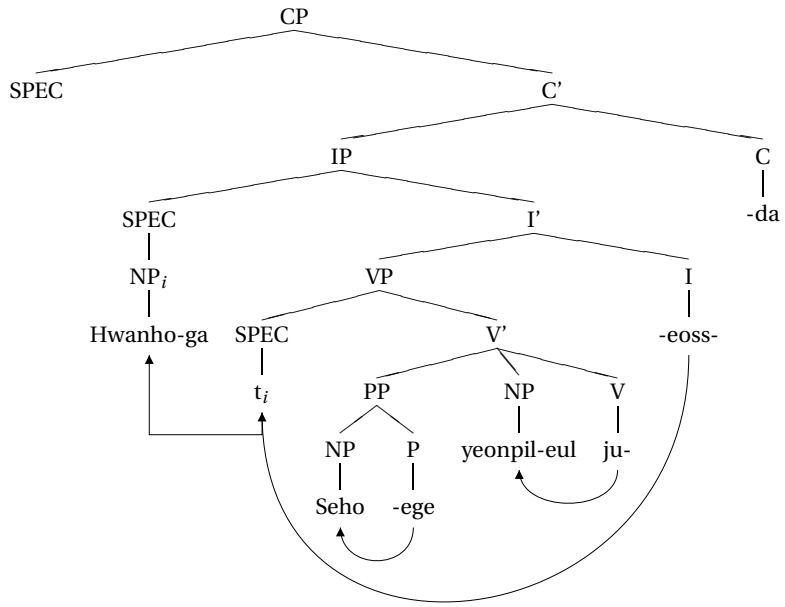


Figure 2.3: *A GB-style phrase structure tree showing the case assignment mechanism in Korean*

c. Contextual case ([NP_X]): GENITIVE *-ui*

d. Case assignment and realisation are concurrent and completed after movement from D-structure to S-structure before scrambling.

e. If structural conditions are satisfied, case can also be assigned to optional constituents.

f. Case particles *-i/-ga* and *-eul/-leul* are morphological realisations of structurally determined abstract cases. Other case particles have concrete meanings.

g. The feature of a governor percolates into its maximal projection.

The phrase structure and the case assignment procedure for a sentence (26) conforming to the Case Assignment Principle are illustrated in Figure 2.3.

(26) Hwanho-ga Seho-ege yeonpil-eul ju-eosss-da.
Hwanho-NOM Seho-DAT pencil-ACC give-PST-DCL
'Hwanho gave a pencil to Seho.'

In Figure 2.3, noun phrase *Hwanho* 'Hwanho' is assigned a NOMINATIVE case by a non-terminal ending *-eoss-* PST which governs the noun phrase. This noun phrase is moved to its final position after the case assignment. Similarly, the verb *ju-* 'give' assigns an ACCUSATIVE case to *yeonpil* 'pencil'. The assigned cases are morphologically realised by the case particles *-ga* and *-eul*. In contrast to the structural case assignments, a noun phrase

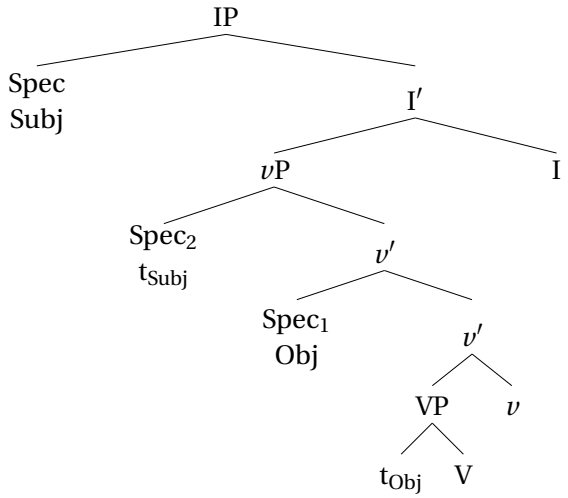


Figure 2.4: *A Minimalist Program-style phrase structure tree for a Korean transitive sentence*

Seho gets assigned DATIVE case by a *postposition -ege* which has a concrete lexical meaning.[10]

There are also attempts to explain case-related phenomena in Korean (e.g. Yu 1995; Kang 1996; Kim 1999a) based on the Minimalist Program (Chomsky, 1993, 1995). In the Minimalist Program framework, case assignment is replaced by the *case checking* operation. However, the fundamental idea on structural/inherent case marking is preserved. Figure 2.4 is a Minimalist Program-style phrase structure analysis of a transitive sentence in Korean given in Kim (1999a).

2.2.2.4 Head-Driven Phrase Structure Grammar

Head-Driven Phrase Structure Grammar (Pollard and Sag, 1988, 1994) is a highly lexicalised grammar formalism. In the original HPSG, there is no explicit case assignment operation and case assignment is treated as a matter of lexical selection. Case is realised as one of the many properties of a dependent which are governed by a head. The various relationships between a lexical head and its complements are encoded in the feature SUBCAT. The flow of subcategorisation information is handled by the 'Subcategorisation Principle' shown in (27).[11]

[10] There are variations on the treatment of oblique case marking. Kang (1988) considers oblique cases as structural cases. On the other hand, Kim (1999b) distinguishes two different usages of oblique case and treats them differently: If a noun phrase marked as an oblique case is used as an argument, the oblique case is assigned by the governing verb and the case marker is just marking the case. If the noun phrase is used as a non-argument, the oblique case marker also assigns the case.

[11] The Subcategorisation Principle has been replaced by the Valance Principle in Pollard and Sag (1994).

(27) Subcategorisation Principle
In a headed phrase, the list value of DAUGHTERS | HEAD-DAUGHTER | SYNSEM | LOCAL | CATEGORY | SUBCAT is the concatenation of the list value of SYNSEM | LOCAL | CATEGORY | SUBCAT with the list consisting of the SYNSEM value in order of the elements of the list value of DAUGHTERS | COMPLEMENT-DAUGHTERS.

Chang (1993) presents a fairly comprehensive syntactic/semantic analysis of Korean within the HPSG framework.[12] This study does not approve the notion of case for Korean. Instead, grammatical function is treated as a primitive grammatical element of Korean. According to this study, the case particles are marking grammatical functions not cases.[13] Case particles are classified into two groups. The first group consists of NOMINATIVE, ACCUSATIVE, and QUOTATIVE particles. These particles function as markers and form head-marker structures with headwords. The second group of case particles are equivalent to the oblique case particles such as *-e*, *-eulo/-lo*, and *-gwa/-wa*. These particles function as heads and form particle phrases with their complements. Figure 2.5 and Figure 2.6 show feature structures for a NOMINATIVE noun phrase *Hwanho-ga* 'Hwanho-NOM' and a DATIVE particle phrase *Seho-ege* 'Seho-DAT'. In Figure 2.5, we can see that the feature GF (grammatical function) is introduced. Possible values for the feature are SUBJECT and OBJECT. Once noun phrases and particle phrases are formed, they can be combined with a verb which has a concordant SUBCAT feature, for instance, *ju-* 'give' in Figure 2.7.

When the arguments are combined with a head, the order of combination is determined by the *obliqueness hierarchies*. (28) is the 'Obliqueness Hierarchy of Grammatical Functions in Korean' proposed in Chang (1993).

(28) Obliqueness Hierarchy of Grammatical Functions in Korean
SUBJECT < SUBJECT-2 < OBJECT < OBJECT-2 < LOCATIVE OBJECT < other oblique objects

Figure 2.8 is a feature structure of the sentence (26), which is repeated here as (29).

(29) Hwanho-ga Seho-ege yeonpil-eul ju-eoss-da.
Hwanho-NOM Seho-DAT yeonpil-ACC give-PST-DCL
'Hwanho gave a pencil to Seho.'

Unlike Chang (1993), Yoo (1993) incorporated the case assignment operation in the notion of the structural case into the HPSG-based analysis of Korean (cf. Pollard 1994; Heinz and

[12] Chang (1993) is largely based on Pollard and Sag (1988) and partially incorporates the revised version of HPSG in Pollard and Sag (1994).

[13] Strictly speaking, we cannot use the term 'case particle' for this work. However, we will use the term for the convenience.

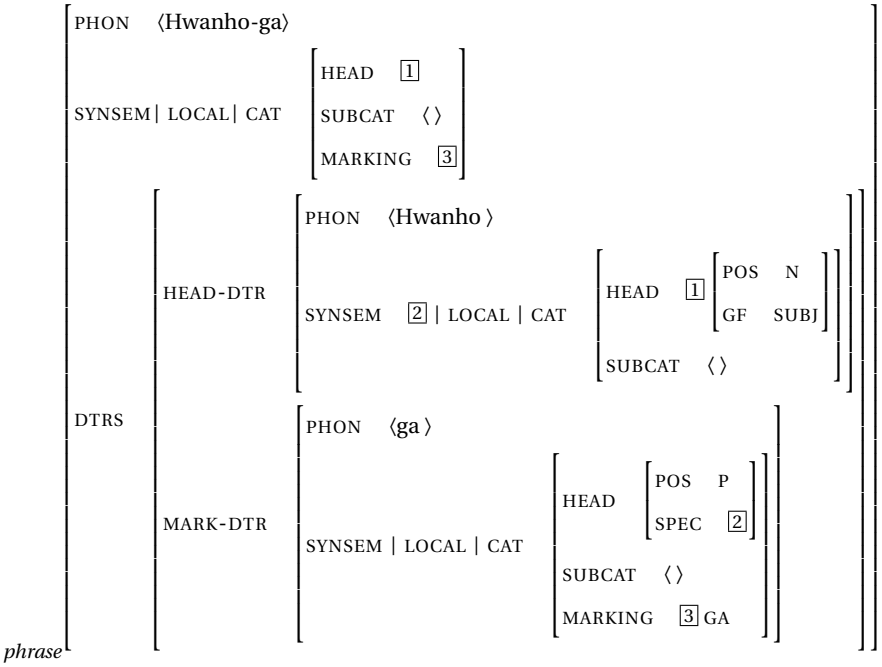


Figure 2.5: *The feature structure of the noun phrase* Hwanho-ga *'Hwanho-*NOM*'*

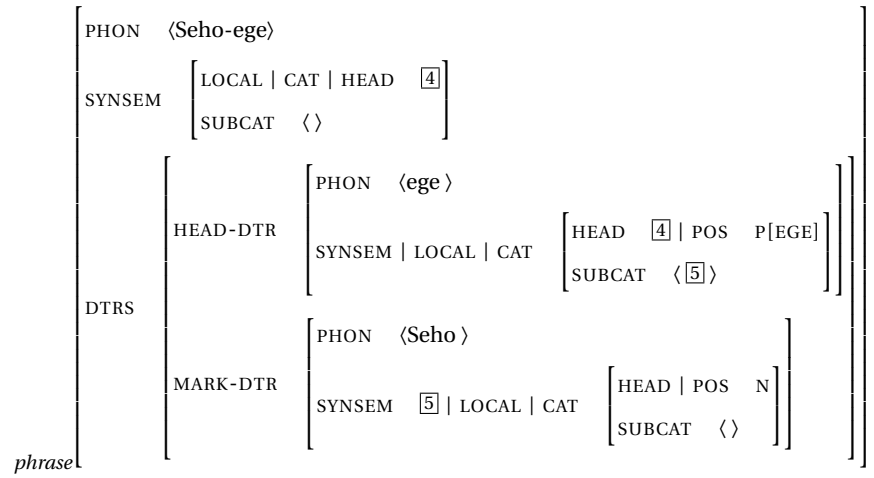


Figure 2.6: *The feature structure of the particle phrase* Seho-ege *'Seho-*DAT*'*

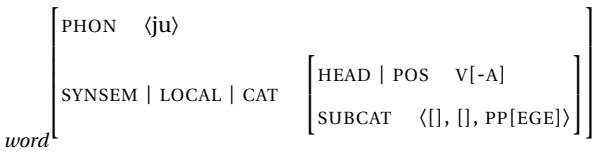


Figure 2.7: *The feature structure of the verb* ju- *'give'*

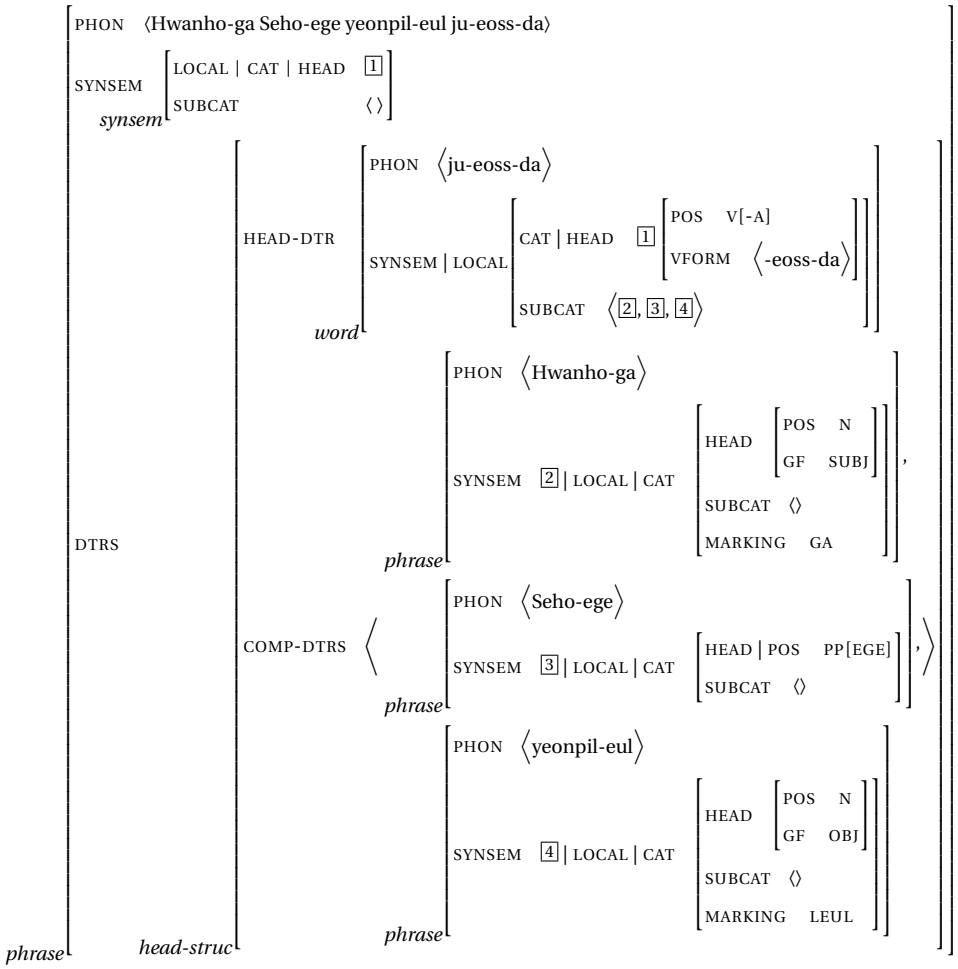


Figure 2.8: *The feature structure of the sentence (29)*

Matiasek 1994). This study also adopted the distinction of the structural case and the inherent case. (30) is the Case Principle provided in Yoo (1993) for structural case realization.

(30) Case Principle

An unresolved structural NP, which is a daughter of a phrase α, is [nom] if it is a SUBJ-DTR of α and [acc] if it is a COMP-DTR of α.

This framework is similar to the structural case assignment in GB theory, in which the structural case assignment is purely based on syntactic configuration. However, within HPSG, structural case are still lexically assigned in the lexical entry of a predicate even though it requires some syntactic information specified in the Case Principle.

The syntactic combination of a noun phrase and a case marker can be handled by the HEAD-MARK schema (Pollard and Sag, 1994) as in Chang (1993) or a similar schema. For example, Lee (2004) introduced the HEAD-C(ASE)MARK schema, which is illustrated Figure 2.9.

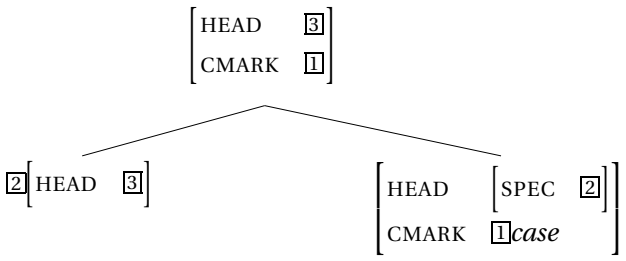


Figure 2.9: *The HEAD-CMARK schema*

2.3 Case Ambiguity in Korean

In this section, we look into the two phenomena that cause the case ambiguity in Korean: case particle deletion and case particle unrealisation. We also cautiously explore the conditions of the case particle deletion and unrealisation.[14]

2.3.1 Case Particle Deletion

As presented in the previous sections, the cases for noun phrases are marked by case particles in Korean. There are, however, many instances in which the case particles are deleted when they are followed by the auxiliary particles. Consider the following examples.

(31) a. Bi-*ga* naeli-nda.
rain-NOM fall-DCL
'It rains.'

b. Bi-∅-*neun* naeli-nda.
rain-TOP fall-DCL
'(lit.) As for the rain, it falls.'

c. Bi-∅-*do* naeli-nda.
rain-also fall-DCL
'(lit.) We are also having a rain (and other features like a strong wind).'

d. Bi-∅-*man* naeli-nda.
rain-only fall-DCL
'(lit.) We are only having a rain (and not other features like a strong wind).'

In (31b)-(31d), the NOMINATIVE case particle *-ga* is missing. Instead, the auxiliary particles *-neun* TOPIC, *-do* 'also', and *-man* 'only' are used without the case particles. These auxiliary particles are not related to case marking. They only add semantic/pragmatic meanings such as emphasis and focus to the sentence. Therefore, the same set of auxiliary particles

[14] This section is based on Hong (1987), Kim (1998), Chung (1998), and Choi (1999).

can be used in other places. In (32), auxiliary particles are used in DIRECT OBJECT positions without the ACCUSATIVE case particle.

(32) a. Seho-ga uyu-*leul* manhi masi-eoss-da.
Seho-NOM milk-ACC much drink-PST-DCL
'Seho had plenty of milk.'

b. Seho-ga uyu-*neun* manhi masi-eoss-da.
Seho-NOM milk-TOP much drink-PST-DCL
'(lit.) As for the milk, Seho had plenty of it.'

c. Seho-ga uyu-*do* manhi masi-eoss-da.
Seho-NOM milk-also much drink-PST-DCL
'Seho also had plenty of milk.'

d. Seho-ga uyu-*man* manhi masi-eoss-da.
Seho-NOM milk-only much drink-PST-DCL
'Seho only had plenty of milk.'

From (31) and (32), we can reason that if the NOMINATIVE or the ACCUSATIVE case particle is followed by an auxiliary particle, they are deleted. In (33), we confirm that this deletion is obligatory.

(33) *Bi-*ga*-{neun, do, man} naeli-nda.
rain-NOM-{TOP, also, only} fall-DCL
'It rains.'

a. *Seho-ga uyu-*leul*-{eun, do, man} manhi masi-eoss-da.
Seho-NOM milk-acc-{TOP, also, only} much drink-PST-DCL
'Seho had a plenty of milk.'

Not all case particles are deleted when they are followed by auxiliary particles. Two other NOMINATIVE case particles *-kkeseo* and *-eseo* can co-occur with auxiliary particles as shown in (34)-(35).[15] For these particles, case particle deletion is an optional process.

(34) a. Seonsaeng-nim-*kkeseo*-{neun, do, man} o-si-eoss-da.
Teacher-HON-NOM-{TOP, also, only} come-HON-PST-DCL
'The teacher came.'

b. Gyohoe-*eseo*-{neun, do, man} guhopum-eul bunjaeng jiyeog-e
Church-NOM-{TOP, also, only} relief supplies-ACC troubled areas-LOC
bonae-eoss-da.
send-PST-DCL
'Church sent the relief supplies to the troubled areas.'

[15] It is also possible to understand that particle *-i/-ga* is deleted in (35b).

(35) a. Seonsaeng-nim-∅-{eun, do, man} o-si-eoss-da.
Teacher-HON-∅-{TOP, also, only} come-HON-PST-DCL
'The teacher came.'

b. Gyohoe-∅-{neun, do, man} guhopum-eul bunjaeng jiyeog-e
Church-∅-{TOP, also, only} relief supplies-ACC troubled areas-LOC
bonae-eoss-da.
send-PST-DCL
'Church sent the relief supplies to the troubled areas.'

Particles *-e* LOCATIVE and *-ege* DATIVE are other case particles that are optionally deleted when they are used with auxiliary particles as shown in (36)-(37)

(36) a. Seho-ga yuchiwon-*e*-{neun, do, man} dani-nda.
Seho-NOM nursery-LOC-{TOP, also, only} attend-DCL
'Seho attends a nursery.'

b. Hwanho-ga jangnangam-eul Seho-*ege*-{neun, do, man} ju-eoss-da.
Hwanho-NOM toy-ACC Seho-DAT-{TOP, also, only} give-PST-DCL
'Hwanho gave a toy to Seho.'

(37) a. Seho-ga yuchiwon-∅-{neun, do, man} dani-nda.
Seho-NOM nursery-∅-{TOP, also, only} attend-DCL
'Seho attends a nursery.'

b. Hwanho-ga jangnangam-eul Seho-∅-{neun, do, man} ju-eoss-da.
Hwanho-NOM toy-ACC Seho-∅-{TOP, also, only} give-PST-DCL
'Hwanho gave a toy to Seho.'

Other case particles should be retained when auxiliary particles are attached to the case marked noun phrases.

(38) a. Seho-ga anbang-*eseo*-{neun, do, man} jal ja-nda.
Seho-NOM master bedroom-LOC-{TOP, also, only} well sleep-DCL
'Seho sleeps well in the master bedroom.'

b. Hwanho-ga gawi-*lo*-{neun, do, man} joingi-leul jaleu-eoss-da.
Hwanho-NOM scissors-INST-{TOP, also, only} paper-ACC cut-PST-DCL
'Hwanho cut the paper with scissors.'

c. Hwanho-ga Cameron-*gwa*-{neun, do, man} manna-ss-da.
Hwanho-NOM Cameron-COM-{TOP, also, only} meet-PST-DCL
'Hwanho met Cameron.'

d. Gae-ga goyangi-*boda*-{neun, do, man} ttogttogha-da.
Dog-NOM cat-COMP-{TOP, also, only} smart-DCL
'Dogs are smarter than cats.'

(39) *Seho-ga anbang-∅-{neun, do, man} jal ja-nda.
Seho-NOM master bedroom-∅-{TOP, also, only} well sleep-DCL
'Seho sleeps well in the master bedroom.'

a. *Hwanho-ga gawi-∅-{neun, do, man} joingi-leul jaleu-eoss-da.
Hwanho-NOM scissors-∅-{TOP, also, only} paper-ACC cut-PST-DCL
'Hwanho cut the paper with scissors.'

b. *Hwanho-ga Cameron-∅-{neun, do, man} manna-ss-da.
Hwanho-NOM Cameron-∅-{TOP, also, only} meet-PST-DCL
'Hwanho met Cameron.'

c. *Gae-ga goyangi-∅-{neun, do, man} ttogttogha-da.
Dog-NOM cat-∅-{TOP, also, only} smart-DCL
'Dogs are smarter than cats.'

We can summarise the case particle deletion phenomenon as (40).

(40) Case particle deletion

a. Obligatory deletion
If the NOMINATIVE case particle *-i/-ga* or the ACCUSATIVE case particle *-eul/-leul* is followed by an auxiliary particle, the case particle is obligatorily deleted.

b. Optional deletion
if the nominative case particles *-kkeso*, *-eseo*, the LOCATIVE case particle *-e* or the DATIVE case particle *-ege* is followed by an auxiliary particle, the case particle is optionally deleted.

2.3.2 Case Particle Unrealisation

Together with the case particle deletion presented in the previous section, *case particle unrealisation* is also a source of case ambiguity in Korean. Consider the following examples.

(41) a. Keu-n il-*i* na-ass-da.
Big-ADN event-NOM happen-PST-DCL
'A big incident has happened.'

b. Imo-nim-*kkeseo* o-si-eoss-da.
Aunt-HON-NOM come-HON-PST-DCL
'Aunt came.'

c. Seho-ga geu sangja-*leul* yeol-eoss-da.
Seho-NOM the box-ACC open-PST-DCL
'Seho opened the box.'

d. Hwanho-ga chingujib-*e* ga-ass-da.
Hwanho-NOM friend-house-LOC go-PST-DCL
'Hwanho went to a friend's house.'

(42) a. Keun il-∅ na-ass-da.
Big event-∅ happen-PST-DCL
'A big incident has happened.'

b. Imo-nim-∅ o-si-eoss-da.
Aunt-HON-∅ come-HON-PST-DCL
'Aunt came.'

c. Seho-ga geu sangja-∅ yeol-eoss-da.
Seho-NOM the box-∅ open-PST-DCL
'Seho opened the box.'

d. Hwanho-∅ chingujib-∅ ga-ass-da.
Hwanho-∅ friend-house-∅ go-PST-DCL
'Hwanho went to friend's house.'

Case particles *-i/-ga*, *-kkeseo* NOMINATIVE, *-eul/-leul* ACCUSATIVE, and *-e* LOCATIVE in (41) are not realised in (42) and the noun phrases *il* 'event', *sangja* 'box', and *chingujib* 'friend's house' are used without any particles. Note that two noun phrases are occurring without case particles in (42d).

In addition to the above case particles, *-ege* DATIVE, *-eulo/-lo* FUNCTION, and *-gwa/-wa* COMITATIVE can also be optionally unrealised as shown in (43) and (44).[16]

(43) a. Seonsaeng-nim-kkeseo seonmul-eul Hwanho-*ege* ju-si-eoss-da.
Teacher-HON-NOM present-ACC Hwanho-DAT give-HON-PST-DCL
'The teacher gave a present to Hwanho.'

b. Hwanho-ga Sean-eul chingu-*lo* sam-ass-da.
Hwanho-NOM Sean-ACC friend-FUNC make-PST-DCL
'Hwanho made Sean as his friend.'

c. Hwanho-ga halabeoji-*wa* dalm-ass-da.
Hwanho-NOM grandfather-COM look-like-PST-DCL
'Hwanho looked like his grandfather.'

(44) a. Seonsaeng-nim-kkeseo seonmul-eul Hwanho-∅ ju-si-eoss-da.
Teacher-HON-NOM present-ACC Hwanho-∅ give-HON-PST-DCL
'The teacher gave a present to Hwanho.'

b. Hwanho-ga Sean-eul chingu-∅ sam-ass-da.
Hwanho-NOM Sean-ACC friend-∅ make-PST-DCL
'Hwanho made Sean as his friend.'

[16] Case particle unrealisation is an optional process.

c. Hwanho-ga halabeoji-∅ biseusha-da.
Hwanho-NOM grandfather-∅ look like-DCL
'Hwanho looks like his grandfather.'

Unlike particles *-i/-ga* NOMINATIVE and *-eul/-leul* ACCUSATIVE that can be unrealsed quite freely, particles *-e* LOCATIVE *-ege* DATIVE, *-eulo/-lo* FUNCTION, and *-gwa/-wa* COMITATIVE cannot be unrealised in many situations as shown in (45) and (46).

(45) a. Yeho-ga inhyeong-eul gabang-*e* neoh-eoss-da.
Yeho-NOM doll-ACC bag-LOC put-PST-DCL
'Yeho put a doll in a bag.'

b. Uli-ga cha-leul ius-*ege* pal-ass-da.
We-NOM car-ACC neighbour-DAT sell-PST-DCL
'We sold a car to a neighbour.'

c. Hwanho-ga gugsu-leul achim-*eulo* meog-eoss-da.
Hwanho-NOM noodle-ACC breakfast-FUNC eat-PST-DCL
'Hwanho ate noodle as breakfast.'

d. Seho-ga Seonho-*wa* nol-ass-da.
Seho-NOM Seonho-COM play-PST-DCL
'Seho played with Seonho.'

(46) a. *Yeho-ga inhyeong-eul gabang-∅ neoh-eoss-da.
Yeho-NOM doll-ACC bag-∅ put-PST-DCL
'Yeho put a doll in a bag.'

b. *Uli-ga cha-leul ius-∅ pal-ass-da.
We-NOM car-ACC neighbour-∅ sell-PST-DCL
'We sold a car to a neighbour.'

c. *Hwanho-ga gugsu-leul achim-∅ meog-eoss-da.
Hwanho-NOM noodle-ACC breakfast-∅ eat-PST-DCL
'Hwanho ate noodle as breakfast.'

d. Seho-ga Seonho-∅ nol-ass-da.
Seho-NOM Seonho-∅ play-PST-DCL
'Seho played with Seonho.'

Other case particles such as *-eseo* NOMINATIVE, *-eseo* LOCATIVE, *-eulo/-lo* INSTRUMENTAL/DIRECTION and *-boda* COMPARATIVE should be always realised and the cases must be explicitly marked.

(47) a. Samsung-*eseo* sinjepum-eul sipanha-yeoss-da.
Samsung-NOM new product-ACC launch-PST-DCL
'Samsung lanched a new product.'

b. Hwanho-ga gyohoe-*eseo* Eva-leul manna-ass-da.
Hwanho-NOM church-LOC Eva-ACC meet-PST-DCL
'Hwanho met Eva at the church.'

c. Seho-ga saegyeonpil-*lo* geulim-eul geuli-nda.
Seho-NOM colour pencil-INST picture-ACC draw-DCL
'Seho is drawing a picture with a colour pencil.'

d. Hwanho-ga gong-eul ulijjog-*eulo* cha-ass-da.
Hwanho-NOM ball-ACC our side-DIR kick-PST-DCL
'Hwanho kicked the ball toward us'

e. Seho-ga mul-eul juseu-*boda* johaha-nda.
Seho-NOM water-ACC juice-COMP like-DCL
'Seho likes water better than juice.'

(48) a. Samsung-∅ sinjepum-eul sipanha-yeoss-da.
Samsung-∅ new product-ACC launch-PST-DCL
'Samsung lanched a new product.'

b. *Hwanho-ga gyohoe-∅ Eva-leul manna-ass-da.
Hwanho-NOM church-∅ Eva-ACC meet-PST-DCL
'Hwanho met Eva at the church.'

c. *Seho-ga saegyeonpil-∅ geulim-eul geuli-nda.
Seho-NOM colour pencil-∅ picture-ACC draw-DCL
'Seho is drawing a picture with a colour pencil.'

d. *Hwanho-ga gong-eul ulijjog-∅ cha-ass-da.
Hwanho-NOM ball-ACC our side-∅ kick-PST-DCL
'Hwanho kicked the ball toward us'

e. *Seho-ga mul-eul juseu-∅ johaha-nda.
Seho-NOM water-ACC juice-∅ like-DCL
'Seho likes water better than juice.'

However, when the particle *-eulo*/*-lo* INSTRUMENTAL is used to denote the manner/mode of an event, it can be unrealised in certain environment (See Section 2.3.3). Similar instances of case particle unrealisation are also observed when the particle *-e* LOCATIVE is used to denote the time of an event. This type of case particle unrealisation is illustrated in (49)-(50).

(49) a. Na-neun pyeongso-*e* geudeul-eul demyeondemyoenhage
I-TOP ordinary times-LOC they-ACC inattentively
daeha-yess-da.
confront-PST-DCL
'I usually confronted them inattentively.'

b. Haggyo-eseo choedaehan-*eulo* jiwon-eul ha-nda.
School-NOM maximum-INST support-ACC do-DCL

'The school gives a maximum support.'

(50) a. Na-neun pyeongso-∅ geudeul-eul demyeondemyoenhage
I-TOP ordinary times-∅ they-ACC inattentively
daeha-yess-da.
confront-PST-DCL
'I usually confronted them inattentively.'

b. Haggyo-eseo choedaehan-∅ jiwon-eul ha-nda.
School-NOM maximum-∅ support-ACC do-DCL
'The school gives a maximum support.'

We can summarise the case particle unrealisation phenomenon as (51).

(51) Case particle unrealisation
Case particles *-i/-ga*, *-kkeso* NOMINATIVE, *-eul/-leul* ACCUSATIVE, *-e* LOCATIVE, *-ege* DATIVE, *-eulo/-lo* INSTRUMENTAL/FUNCTION, and *-gwa/-wa* COMINATIVE can be optionally unrealsed under certain conditions.

2.3.3 Conditions of the Case Particle Deletion and Unrealisation

Kim (1998) claims that the case particle unrealisation is only possible for the noun phrases used as arguments of predicates. This claim is supported by the fact that the noun phrases that do not permit the case particle unrealisation in (47) are all non-arguments and the sentences without the noun phrases are perfectly acceptable sentences as shown in (52).

(52) a. Seho-ga geulim-eul geuli-nda.
Seho-NOM picture-ACC draw-DCL
'Seho is drawing a picture.'

b. Hwanho-ga gong-eul cha-ass-da.
Hwanho-NOM ball-ACC kick-PST-DCL
'Hwanho kicked the ball.'

c. Seho-ga mul-eul johaha-nda.
Seho-NOM water-ACC like-DCL
'Seho likes water.'

According to Kim (1998), case particle unrealisation is possible for argument noun phrases since the cases are structurally determined even without the case particles. In other words, the relationships between argument noun phrases and the governing predicate can be recognised without explicit markings. This explanation is very persuasive. However, it is not sufficient. Consider the following examples.

(53) a. Park seonsaeng-i sikkeuleob-eun hwangyeong-*e* igsugha-da.
Park teacher-NOM noisy-ADN environment-LOC familiar-DCL
'Mr Park is familiar with noisy environments.'

b. Seho-ga Morgan-*gwa* chinha-da.
Seho-NOM Morgan-COM intimate with-DCL
'Seho is intimate with Morgan.'

(54) a. *Park seonsaeng-i igsugha-da.
Park teacher-NOM familiar-DCL
'Mr Park is familiar with (something).'

b. *Seho-ga chinha-da.
Seho-NOM intimate with-DCL
'Seho is intimate with (somebody).'

(55) a. *Park seonsaeng-i sikkeuleob-eun hwangyeong-∅ igsugha-da.
Park teacher-NOM noisy-ADN environment-∅ familiar-DCL
'Mr Park is familiar with noisy environments.'

b. *Seho-ga Morgan-∅ chinha-da.
Seho-NOM Morgan-∅ intimate with-DCL
'Seho is intimate with Morgan.'

The noun phrases *hwangyeong-e* 'environment-LOC' and *Morgan-gwa* 'Morgan-COM' in (53) are arguments that cannot be dropped as shown in (54). Therefore, we expect that the case particles can be unrealised in these noun phrases. However, the sentences, in which the case particles *-e* LOCATIVE and *-gwa* COMITATIVE, are unrealised are uninterpretable. From this, we can conclude that not all argument noun phrases are subject to the argument noun phrase condition of the case particle unrealisation.[17]

Furthermore, case particle unrealisation in (50) takes place with non-argument noun phrases. According to Chung (1998), this type of case particle unrealisation is due to the semantic properties of the preceding nouns. For example, the noun *pyeongso* 'ordinary times' in (49a) bears a strong sense of 'time'. Consequently, this noun does not have any difficulty in functioning as an adverbial in the sentence even without the LOCATIVE case particle *-e* which denotes 'a point of time'. Similarly, the noun *choedaehan* 'maximum' bears a sense of 'manner of an action' and it can also function as an adverbial without the help of the case particle *-eulo/-lo*. Chung (1998) labelled these nouns as *adverbial nouns*.

The conditions of the case particle unrealisation are mostly applicable to the case particle deletion. However, case particles cannot be deleted from the adverbial noun phrases.

[17] Kim (1998) pointed out that the cases of the noun phrases that permit the case particle unrealisation are all interchangeable with the ACCUSATIVE case particle *-eul/-leul*. See Section 2.3.4.

(56) a. Na-neun pyeongso-*e*-{neun, do, man} geudeul-eul
I-TOP ordinary times-LOC-{TOP, also, only} they-ACC
demyeondemyoenhage daeha-yess-da.
inattentively confront-PST-DCL
'I usually confronted them inattentively.'

b. Haggyo-eseo choedaehan-*eulo*-{neun, do, man} jiwon-eul ha-nda.
School-NOM maximum-INST-{TOP, also, only} support-ACC do-DCL
'The school gives maximum support.'

(57) a. *Na-neun pyeongso-∅-{neun, do, man} geudeul-eul
I-TOP ordinary times-LOC-{TOP, also, only} they-ACC
demyeondemyoenhage daeha-yess-da.
inattentively confront-PST-DCL
'I usually confronted them inattentively.'

b. *Haggyo-eseo choedaehan-∅-{eun, do, man} jiwon-eul ha-nda.
School-NOM maximum-∅-{TOP, also, only} support-ACC do-DCL
'The school gives maximum support.'

In summary, we can tentatively conclude that case particle deletion and unrealisation occur with noun phrases when the unmarked cases are predicted either by the head-dependent relationships of the noun phrases and the predicates or the semantic properties of the noun phrases.

2.3.4 Case Particle Alternations

Diathesis alternations are the changes of the realisation of the argument structure of a verb that are sometimes accompanied by changes in meaning (Levin, 1993). Diathesis alternations are realised as case particle alternations in Korean as illustrated in (58) and (59).

(58) a. Seho-ga Jaehwi-*wa* manna-ass-da.
Seho-NOM Jaehwi-COM meet-PST-DCL
'Seho met Jaehwi.'

b. Hwanho-ga Seho-ege sonmog-*i* jabhi-eoss-da.
Hwanho-NOM Seho-DAT wrist-NOM be held-PST-DCL
'Hwanho's wrist was held by Seho.'

c. Seho-ga yuchiwon-*e* ga-ass-da.
Seho-NOM nursery-LOC go-PST-DCL
'Seho went to the nursery.'

d. Hwanho-neun gyosil-*lo* hyangha-yeoss-da.
Hwanho-TOP classroom-LOC proceed-PST-DCL
'Hwanho proceeded to the classroom.'

(59) a. Seho-ga Jaehwi-*leul* manna-ass-da.
Seho-NOM Jaehwi-ACC meet-PST-DCL
'Seho met Jaehwi.'

b. Hwanho-ga Seho-ege sonmog-*eul* jabhi-eoss-da.
Hwanho-NOM Seho-DAT wrist-ACC be held-PST-DCL
'Hwanho's wrist was held by Seho.'

c. Seho-ga yuchiwon-*eul* ga-ass-da.
Seho-NOM nursery-ACC go-PST-DCL
'Seho went to the nursery.'

d. Hwanho-neun gyosil-*eul* hyangha-yeoss-da.
Hwanho-TOP classroom-ACC proceed-PST-DCL
'Hwanho proceeded to the classroom.'

In the above examples, case particle alternations between the three case particles NOMINATIVE, LOCATIVE, DIRECTIONAL and COMITATIVE, and the ACCUSATIVE case particle are observed. There were several efforts to account for the case particle alternation with regards to topicalisation (Im, 1979; Lee, 1988), focusing (Kim, 1994), and semantic roles (Yu and Lee, 1996). Yoo (2002) covered a variety of case particle alternation patterns shown in (60).

(60) Case particle alternations in Korean

a. Structural case vs. structural case

-i/-ga NOMINATIVE ↔ *-eul/-leul* ACCUSATIVE

-i/-ga NOMINATIVE ↔ *-ui* GENITIVE

-eul/-leul ACCUSATIVE ↔ *-ui* GENITIVE

b. Structural case vs. inherent case

-i/-ga NOMINATIVE ↔ *-e* LOCATIVE

-i/-ga NOMINATIVE ↔ *-ege* DATIVE

-i/-ga NOMINATIVE ↔ *-eulo/-lo* DIRECTIONAL

-eul/-leul ACCUSATIVE ↔ *-eseo* LOCATIVE

-eul/-leul ACCUSATIVE ↔ *-eulo/-lo* DIRECTIONAL

-eul/-leul ACCUSATIVE ↔ *-e* LOCATIVE

-eul/-leul ACCUSATIVE ↔ *-ege* DATIVE

-eul/-leul ACCUSATIVE ↔ *-gwa/-wa* COMITATIVE

c. Inherent case vs. inherent case

-e LOCATIVE ↔ *-eulo/-lo* DIRECTIONAL

-e LOCATIVE ↔ *-gwa/-wa* COMITATIVE

When a human tries to infer the hidden case particle for an ambiguous instance, there can be more than one answer due to the case particle alternation phenomenon. Consequently,

it is more appropriate to evaluate the output of the case ambiguity resolution system on multiple human annotations than a single annotation.

2.3.5 Relative Clause Constructions

A clause which modifies a head nominal is broadly called a *relative* or an *adnominal* clause (Sohn, 1999). In a narrow sense, the relative clause construction is a subtype of the adnominal clause construction distinguished from another subtype, the *appositive clause* construction (Chang, 1993; Nam and Koh, 1993). Consider the following examples.

(61) a. Seho-neun [Hwanho-ga hangug-e ga-n] sasil-eul najung-e
Seho-TOP [Hwanho-NOM Korea-LOC go-ADN] fact-ACC last-LOC
al-ass-da.
know-PST-DCL
'Seho realised the fact that Hwanho went to Korea later.

b. Seho-do [Hwanho-ga dani-n] yuchiwon-eul silheoha-yeoss-da.
Seho-also [Hwanho-NOM attend-ADN] nursery-ACC dislike-PST-DCL
'Seho also disliked the nursery which Hwanho had attended.'

An adnominal clause in Korean is constructed by attaching an *adnominaliser* to the main predicate of the modifying clause as shown in (61). The adnominal clause in (61a) is an appositive clause which maintains a complete sentential form. On the other hand, the adnominal clause is a relative clause which lacks a constituent, i.e. *yuchiwon-e* 'nursery-LOC'. In other words, we regard that the noun phrase *yuchiwon-e* 'nursery-LOC' has been moved out or extracted from the adnominal clause.

Although there are some restrictions, any nominal can be extracted as a head nominal in principle. When a nominal is extracted, it loses the case particle it had and a new case particle is attached to mark the case of the nominal as a constituent of the main clause. Thus, it is not easy to infer the grammatical status of exracted nominal it had in the relative clause before the extraction. This problem can be viewed as another type of case ambiguity. However, we are not dealing with this problem in this thesis.[18]

2.4 Related Work

This section surveys previous work related to this thesis. We especially pay our attention to the work on Korean since it is directly related to the current work. Work on other languages

[18] Some studies attacked this problem with similar methods used in case ambiguity resolution in non-relative clauses/sentences. See Section 2.4.1.4.

salam 'man'	*Hwanho, Seho*: [+human, +animate, -edible]
mul 'water'	*sagwa* 'apple': [-human, -animate, +edible]
mos 'nail'	[-human, -animate, -edible]
jileongi 'earth worm'	[-human +animate, -edible]
meog- 'eat'	NOM: +human:1.0, -human/+animate:0.8, -animate:0.0
	ACC: +edible:1.0, -edible:0.0

Figure 2.10: *Example of semantic feature marking and feature concord information in a lexicon*

is briefly presented.

2.4.1 Work on Korean

2.4.1.1 Knowledge-Based Approaches to Case Ambiguity Resolution

Yoon and Kim (1989a,b) provide a typical example of a knowledge-based case ambiguity resolution method in the context of syntactic analysis within the Lexical Functional Grammar (Kaplan and Bresnan, 1982) framework. The proposed methods are as follows:[19]

- Grammatical Relation Mapping Method
  When there is only one instance of case ambiguity in a clause/sentence, unambiguous arguments are matched with the appropriate slots of the subcategorisation frame of the predicate of the clause/sentence. The remaining slot is matched with the ambiguous argument and the case is mapped from the slot.
- Constituent Comparison Method
  This method requires a lexicon with comprehensive semantic feature marking and feature concord information (Figure 2.10). When there are two or more candidate cases for a nominal, an optimal selection is made according to the semantic feature marking of the nominal and the feature concord information in the lexicon.
- Default Word Order Mapping Method
  This method assumes that there is a predominant word order in Korean although it is a relatively free word order language. This study recognises NOMINATIVE > SUBJECT_2 > OBJECT_1 > OBJECT_θ as a default word order of Korean. Ambiguous cases are decided according to the default word order.

[19] LFG-specific arguments are generalised.

To resolve a case ambiguity, the above methods are applied one by one until a satisfying solution is found. Yoon and Kim (1989b) stated that the above methods were implemented in a syntactic analyser based on LFG framework. However, any further real example or evaluation result has not been reported.

In Yang and Shim (1999), a case ambiguity resolution algorithm using a thesaurus and a subcategorisation frame dictionary was presented. The thesaurus and the subcategorisation dictionary used in this work were still under development (Seo, 1998) when this study was conducted. They contained 91,000 nominal and 12,804 predicate headwords respectively. One peculiar feature of this subcategorisation frame dictionary is that the every subcategorisation entry includes typical nominal words for each argument slot rather than semantic markers or concept classes. These nominal words are generalised using the thesaurus. The proposed case ambiguity resolution method is as follows:

- Compare the input sentence pattern with the subcategorisation frame in the dictionary and assess the confidence of each candidate case. The confidence score is the sum of the weights determined according to the following criterion:
  - w_1: The input nominal matches the semantic information of an argument slot in a subcategorisation frame.
  - w_2: The input sentence pattern completely matches a subcategorisation frame.
  - w_3: The input nominal is a 'time' word and expected to be an adverbial.
  - $w_1 > w_2 + w_3$
- Apply the above procedure to every subcategorisation frame for the predicate of the input sentence. Choose a case which has the highest score.
- If there are unresolved case ambiguities, apply the following heuristic:
  - If the predicate of the input sentence is a predicate which can take multiple nominative arguments, decide the target case as NOMINATIVE.
  - If the target nominal is not accompanied by an auxiliary particle *-eun/-neun* TOPIC and there is no sibling nominals marked as ACCUSATIVE case and the predicate is a transitive verb, decide the target case as ACCUSATIVE.
  - If a NOMINATIVE nominal is present and an ACCUSATIVE nominal is not present among the siblings of the target nominal, decide the target case as ACCUSATIVE.
  - Decide any remaining target cases as NOMINATIVE cases.

	test-set1		test-set2	
	num	%	num	%
baseline correct	429	90.3	226	53.7
baseline incorrect	46	9.7	195	46.3
correct	460	96.8	364	86.5
incorrect	15	3.2	57	13.5

Table 2.2: *Experimental results of Yang and Shim (1999)*

The above case ambiguity resolution methods were applied on two test sets consisting of 475 and 421 ambiguous instances respectively. Weights were set as $w_1 = 4$, $w_2 = 2$ and $w_3 = 1$. Considered target cases were NOMINATIVE, ACCUSATIVE and ADVERBIAL. The experimental result obtained by evaluating the output on human-annotated data is shown in Table 2.2. The baseline strategy was to choose the most frequently used case particle *-ga* NOMINATIVE. The baseline accuracy on the test-set1 reached 90.3%. Test-set2 was constructed deliberately excluding ambiguous nominals occurring with *-eun/-neun* TOPIC for there was a high tendency that the hidden cases of those nominals were NOMINATIVE cases.

2.4.1.2 Statistical Approaches to Case Ambiguity Resolution

Yang and Kim (1994b) is one of the early attempts which adopted statistical methods to resolve case ambiguity in Korean. In this study, only NOMINATIVE and ACCUSATIVE cases were considered. The statistical case decision was guided by *SR* (Statistical Relevance Score), which is the sum of *SS* (Subcategorisation Score) and *CS* (Co-occurrence Score). These scores are calculated using the frequency counts of v (predicate), n (nominal), and j (case particle) obtained from a corpus through the following equations.[20]

$$SR(v,n,j) = SS(v,j) + c * CS(v,n,j), \quad c > 1 \tag{2.1}$$

$$SS(v,j) = \frac{f(v,j)}{f(v)}, \quad j \in \{\text{NOMINATIVE}, \text{ACCUSATIVE}\} \tag{2.2}$$

$$CS(v,n,j) = \frac{f(v,n,j)}{f(n,j) + f(v,j) - f(v,n,j)} \tag{2.3}$$

The Subcategorisation Score (*SS*) (2.2) measures the strength of the association between a predicate and a given case particle. For example, a transitive verb and the ACCUSATIVE case particle will yield a high *SS* value. This score is equivalent to the conditional probability of a case particle given a predicate.

[20] Original equations were slightly modified for a better presentation.

	CS	SS	$SS + 8 \times CS$
correct	113	119	214
incorrect	6	33	18
inapplicable	229	116	116
coverage	34.2%	66.7%	66.7%
accuracy (applicable instances)	95.0%	85.8%	92.2%
accuracy (all instances)	32.5%	57.2%	61.5%

Table 2.3: *Experimental result of Yang and Kim (1994b)*

The Co-occurrence Score (CS) (2.3) is a measure of the degree of co-occurrence between a predicate v and a nominal n under a particular case relation denoted by a case particle j. It is calculated by dividing the frequency of a triple $\langle v, n, j \rangle$ with the subtraction of the frequency of a triple $\langle v, n, j \rangle$ from the sum of the frequencies of pairs $\langle n, j \rangle$ and $\langle v, j \rangle$.

The final Statistical Relevance Score (SR) (2.1) is defined as the weighted sum of the the above two scores. CS has more contribution to SR than SS since CS gets a large weight ($c > 1$). A case particle which maximises the SR is is selected as an answer for a given case ambiguity problem.

For the training data construction, an unspecified syntactic analyser was applied to a 330,000-word corpus of computer science domain. As a result, frequency counts of 19,800 $\langle v, n, j \rangle$ triplets were collected. The accuracy of this data collection method which was measured on 500 sample sentences was 93.3%.

The case ambiguity resolution procedure was tested on 348 ambiguous instances. Since SR does not have any form of smoothing, it could not be applied to 116 instances. The output of the system was compared to a human annotation. The reported accuracy is 92.9%. If we take account of the inapplicable instances, the accuracy becomes 61.5%. The experimental result is summarised in Table 2.3.

Kim (1996b) introduced an *Association Measure* influenced by other work (Resnik, 1993). This measure was defined as the multiplication of the conditional probability of a nominal given a predicate and a case particle, and the conditional mutual information of the predicate and the nominal given the case particle as shown in (2.4)-(2.5). This work used a class-based smoothing technique to cope with the unseen $\langle v, n, j \rangle$. For unseen $\langle v, n, j \rangle$ the nominal words are replaced by their conceptual classes obtained from an experimental thesaurus (Im, 1993) (2.6).

$$Assoc(v, n, j) = P(n|v, j)I(v; n|j) \tag{2.4}$$

verb	correct (word-based)	correct (class-based)	incorrect	accuracy (%)
naeli- 'take down', 'come down'	4	50	12	81.89
mandeul- 'make'	7	65	12	85.71
meog- 'eat'	20	43	11	85.14
bad- 'receive'	12	179	30	86.43
bonae- 'send'	3	37	4	90.90
sseu- 'write', 'put on'	42	117	15	91.37
anj- 'sit'	11	19	4	88.24
yeol-/yeolli- 'open/be opened'	4	31	10	77.27
jis- 'build', 'make'	10	32	12	77.78
ta- 'get on', 'burn'	3	32	5	88.89
heuleu 'flow'	0	1	0	100.00
Total	116	606	115	86.26

Table 2.4: *Experimental result of Kim (1996b)*

$$I(v;n|j) = \log_2 \frac{P(v,n|j)}{P(v|j)P(n|j)} \tag{2.5}$$

$$I(v;n|j) = \log_2 \frac{P(v,class(n)|j)}{P(v|j)P(class(n)|j)} \tag{2.6}$$

Training data consisting of triplets in the form of $\langle v,n,j\rangle$ was constructed through a manual filtering of the initial set of triplets suggested by an automatic procedure which couples a $\langle n,j\rangle$ pair to the nearest possible governing predicate. For smoothing, $\langle v,class(n),j\rangle$ triplets were also collected. If a nominal belongs to multiple classes, correct class was determined by a human judge. Neither the size of the corpus nor the size of the training set has been reported.

This study applied the proposed *Association Measure* to the case ambiguity resolution. 837 ambiguous instances were collected and annotated by a human judge for the test. The reported accuracy is 86.26%. This study considered NOMINATIVE, ACCUSATIVE, and ADVERBIAL cases. ADVERBIAL case includes all cases other than NOMINATIVE and ACCUSATIVE cases. The test set was constructed in a very restricted way. The test instances were collected for only 12 verbs and the numbers of test instances per a verb were not balanced. The experimental result is displayed in Table 2.4.

Chung (1999) used the *Association Measure* (2.7) which was borrowed from Yoon et al. (1997) and Yoon (1998). This work incorporated a class-based smoothing technique utilising the experimental thesaurus of Cho and Ok (1997) as shown in Equations (2.8).

verb	accuracy
naeli- 'take down', 'come down'	83.01%
mandeul- 'make'	81.48%
meog- 'eat'	90.38%
bad- 'receive'	85.96%
bonae- 'send'	86.79%
sseu- 'write', 'put on'	89.65%
anj- 'sit'	78.84%
yeol-/yeolli- 'open/be opened'	90.00%
jis- 'build', 'make'	96.22%
ta- 'get on', 'burn'	79.24%
average	81.16%

Table 2.5: *Experimental result of Chung (1999)*

$$Assoc(v,n,j) = \alpha \times \overline{Assoc}(v,n,j) + (1-\alpha) \times \overline{Assoc}(v,j) \qquad (0.5 \le \alpha \le 1) \tag{2.7}$$

$$\overline{Assoc}(v,n,j) = \max\left(P(n,j|v), \frac{P(class(n),j|v)}{N}\right) \tag{2.8}$$

$$\overline{Assoc}(v,j) = P(j|v) \tag{2.9}$$

Training data was collected by a simple heuristic method similar to the data collection methods of this thesis. The accuracy of this heuristic method was not reported. The data collection heuristic is as follows:

- Split mixed sentences into simple sentences according to the connective ending of the main predicates of the clauses.
- Extract $\langle v,n,j\rangle$ triplets for the last predicate of each simple sentence and nominals preceding the predicates.
- Take the last nominal from a compound nominal word.

From a 8,000,000-word corpus 624,200 $\langle v,n,j\rangle$ triplets were collected. These triplets were generalised using a thesaurus which contains 12,933 headwords. The final result was a set of 5,000 $\langle v,class(n),j\rangle$ triplets and their frequency counts.

For an evaluation, 534 ambiguous instances for 10 verbs chosen in Kim (1996b) were collected both from the training corpus and an independent test corpus. The reported accuracy measured on a human annotation is 86.16%, which is comparable to that of Kim (1996b). The experimental result is summarised in Table 2.5.

Lee et al. (1998) proposed a case ambiguity resolution method based on conceptual pattern and statistical information. From the set of $\langle v, n, j \rangle$ triplets extracted from a corpus, a set of CFP (Conceptual Frequency Patterns) in the form of $\langle\langle\langle c_1, f_1\rangle, \langle c_2, f_2\rangle, \ldots, \langle c_n, f_n\rangle\rangle, j, v\rangle$ was constructed, where c_i is a concept code and f_i is the frequency count of the concept code occurring with v and j. The concept codes are obtained by using Korean-Japanese dictionary for a machine translation system and a Japanese thesaurus (Ohno and Hamanishi, 1981). CFPs are further generalised by filtering out statistically insignificant conceptual codes producing a set of CPs (Conceptual Patterns). A CP has the form of $\langle\langle c_1, c_2, \ldots, c_n\rangle, j, v\rangle$. In addition, CD (Case Distribution) is used to supplement the CP for case ambiguity resolution.

$$CD(v, j) = \frac{f(v, j)}{f(v)}, \quad j \in \{\text{NOMINATIVE, ACCUSATIVE}\} \tag{2.10}$$

The proposed case ambiguity resolution method is as follows:

- Choose the candidate target cases referring to the subcategorisation frame information for the verb of input instance.
- Calculate the similarities between the concept code of the target nominal and each of the concept codes in the CPs containing candidate target cases and the verb.
- Pick the CP which contains the concept code most similar to the concept code of the target nominal word. Decide the target case as the case in the CP.
- If multiple CPs have the same similarities, select a case guided by the CD value of the input verb and candidate case particles.

For an experiment, a 6,000,000-word corpus is analysed by an unspecified partial parser and 5,138,000 $\langle v, n, j \rangle$ triplets for 84 high frequency verbs and NOMINATIVE and ACCUSATIVE case particles. The above method was applied to 284 sentences containing the 84 high frequency verbs. Each sentence had 3 or 4 ambiguous instances. The reported accuracy is 92%.

2.4.1.3 Case Ambiguity Resolution in Full Parsing

In Yang and Kim (1994a), a statistical case ambiguity module was integrated into a dependency parser. This module uses the following association measure, which is a modification of pointwise mutual information, to resolve the case ambiguity.

$$I(v, n, j) = \log_2 \frac{P(v, n, j)}{P(v)P(n)}, \quad j \in \{\text{NOMINATIVE, ACCUSATIVE}\} \tag{2.11}$$

The case ambiguity resolution module was trained on 800,000-word corpus. The parser including this module was tested on 185 sentences, and 174 sentences were correctly analysed (92.4%).

Eom et al. (1996) applied the same case ambiguity resolution module in a parser based on an extended context-free grammar formalism. This module was trained on a small corpus (300,000 words) and tested on 100 sentences. The ambiguity resolution module was applicable on only 33 sentences. The reported accuracy is 91%.

Yoon et al. (1997) and Yoon (1998) developed a statistical dependency parser in which attachment ambiguity and case ambiguity are resolved based on the Association Measure defined in (2.12).

$$Assoc(v, n, j) = \lambda_1 P(n, j|v) + \lambda_2 P(j|v), \quad \lambda_1 \gg \lambda_2 \tag{2.12}$$

The parser was trained on a 30,000,000-word corpus and tested on 408 sentences. The case ambiguity resolution module was applied on 256 instances and obtained 86.3% accuracy.

2.4.1.4 Case Decision for Head Nominals of Relative Clauses

A very similar task which closely resembles the case ambiguity resolution task is the task of recovering the original case of the head nominal of a relative clause as noted in Section 2.3.5.

Yang and Kim (1993) and Li et al. (1998) tackled this task using essentially the same techniques developed for the case ambiguity resolution task.

Lee et al. (2001) proposed a conditional probability model for case decision for head nominals of relative clauses as shown in (2.13).

$$\arg\max_{j \in J} P(j|v, e, n) \tag{2.13}$$

Besides the usual v and n, this work introduced the adnominal ending e as a feature. To estimate the conditional probability, Collins and Brooks (1995) style back-off strategy was adopted. With the same feature set, Lee et al. (2002) utilised Support Vector Machines (Vapnik, 1995) as a learning method.

In both studies, training data was collected from the KAIST Treebank and the system is evaluated on 1,595 test instances. The experimental results are displayed in Table 2.6.

	accuracy (%)				
	NOMINATIVE	ACCUSATIVE	ADVERBIAL	appositive	total
Conditional Probability	86.2	42.0	62.0	91.7	83.5
Support Vector Machines	84.4	62.9	92.0	97.5	88.7

Table 2.6: *Experimental results of Lee et al. (2001) and Lee et al. (2002)*

2.4.2 Work on Other Languages

2.4.2.1 Work on English

Ferro et al. (1999) introduced a grammatical relation finding model based on the Transformation-Based Learning framework. The grammatical relation function tagset consisted of 19 tags (subject, object, location object, location modifier, etc.). The model was trained on 1,963 tuples and tested on 748 tuples. The model yielded 77.3% precision and 63.6% recall (F-measure 69.8).

Blaheta and Charniak (2000) presented a maximum-entropy-inspired feature-tree based statistical model for function tag assignment. The task was recovering 20 function tags that can be appended to constituent labels (S, VP, NP, PP, etc.) in the Penn Treebank II (Bies et al., 1995). This model was trained on the section 2-21 of the Penn Treebank and tested on section 2 of the treebank. The proposed method achieved 88.450% precision and 88.493% recall (F-measure 88.472) when this method was applied to the parses in the test set. This model was also combined with a parser and produced 87.173% precision and 87.371% recall (F-measure 87.277) on the correctly labelled constituents output by the parser.

Buchholz (2002) adapted the Memory-Based Learning framework to the task of finding grammatical relations to head of verb chunks. A 10-fold cross validation experiment was performed on the sections 10-19 of the Wall Street Journal Corpus of the Penn Treebank II containing 21,747 sentences. This method reached 82.12% precision and 79.99% recall (F-measure 82.94). This model was also integrated in the Memory-Based Shallow Parser and yielded 79.96% precision and 66.47% recall (F-measure 72.59).

2.4.2.2 Work on German

de Lima (1997) proposed a simple grammatical relation assignment method based on a back-off model for German. Training data was constructed using a standard CFG parser with a hand-written grammar and a simple data collection heuristic. This method was applied to the task of distinguishing NOMINATIVE and ACCUSATIVE cases for nominal con-

stituents. As a result, this model produced 90.49% accuracy when trained on 47,547 tuples and tested on 24,178 test tuples.

In the context of an automatic creation of a syntactically and semantically annotated corpus of German, Brants et al. (1997) suggested a grammatical function assignment method based on a Markov tagging model. The task was tagging 17 grammatical function tags including subject, accusative object, dative, etc. to 9 phrasal categories (S, VP, NP, PP, etc.) identified by human annotators. This model was tested on a 1,200-sentence (24,000 words) German newspaper treebank using 10-fold cross validation. The average tagging accuracy was 94.2%.

2.5 Summary

In this chapter, we studied the theoretical work on case-related issues in Korean and identified six target case particles: *-i/-ga* NOMINATIVE, *-eul/-leul* ACCUSATIVE, *-e* LOCATIVE, *-ege* DATIVE, *-eulo/-lo* INSTRUMENTAL/DIRECTIONAL/FUNCTION and *-gwa/-wa* COMITATIVE. A careful investigation of the conditions of the case ambiguity revealed the plausibility of our approach to case ambiguity resolution.

We also surveyed the related work focusing on work on Korean. We saw that there still is a margin for more work: (1) Previous statistical approaches have used only minimal feature sets and tried to incorporate class-based smoothing methods to improve the ambiguity resolution models; (2) The number of target case particle was very limited; (3) Data collection methods did not get much attention; (4) In most cases, the size of the test set was relatively small and evaluation was performed on a single human annotation.

Chapter 3

Methodology

In this chapter, we focus on methodological issues concerning the training data collection method and statistical modelling for case ambiguity resolution in Korean. First of all, our task of case ambiguity resolution is defined in Section 3.1. Next, Section 3.2 introduces the corpora we use for our data collection and evaluation. Section 3.3 describes the statistical models for our task. Section 3.4 presents the data collection strategy we use and Section 3.5 briefly sketches the evaluation methods we adopt for the evaluation of our data collection methods and statistical models. Finally, Section 3.6 summarises this chapter.

3.1 The Task

Our task is to resolve case ambiguity in Korean caused by the case particle deletion or the case particle unrealisation described in Section 2.3. Specifically, the case ambiguity resolution task is to choose a case particle (j) for a nominal (n) which is used as either an argument or an adjunct of a predicate (v) without any accompanying case particle in a clause or in a sentence.[1] We call this operation *case decision*.[2] In other words, we resolve case ambiguity using a tool called case decision operation.

Table 3.1 shows the case particles involved in either deletion or unrealisation or both as identifed in Section 2.3. Based on this table, we establish six case particles in (62) as the target case particles for case ambiguity resolution.

[1] From now on we do not distiguish sentences from clauses unless specified.

[2] We have deliberately chosen the term 'case decision' to avoid the term 'case assignment' which is widely used in linguistic theories.

Case particle		Deletion	Unrealisation
NOMINATIVE	*-i/-ga*	✓	✓
	-eseo	✓	×
ACCUSATIVE	*-eul/-leul*	✓	✓
LOCATIVE	*-e*	✓	✓
DATIVE	*-ege*	✓	✓
INSTRUMENTAL	*-eulo/-lo*	×	✓
COMITATIVE	*-gwa/-wa*	×	✓

Table 3.1: *Case particles involved in deletion and unrealisation*

(62) Target case particles for the case ambiguity resolution task

a. *-i/-ga* NOMINATVIE

b. *-eul/-leul* ACCUSATIVE

c. *-e* LOCATIVE

d. *-ege* DATIVE

e. *-eulo/-lo* INSTRUMENTAL

f. *-gwa/-wa* COMITATIVE

As shown in (62), we are excluding the particle *-eseo* NOMINATIVE which is lexically ambiguous with *-eseo* LOCATIVE because our training data cannot provide training examples for the particle *-eseo* used as a NOMINATIVE case particle due to its limited annotation. The particle *-eseo* is annotated only as an adverbial case particle in our training data. Thus all instances of *-eseo* are interpreted as NOMINATIVE. As *-eseo* cannot be deleted or unrealised when it is used as an LOCATIVE case particle, we do not regard this particle as a target case particle. This treatment does not bring up any problem since *-eseo* is freely interchangeable with *-i/-ga* when it is used as a NOMINATIVE case particle as presented in Section 2.2.1.1. Human annotators will be able to choose *-i/-ga* NOMINATIVE instead of *-eseo* NOMINATIVE.

The particle *-gwa/-wa* COMITATIVE is also lexically ambiguous with *-gwa/-wa* CONJUNCTIVE particle as noted in Section 2.2.1.6. We decided to include this particle in the target case particles because it is possible to resolve this lexical ambiguity in the training data, although we do not expect that the disambiguation is perfect.[3]

Phonological and stylistic variants of the case particles are all consolidated into the representative forms. The same process is applied during the training stage.

[3]The part-of-speech tagger we use for the training data construction attempts to resolve this ambiguity. We also use a simple heuristic which considers the particle *-gwa/-wa* as a COMITATIVE case particle only when it is adjacent to the main predicate of a sentence.

Strictly speaking, we are not directly tackling the case ambiguity problem but indirectly by recasting the case ambiguity problem to the case particle ambiguity problem. Therefore, refined disambiguation is not possible for the case particles *-e* locative/dative and *-eulo*/*-lo* INSTUMENTAL/DIRECTIONAL/FUNCTION that can mark more than one cases. Nevertheless, this method is still useful since the refined cases can be recognised by the contexts as described in Section 2.2.

Finally, the case decision operation is formally defined as (3.1).

$$CD : \langle n, v, \vec{c} \rangle \rightarrow j, \quad j \in J \tag{3.1}$$

where

- J is the set of candidate case particles.
  $J = \{$*-i*/*-ga* NOM, *-eul*/*-leul* ACC, *-e* LOC, *-ege* DAT, *-eulo*/*-lo* INST, *-gwa*/*-wa* COM$\}$
- n is the focus nominal.
- v is the predicate.
- $\vec{c}$ is the vector of contextual information which can be obtained from the sentence which n and v belong to.

The candidicate case particles are selected according to the case particles that are either 'deleted' or 'unrealised' as presented in Section 2.3. The vector $\vec{c}$ contains contextual information which can be gathered from the sentence.

Case ambiguity resolution task is essentially a *classification* task since the case decision operation is involved in mutually exclusive categorial assignment. Classification or categorisation is defined as the task of assigning objects from a universe to two or more pre-defined *classes* or *categories* (Mitchell, 1997; Manning and Schütze, 1999; Dagan and Wintner, 2004).

3.2 Corpora

The primary language resources we use are raw and part-of-speech tagged corpora. We also use syntactically annotated corpora (treebanks) for training data collection method evaluation and test data preparation.

3.2.1 The Yonsei Corpora

The Yonsei Corpora (Seo, 1999)[4] are a set of modern Korean corpora compiled for corpus-based lexicography and language researches by Yonsei University in Korea. The Yonsei Corpora are composed of nine sub-corpora and the corpora as a whole contain 41,240,000 words[5] of written text and 760,000 words of transcribed speech. Two of the sub-corpora, YSC-1 (2,880,000 words) and YSC-2 (1,100,000 words), are balanced corpora that contain texts from a range of genres (newspaper, magazines, books, etc.) and subjects (general, philosophy, religion, social science, natural science, art, literature and history, etc.) YSC-3 and YSC-5 through YSC-7 were compiled from texts of specific periods. YSC-3 (5,900,000 words) contains written texts published in the 1980s. YSC-5 (8,620,000 words) is from the 1970s, YSC-6 (7,256,000 words) is from the 1960s, and YSC-7 (13,710,000 words) is from the 1990s. YSC-8 (898,000 words) and YSC-9 (1,499,000 words) are special purpose corpora of school children's textbooks and general books. Finally, YSC-4 (760,000 words) is a corpus of transcribed speech and pseudo-speech. Most of the Yonsei Corpora were manually encoded and proof-read to ensure the high quality of the corpora. The Yonsei Corpora have XML-style mark-ups. In the body part of each file, individual sentences were delimited by new line characters.

3.2.2 The Sejong Corpora

The Sejong Corpora (Kang and Kim, 2001, 2004) are products of a long-term on-going government-funded Korean language resource construction project called 'The 21st Century Sejong Project' (`http://www.sejong.or.kr`). We use the 10,000,000-word raw corpus distributed for educational and research purposes in 2000 (SJC-1, Kim et al. 2000) and the 2001 distribution of 7,000,000-word raw corpus and 2,000,000-word part-of-speech tagged corpus (SJC-2 and SJC-3, Kim et al. 2001). The Sejong Corpora are also balanced corpora that consist of the texts from a variety of genres (books, magazines, newspapers, etc.) and subjects (general, news, education, imaginary, descriptive, humanity, society, science, art and life). The Sejong Corpora have been marked up using an extended TEI-Lite encoding scheme (Kang et al., 1998). We need to split out the individual sentences from each paragraph since sentence boundaries are not marked up in the text.

[4] In this thesis, the Yonsei Corpora refers to the 1998 edition of the corpora. We only use the written text part of the corpora.

[5] The term 'word' is slightly abused here. The linguistic unit delimited by white spaces are called *eojeol* 'wordform' in Korean and it is not identical with 'word'.

3.2.3 The KAIST Treebank

KAIST (Korean Advanced Institute of Science and Technology) developed a 30,000-sentence syntactically analysed corpus of Korean through a number of research projects. This treebank can be licensed from the KORTERM (Korean Terminology Research Center, `http://www.korterm.org`). Due to the high licensing cost, we use a subset of the corpus (12,084 sentences) which was publicly released.[6]

The KAIST Treebank adopted a phrase structure grammar which has strict restrictions on the form of rewrite rules as its annotation scheme to prevent the rapid increase of the number of the rewrite rules while effectively coping with the partial free word order of Korean (Lee et al., 1997b,c). Accordingly the KAIST Treebank considers morphemes as basic units of syntactic analysis. Morphemes that have syntactic roles such as particles and endings occupy individual nodes in parse trees to indicate the syntactic functions explicitly. However, the grammatical functions of nominals are not encoded in the treebank and there is no distinction between arguments and adjuncts.

In the KAIST Treebank, embedded clauses are not distinguished from verb phrases. Phrasal tag `S` is only used to mark the top-level sentence. Figure 3.1 shows an example of a parse tree and its encoding in the treebank for a sentence (63).[7]

(63) Ibhuboja-deul-i choeseon-eul daha-ess-seubnida.
candidates-PL-NOM best-ACC do-PST-DCL
'Candidates did their best.'

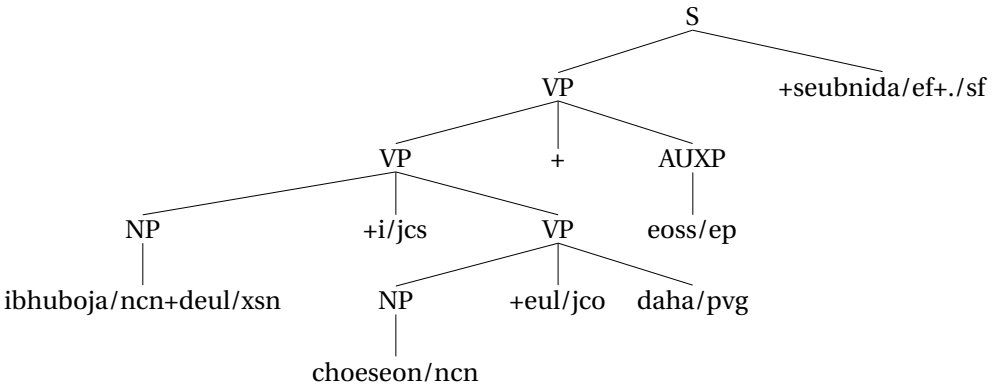
3.2.4 The Sejong Treebank

The Sejong Treebank is also a product of an on-going language resource development effort in the context of 'The 21st Century Sejong Project'. We use the 2003 distribution containing 13,174 syntactically annotated sentences (Kim and Rim, 2003).

In contrast to the KAIST Treebank, the syntactic analysis unit of the Sejong Treebank is *eojeol* 'wordform' and syntactic functions are encoded in parse trees, although they are quite limited. The function tags indicate whether a constituent is a SUBJECT or an OBJECT or an ADJUNCT of a head. In other words, only SUBJECT and OBJECT are treated as arguments in the Sejong Treebank. The phrase structure grammar adopted in the Sejong Treebank does not have any restriction on the form of rewrite rules.

[6] We obtained the treebank from `http://bi.snu.ac.kr/~sbpark/Step2000/`

[7] See Appendix B for the full lists of the KIAST part-of-speech and phrasal tags.




```
(S
   (VP
       (VP
           (NP ibhuboja/ncn+deul/xsn )+i/jcs
           (VP (NP choeseon/ncn )+eul/jco daha/pvg ))+(AUXP eoss/ep ))
    +seubnida/ef+./sf )
```

Figure 3.1: *An example parse tree and its encoding in the KAIST Treebank*

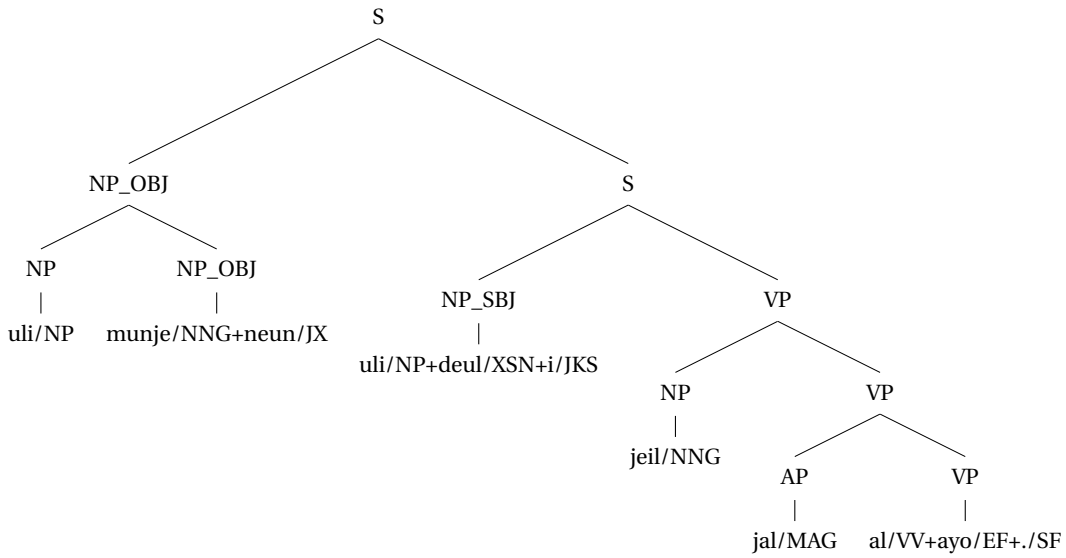
The Sejong Treebank attempts to distinguish embedded clauses from verb phrases. The phrasal tag S is used when an embedded verbal phrase has a subject.[8] However, this distinction is not very effective and there can be many subject-missing embedded clauses as subject dropping is quite common in Korean. An example of a parse tree and its encoding for a sentence (64) is illustrated in Figure 3.2.

(64) Uli munje-neun uli-deul-i jeil jal al-ayo.
our problem-TOP we-PL-NOM best well know-DCL
'We know our problem best.'

3.3 Statistical Models for Case Ambiguity Resolution

In this thesis, we suggest two case decision methods: the *discrete case decision* and the *sequential case decision*. In the discrete case decision, each instance of case ambiguities in a sentence is treated in isolation. Any existing information in the sentence can be used as clues for the case decision. However, if there exist two or more ambiguous instances in a sentence, each decision is independent from the other. In the sequential case decision, each case decision is performed one by one in a sequence. We choose to begin a case decision sequence from the ambiguous instance closest to the predicate of a sentence. The

[8] The full lists of the Sejong part-of-speech and phrasal tags are presented in Appendix C.




```
(S   (NP_OBJ   (NP uli/NP)
               (NP_OBJ munje/NNG + neun/JX))
     (S   (NP_SBJ uli/NP + deul/XSN + i/JKS)
          (VP   (NP jeil/NNG)
                (VP   (AP jal/MAG)
                      (VP al/VV + ayo/EF + ./SF)))))
```

Figure 3.2: *An example parse tree and its encoding in the Sejong Treebank*

result of a case decision is used as one of the clues for subsequent case decisions in the sentence. To model the two case decision methods, we use simple joint probabilistic models and a Markov chain tagging model.

3.3.1 Discrete Case Decision

To model the discrete case decision method, we represent a case decision operation as a straight-forward joint probabilistic event as shown in (3.2).

$$DCD(n, v, \vec{c}) = \underset{j \in J}{\arg\max}\, P(n, v, \vec{c}, j) \tag{3.2}$$

When we use a joint probability to represent an event that involves more than two variables. The ordering of the variables is very important for the following reasons:

First, when we estimate a joint probability with many variables, the joint probability needs be factored out as a product of a prior probability and a series of conditional probabilities using the chain rule. If the sub-events are equally related each other, the variable order-

ing won't affect the whole event. However, in a realistic problem, it matters which variable depends on which variable. Second, if we do not have a correct ordering of the variables, we cannot make any independence assumption to simplify the conditional probabilities decomposed from the joint probability. Independence assumptions are, of course, not always possible.

Regarding the variable ordering, Collins (1999) and Lapata (2001) introduce the following example which is presented in Russell and Norvig (1995), which we also briefly repeat.

The given situation is as follows:

- A person has a house with a burglar alarm and it works normally.
- She has two neighbours, John and Mary, who are fairly reliable at calling her at work when the alarm goes off.
- The alarm is triggered by two causes: a burglary or an earthquake.

The task is to build a model that supports queries such as "If Mary has called, what is the probability that there was a burglary?" or "If there is an earthquake, what is the probability that both John and Mary will call?"

To model the problem, we use 5 boolean-valued random variables: A alarm goes off or not, E there is an earthquake or not, B there is a burglary or not, J John calls or not, M Mary calls or not. To support all possible inferences, the model requires the joint probability $P(A,B,E,J,M)$. Now we simplify this joint probability.

The first step is decomposing the joint probability using the chain rule with the variable order $\langle B,E,A,J,M\rangle$ as shown in (3.3).

$$P(B,E,A,J,M) = P(B)P(E|B)P(A|E,B)P(J|A,E,B)P(M|A,E,B,J) \tag{3.3}$$

The next step is to make some independence assumptions to reduce the number of parameters following our real-world knowledge of *causality* such as:

- Earthquakes (E) and burglaries (B) usually do not have causal links.
- Earthquakes (E) and burglaries (B) both have strong links to the alarm (A).
- John's calling (J) has no direct link to earthquakes (E) and burglaries (B). John's calling (J) is only directly linked to the alarm (A).
- Similarly, Mary's calling (M) is only directly linked to the alarm (A).

The above reasoning of causality is translated into (3.4)–(3.7). From these we obtain the parameter-reduced version (3.8) of the initial joint probability (3.3).

$$P(E|B) = P(E) \tag{3.4}$$

$$P(A|E,B) = P(A|E,B) \tag{3.5}$$

$$P(J|A,E,B) = P(J|A) \tag{3.6}$$

$$P(M|A,E,B,J) = P(M|A) \tag{3.7}$$

$$P(B,E,A,J,M) = P(B)P(E)P(A|E,B)P(J|A)P(M|A) \tag{3.8}$$

(3.8) is a far more compact model with 10 parameters compared to the original one (3.3) with 31 parameters in worst case.

As illustrated in the above examples, when we decide a variable ordering, we need to follow the *causal relations* between the variables. In practice, though, the causal relations between the variables could not be as clear as the above example in many situations.

If we assume that we are only using v and n as features without any contextual information, and set the variable ordering as (v, j, n), (3.2) is formalised as follows:

$$DCD(n,v) = \arg\max_{j \in J} P(v,j,n) \tag{3.9}$$

$$= \arg\max_{j \in J} P(v)P(j|v)P(n|v,j) \tag{3.10}$$

$$= \arg\max_{j \in J} P(j|v)P(n|v,j) \tag{3.11}$$

The variable ordering (v, j, n) is based on our reasoning about the causal relations among the individual events v, j, and n participating in the joint event of a case decision operation. That is (1) a predicate (v) is selected, (2) a case slot (j) is provided, (3) a nominal word (n) is chosen and fill the case slot. The joint probability in (3.9) is factored as (3.10) and simplified as (3.11) since the first probability term $P(v)$ is a constant and does not affect the whole probability value.

To esitimate the probability, we use frequency counts from a corpus as (3.12) and (3.13), where $freq$ is the frequency count function.

$$P(j|v) = \frac{freq(j,v)}{freq(v)} \tag{3.12}$$

$$P(n|v,j) = \frac{freq(n,v,j)}{freq(v,j)} \tag{3.13}$$

1. If $freq(n, v, j) > k$
$$P(n|v, j) = \frac{freq(n, v, j)}{freq(v, j)}$$
2. Else if $freq(n, v) + freq(n, j) > k$
$$P(n|v, j) = \frac{freq(n, v) + freq(n, j)}{freq(v) + freq(j)}$$
3. Else
$$P(n|v, j) = \begin{cases} 1.0 & \text{if } j = \text{NOMINATIVE} \\ 0.0 & \text{Otherwise} \end{cases}$$

Figure 3.3: *Back-off strategy for probability estimation*

Unfortunately, however, the above estimates could be useless due to the sparse data problems. A particular combination of features appearing in test data might never be seen in training data and then it will not be possible to estimate the probability for the combination. To prevent this unpleasant problem, we use the *back-off* smoothing.

In back-off smoothing (Katz, 1987), we move onto another count that has fewer variables recursively when we encounter a low frequency count. Following Collins and Brooks (1995) and de Lima (1997), we use the back-off strategy illustrated in Figure 3.3 for all of the discrete case decision models. The step 3 in Figure 3.3 is our default case decision, the NOMINATIVE case which is the most frequently used case.

The count combination method in Figure 3.3, although it is rather ad-hoc, works quite well in practice. It should be also noted that without discounting, the sum of the probabilities estimated by using the back-off smooting shown in Figure 3.3 will not be 1. However, because we are only interested in picking up the case particle which makes the highest probability value, we do not need to worry about the accuracy of the probability value. The constant k is a cut-off frequency for a back-off stage and normally set to 0 or 1 (Manning and Schütze, 1999).

3.3.2 Sequential Case Decision

To model the sequential case decision, we adopt a Markov chain tagging model which was applied to a similar task in other languages (Brants et al., 1997). In a Markov chain tagging model, a sequence of tagging events is considered as a Markov chain which has the following properties (Manning and Schütze, 1999):

- Limited horizon: $P(X_{i+1} = t^j|X_1, \ldots, X_i) = P(X_{i+1} = t^j|X_i)$

- Time invariant (stationary): $P(X_{i+1} = t^j | X_i) = P(X_2 = t^j | X_1)$

In our case, we assume that the case particle of a nominal only depends on the previous case particle (limited horizon) and this dependency does not change over time (time invariant). Since a case decision is only dependent on the previous case decision, long-distance relationships cannot be modelled.

More specifically, we represent a sequential case decision (SCD) as an joint event of a predicate v, a sequence of nominals $N = \langle n_1, \ldots, n_n \rangle$, and a sequence of case particles $J = \langle j_1, \ldots, j_n \rangle$ as (3.14). In our model, the sequences work backwords from the predicate. (3.14) is factored out as (3.15) and simplified as (3.16) by omitting $P(v)$ and making an independence assumption that the predicate v and the case particle sequence J are mutually independent.

$$SCD(N, v) = \arg\max_J P(v, J, N) \tag{3.14}$$

$$= \arg\max_J P(v) P(J|v) P(N|J, v) \tag{3.15}$$

$$= \arg\max_J P(J) P(N|J, v) \tag{3.16}$$

To reduce the parameters of (3.16), we make the following two assumptions about nominals in addition to the limited horizon assumption.

- Nominals are independent of each other, and
- A nominal's identity only depends on its case particle and the predicate.

Finally, we get (3.17) as a sequential case decision model based on a Markov chain tagging model SCD.

$$SCD(N, v) = \arg\max_{j_{1,n}} \prod_i P(n_i | j_i, v) P(j_i | j_{i-1}) \tag{3.17}$$

Probabilities are estimated using the deleted interpolation (Jelinek and Mercer, 1980) which linearly combines multiple probability estimates as shown in (3.18) and (3.19).

$$P(n_i | j_i, v) = \lambda_1 P(n_i | j_i, v) + \lambda_2 P(n_i | j_i) + \lambda_3 P(n_i | v) + \lambda_4 P(n) \tag{3.18}$$

$$P(j_i | j_{i-1}) = \mu_1 P(j_i, j_{i-1}) + \mu_2 P(j_i) \tag{3.19}$$

$$\text{where, } \sum_i \lambda_i = \sum_i \mu_i = 1$$

The weights are determined by the Expectation Maximisation (EM) algorithm (Dempster et al., 1977; Jelinek and Mercer, 1980).

The initial case decision is made using the simplest discrete case decision model which uses the predicate and the focus nomal as features. To choose the best case particle sequence out of all possible case particle sequences, we use the well-known Viterbi algorithm (Viterbi, 1967). If we have unambiguous case particles in a sequence, the search space is greatly reduced.

3.4 Knowledge-Lean Data Collection

The statistical modelling methods introduced in Section 3.3 require a considerable amount of training data consisting of training examples. To construct the training data, we need to collect a set of what we call *case decision instances* (*CDI*s). Each case decision instance contains nominals with their case particles and a predicate which is associated with the nominals and the case particles.[9] Thus, an instance of case decision is an approximation of a sentence. Consider the following example.

(65) a. Eoje jeonyeog-e-do Hwanho-ga Seho-ege-neun gom inhyeong-eul
Yesterday evening-LOC-also Hwanho-NOM Seho-DAT-TOP bear doll-ACC
ju-eoss-da.
give-PST-DCL
'As for Seho, Hwanho gave him a teddy bear yesterday evening'

b. (jeoneyog, -e, Hwanho, -ga, Seho, -ege, inhyeong, -eul, ju-)
(evening, LOC, Hwanho, NOM, Seho, DAT, doll, ACC, give)

(66) a. Eoje jeonyeog-∅ Hwanho-neun Seho-ege-man gom inhyeong-eul
Yesterday evening-∅ Hwanho-TOP Seho-DAT-TOP bear doll-ACC
ju-eoss-da.
give-PST-DCL
'As for Hwanho, he gave a teddy bear only to Seho yesterday evening'

b. (Seho, -ege, inhyeong, -eul, ju-)
(Seho, DAT, doll, ACC, give)

In sentence (65a), all the nominals used as arguments or adjuncts of the predicate *ju-* 'give' are accompanied by case particles. We can extract a case decision instance (65b) from the sentence. By constrast, only two nominals are accompanied by case particles in (66a), and (66b), which is incomplete, is the case decision instance extracted from the sentence.

[9] In principle, other words such as adverbs could helpful for case ambiguity resolution. However, we exclude them for the current work.

If we have fully annotated language resources, typically syntactically analysed corpora (treebanks), we do not need to worry too much about the incomplete case decision instances like (66b).[10] As described in 3.2.3 and 3.2.4, currently available Korean treebanks have none or only partial grammatical function encodings. Consequently, we have to use an unannotated corpus and collect required data from the corpus for the moment. Our hope is that the incomplete case decision instances would be still useful for the case ambiguity resolution.

It is relatively easy to extract case marking instances from simple sentences as shown in (65a) and (66a). However, it is not a trivial job to automatically process *mixed sentences*, in which two or more predicates are present. We have to face the *noun phrase attachment ambiguity*, a situation in which a noun phrase can be associated with two or more predicates, in mixed sentences as depicted in (67).

(67) a. na-neun mangwongyeong-eulo golae-deul-i idongha-neun geos-eul bo-ass-da.
Na-neun mangwongyeong-eulo golae-deul-i idongha-neun geos-eul
I-TOP telescope-INST whale-PL-NOM move-ADN thing-ACC
bo-ass-da.
see-PST-dcl
'I saw the whales moving with a telescope.'

b. (mangwongyeong, -eulo, golae-PL, -i, idongha-)
(telescope, INST, whales, NOM, move)
(mangwongyeong, -eulo, geos, -eul, bo-)
(telescope, INST, thing, ACC, see)

The standard way of overcoming the attachment ambiguity problem would be using a parser. However, at the time of writing, we are not aware of any existence of publicly released robust parser for the Korean language which can be used in relatively large scale projects such as the current work. There are several experimental parsers reported in literature (e.g., Lee et al. 1997d, Seo et al. 1999, Cha et al. 2002, Chung and Rim 2004). These parsers are typically built on small knowledge-base and/or trained on a small-size training data. Hence, it is very unlikely that they can cope with large-scale real-world data.[11]

[10] However, we can still get incomplete case decision instances as argument dropping is very common in Korean.

[11] We don't have any concrete large-scale parsing experiment results. When the mixed sentence example (67) was fed into three different parsing demonstration systems on the Internet (`http://nlp.kookmin.ac.kr/cgi-bin/parse.cgi`, `http://isoft.postech.ac.kr/Research/POSPAR20/demoframe.html`, `http://nlp2.korea.ac.kr/~hjchung/parserdemo/`), two of them returned wrong parsing results.

One possible solution to avoid the attachment ambiguity is using only simple sentences. It is, in every way, not a realistic solution because simple sentences are quite rare in a naturally occurring text. Besides, even if we have a vast amount of simple sentences, it does not automatically guarantee that we can get a reliable set of case decision instances.

Another option is using a system which can minimise the noun phrase attachment ambiguity rather than a full parser. An example of such system is a *clause segmentation* system. A *clause* is "a grammatical unit that includes, at the minimum, a predicate and an explicit or implied subject, and expresses a proposition." (Loos et al., 1997) Thus, if we can segment clauses from a mixed sentence, the noun phrase attachment problem can be avoided.

There have been a few attempts to build clause segmentation systems for Korean (Kim et al., 1993; Kim, 1996a; Lee et al., 1997a; Park, 2000). These systems all depend on rich linguistic knowledge such as subcategorisation frames and semantic hierarchies for nominals making them hard to scale up being hampered by the typical knowledge-bottle-neck problems.

An alternative approach to clause segmentation is a learning approach (Carreras and Màrquez, 2001; Déjean, 2001; Hammerton, 2001; Molina and Pla, 2001; Patrick and Goyal, 2001; Kim Sang., 2001; Hachey, 2002). This approach is very attractive to us, as the only requirement for this approach is an appropriate treebank. Nevertheless, there are still a couple of obstacles if we want to apply a learning approach to clause segmentation in Korean. First, as noted in section 3.2.3, distinguishing clauses from verb phrases is not trivial in Korean and this difficulty is reflected in the annotation schemes of the two treebanks we use. Thus if we want to build a clause recognition learner, we have to reannotate the treebanks to prepare the training data for the learner. Second, even if we successfully provide the training data, actual modelling and training work would consume a considerable amount of time and effort, which we cannot afford in current work.

The remaining option, which is our approach, is not to be worried too much about the noun phrase ambiguity and find a way of using a large number of mixed and simple sentences in a very simple and knowledge-lean manner.

Our clause segmentation method is based on the following observation: Since Korean is a head-final and right-branching language, noun phrase attachment ambiguities always occur on the left side of a predicate. Although certain types of embedded clause can move into another clause causing the attachment ambiguities, there are many occurrences of embedded clauses that keep their original positions in the sentences they belong to. Thus if we are lucky enough we can still get correct attachment decisions yielding valid case decision instances even if our heuristic is not very smart.

	CLS/*CDI* in a parse tree	*CLS*/*CDI* not in a parse tree
Suggested *CLS*/*CDI*	TP	FP
Not suggested *CLS*/*CDI*	FN	TN

Table 3.2: *Definition of true and false positives/negatives*

The case decision instances obtained by our heuristic method will not be completely accurate since wrong attachment decisions will be made by the method. In addition, it is not ensured that the data will give us enough information required for case ambiguity resolution as the data is gathered from unambiguous instances. For example, the data constructed from unambiguous instances do not tell us about the auxiliary particle preference of a particular case particle, which could be useful for case ambiguity resolution. A quick look at the Sejong Treebank reveals that the nominative case particle is frequently replaced by a topic marker. Our hope is that imperfect but abundant training data will eventually contribute quite meaningfully toward our task (Ratnaparkhi, 1998).

A detailed description of our clause segmentation and case decision instance collection methods and their evaluation are given in Chapter 4.

3.5 Evaluation

In this section, we introduce the evaluation measures that we use to evaluate our data collection methods and case ambiguity resolution system.

3.5.1 Precision, Recall and F-measure

For the evaluation of the clause segmentation and the case decision instance extraction procedures, we apply our procedures to the KAIST Treebank and the Sejong Treebank and attempt to recover the clauses and case marking instances in the parse trees. To measure the performance, we use *precision* (P) and *recall* (R). Precision and recall are defined in terms of the number of *true* and *false positives* (TP and FP, respectively) and *true* and *false negatives* (TN and FN, respectively). For clause (*CLS*) segmentation and case decision instance (*CDI*) extraction procedures, these quantities are defined as Table 3.2

Precision measures the proportion of the correct suggested objects amongst all suggested objects. It is defined as the number of true positives (TP) divided by the sum of true positives (TP) and false positives (FP).

Recall measures the proportion of the correct suggested objects among all standard objects.

It is defined as the number of true positives (TP) divided by the sum of true positives (TP) and false negatives (FN).

(3.20) and (3.21) are formal definitions of precision and recall.

$$Precision = \frac{TP}{TP + FP} \tag{3.20}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.21}$$

We also use precision and recall to measure the performances of statistical case ambiguity resolution models. In this case, precision and recall are calculated for each target case particle for each annotation. For this calculation, it is helpful to understand the two measures as following:

$$Precision = \frac{RetRel}{Ret} \tag{3.22}$$

$$Recall = \frac{RetRel}{Rel} \tag{3.23}$$

where

- Ret is the set of all case particles the system has returned for test instances annotated as instances of one of the target case particle.
- Rel is the set of annotations for a specific target case particle.
- $RetRel$ is the set of case particles that agree with the annotations for a specific target case particle.

Precision and recall measures usually show a trade-off between them. Thus when we compare the performances of multiple procedures, it is desirable to have a single measure which combines precision and recall. This combined measure is the *F-measure* (van Rijsbergen, 1979). F-measure is calculated as (3.24).

$$F_\beta = \frac{(\beta^2 + 1) \cdot P \cdot R}{\beta^2 \cdot P + R} \tag{3.24}$$

The parameter β is a weight which determines the relative importance of precision and recall. We set β as 1 to give no preference to either precision or recall. As the result, we have the following F-measure formula, which is the harmonic mean of precision and recall.

$$F = \frac{2PR}{P + R} \tag{3.25}$$

Precision, recall and F-measure are frequently used for the evaluation of Information Extraction systems. It has also been applied to the evaluation of a vast number of NLP tasks. Clause recognition (Carreras and Màrquez, 2001; Hachey, 2002) and grammatical relation finding (Buchholz, 2002) are some of such tasks related to current work.

3.5.2 The Kappa Statistic

As mentioned in 3.3.1, our task of case ambiguity resolution is a classification task which involves assigning mutually exclusive categorial judgements to given questions. To evaluate the performance of our system, we need to apply our system to a test set and compare the output of the system with a gold standard. Since we don't have a ready-made publicly released test set for our task, we have to rely on a human annotation. The pitfall is that we cannot exclude the possibility of getting agreement by chance when we compare the output of the system and the human annotation.

To measure the agreement between the multiple human annotation and the output of our system, we use the Kappa Coefficient.[12] The Kappa Coefficient is the proportion of agreement corrected for chance between two judges assigning cases to a set of categorial assignment as shown in (3.26) (Cohen, 1960). The Kappa Coefficient K is computed as (3.26), where $P(A)$ is the observed agreement among the annotators, and $P(E)$ is the expected agreement representing the agreement by chance.

$$K = \frac{P(A) - P(E)}{1 - P(E)} \tag{3.26}$$

The value of K ranges from -1 to 1. If $K = 1$, then annotators have a perfect agreement. If $K = 0$, then the agreement is equal to chance. If annotators perfectly disagree, then $K = -1$.

In accessing the K value, Landis and Koch (1977) proposed the following scales in the context of a bio-medical study: $.00 \leq K \leq .20$ is slight, $.21 \leq K \leq .40$ is fair, $.41 \leq K \leq .60$ is moderate, $0.61 \leq K \leq 0.80$ is substantial, and $0.81 \leq K \leq 1.00$ is almost perfect. Krippendorff (1980) gives a different assessment of K values drawn from his and his colleagues' content analysis work: discount when $K < .67$, allow tentative conclusions when $.67 \leq K < .8$, and definite conclusions when $K \geq .8$. However, these assessment scales should be considered only as a plausible standard (Carletta et al., 1997).

There are two main ways of computing the expected agreement $P(E)$: The method presented in Siegel and Castellan (1988) assumes that the distribution of proportions over the categories are equal for annotators. On the other hand, the method in Cohen (1960) does

[12] This section is heavily indebted to Lapata (2001) and Eugenio and Glass (2004).

not have such an assumption (Eugenio and Glass, 2004). We use the second method since it is hard to expect that the distribution of case particles is equal for all annotators.

Since being brought to attention in computational linguistics and natural language processing community by Carletta (1996), the Kappa coefficient is the de facto standard to assess inter-annotator agreement (Eugenio and Glass, 2004). It has also been used for the evaluation of NLP systems. Teufel (2000) uses Kappa for the evaluation of a summarisation system. Stevenson and Merlo (2000) assesses the agreement between the output of a system and a human judgement on a task of semantic classification of verbs. Lapata (2001) also uses Kappa to evaluate the performance of a series of automatic lexical/semantic classification procedures.

3.6 Summary

We have laid out the methodological foundations for the task of case ambiguity resolution in Korean pursued in this thesis. We introduced the corpora we use for the training data and test data collection together with our choice of data collection method. We also described the statistical models for case ambiguity resolution. Finally we presented the methodology for the evaluation of our data collection methods and statistical models.

Chapter 4

Data Preparation and Experimental Setup

This chapter describes the training data construction process and various experimental setups including the test set and the performance bounds. In Section 4.1, the details of the individual subprocesses of the training data construction process are described and the evaluation result for the heuristic data collection method is presented. Section 4.2 begins with the test set construction and analysis and establishes the performance bounds. Finally, Section 4.3 summarises this chapter.

4.1 Training Data Construction

This section describes the individual components of the training data construction process illustrated in Figure 4.1. The input of the whole process is the Yonsei and the Sejong raw corpora and the Sejong part-of-speech tagged corpora introduced in Chapter 3. The output of the training data construction processes is a set of case decision instances, which is used as the training material for our statistical case decision models.

4.1.1 Sentence Splitting

As noted in Section 3.2.2, the raw corpora part of the Sejong Corpora does not have sentence boundary markings. We apply a very simple sentence splitting procedure to the corpora to get a sentence-splitted version of the corpora. This procedure only relies on a small set of rules to recognise sentence boundaries.

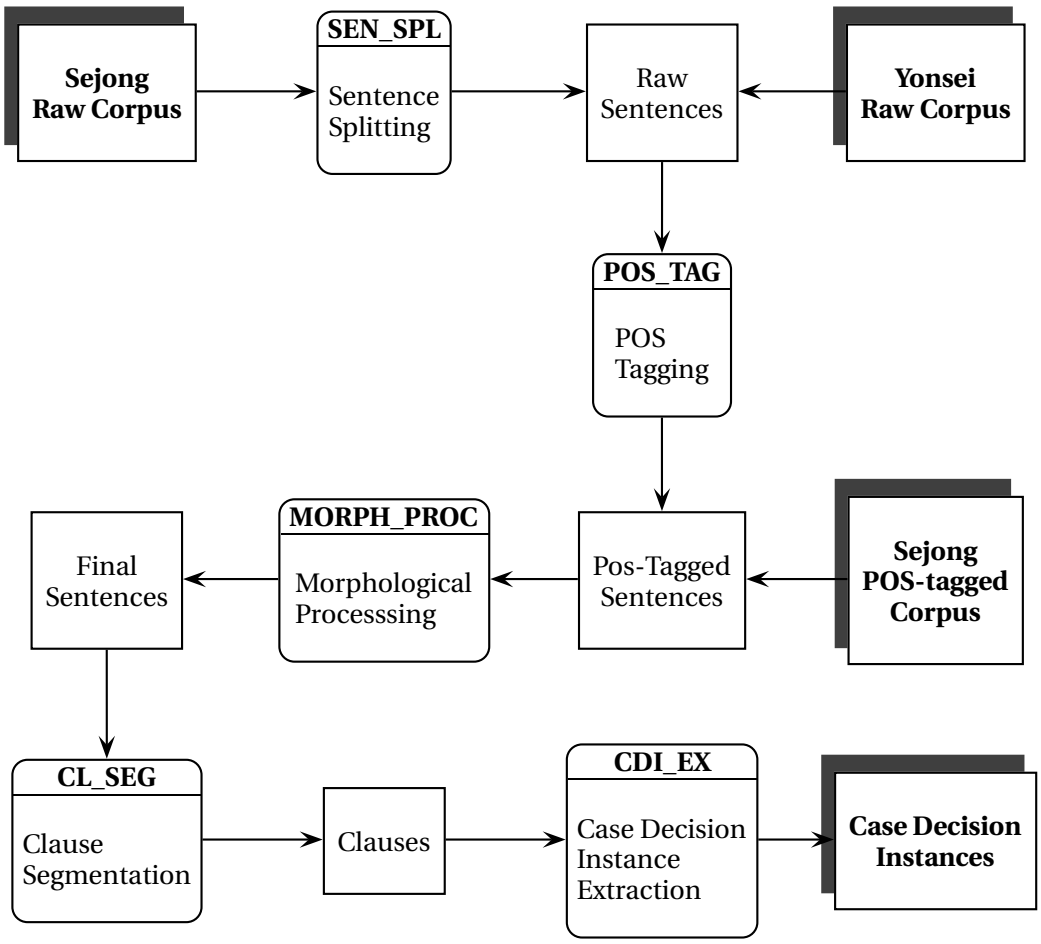


Figure 4.1: *Data flow diagram for the training data construction process*

Input:	Eoneu nal, sonyeon-eun gil ilh-eun hui-n mangaji-leul deli-go jib-eulo o-ass-seubnida. one day, boy-TOP way lose-ADN white-ADN foal-ACC bring-COCON home-LOC come-PST-DCL 'One day, the boy brought home a white lost foal.'
Output:	eoneu/MM nal/NNG ,/SP sonyeon/NNG+eun/JX gil/NNG ilh/VV+eun/ETM hui/VA+eun/ETM mangaji/NNG+leul/JKO deli/VV+go/EC jib/NNG+eulo/JKB o/VV+ass/EP+seubnida/EF ./SF

Figure 4.2: *An example input/output of the Sejong part-of-speech tagger*

4.1.2 Part-of-Speech Tagging

To produce a part-of-speech tagged corpus, we use the Sejong Tagger which is supplied with the Sejong Corpus. This tagger takes a sentence-divided corpus as it's input and returns a part-of-speech tagged corpus as shown in Figure 4.2.

In Figure 4.2, nouns *mangaji* 'foal' and *jib* 'home' are accompanied by case particles *-leul* ACCUSATIVE and *-eulo* DIRECTIONAL whereas nouns *nal* 'day', *sonyeon* 'boy' and *gil* 'way' occur without any case particles. The latter set of nouns are the nominal words that get our attention.

The tagger uses a Korean tagset composed of 47 tags (Kim et al., 2000).[1] The reported tagging accuracy is 94%.

4.1.3 Morphological Processing

The tagger's output undoubtedly contains various tagging and morphological analysis errors. It is impossible to track down and take care of all the errors. Nonetheless, we decided to correct some obvious morphological analysis errors reported in Cho (2002).

The level of the Sejong tagger/morphological analyser's morphological analysis goes down to pre-lexical level and the tagger splits derivational suffixes from their roots. We performed a morphological process to merge these over-segmented morphemes since we are only interested in lexical level information.

Figure 4.3 shows some examples of the tagging error correction and the morphological processing.

[1] See Appendix C for the full list of the Sejong tagset.

Morphological analysis error correction	*/VV+deusi/NNB ⇒ */VV+deusi/EC
	geuleus/NNB ⇒ geuleus/NNG
	maeum/NNG+daelo/JX ⇒ maeumdaelo/MAG
	hamgge/NNG ⇒ hamgge/MAG
Morphological processing (morpheme merging)	*/NNG+gan/XSN ⇒ *gan/NNG
	*/NNG+buti/XSN ⇒ *buti/NNG
	*/NNG+ha/XSV ⇒ *ha/VV
	*/NNG+lob/XSA ⇒ *lob/VA

Figure 4.3: *Examples of tagging error correction/morphological processing*

4.1.4 Clause Segmentation

As stated in Chapter 3, we are using a knowledge-lean method for the data collection and our clause segmentaion method is not an exception. Our clause segmentation method (ClSeg) segments a sentence into fragments considering the predicates in the sentence as delimiters. In other words, this method does not attempt to resolve any noun phrase attachment ambiguities and attaches noun phrases to the nearest right-side predicate. The segmentation result of this method will contain many false clauses. However, the chance of obtaining the correct clauses as often as possible is maximised. At the other extreme, we can think of a method which only takes the right most clause of a sentence discarding other parts of the sentence that may have noun phrase attachment ambiguities. The accuracy of the segmentation result of this method will be very high. However, this method requires a huge amount of raw material because it only uses a very small part of a sentence. Figure 4.4 shows the clause segmentaion results of our method (ClSeg) and the method which takes the right most clause (RMC).

To evaluate the clause segmentation method, we applied the method to the KAIST Treebank and the Sejong Treebank containing 12,084 and 13,174 sentences respectively (total 25,258 sentences). The output of the method was compared with the gold standard retrieved from the treebanks. The evaluation result is shown in Table 4.1. We also included the evaluation results for the RMC and SMP which takes only simple sentences from the treebanks for the comparison.

Not surprisingly, the precision of the method which uses simple sentences only is the highest. Conversely the recall is the lowest. Our method ClSeg, performs quite well with the best $F_{\beta=1}$ score (54.16). The method RMC's performance is not impressive at all. It failed to improve on the precision on this particular test data. We expect that our clause segmentation method which has reasonablly balanced precision and recall will cope well with the

Input Sentence	Nam sweteulandeugundo-neun namgeugbando-e pyeonghaengha-ge baldalha-n 20-yeogae-ui seom-eulo, namgeug-euloseo-neun gajang meonjeo 1819-nyeon-e balgyeondoe-eoss-go mulgae-wa golaejabi-ui geungeoji-ga doe-ess-da. south Shetland-isles-TOP south-antarctic-peninsula-LOC parallel-ADV develop-ADN 20-or-so-GEN island-FUNC, antarctic-FUNC-TOP most earlier 1819-year-LOC be found-PST-COCON seal-and whale-hunting-GEN base-NOM become-PST-DCL 'Southern Shetland isles, which are a group of 20 or so islands that developed parallel to antarctic peninsula, was found first in antarctic area in 1819 and became a base for seal and whale hunting.'
Correct segmentation	(south Shetland-isels-TOP 20-or-so-GEN island-FUNC seal-and whale-hunting-GEN base-NOM become-PST-DCL) + (south-antarctic-peninsula-LOC parallel-ADV) + (develop-ADN) + (antarctic-FUNC-TOP most earlier 1819-year-LOC be found-PST-COCON)
ClSeg output	(south Shetland-isles-TOP south-antarctic-peninsula-LOC parallel-ADV) + (develop-ADN) + (20-or-so-GEN island-FUNC, antarctic-FUNC-TOP most earlier 1819-year-LOC be found-PST-COCON) + (seal-and whale-hunting-GEN base-NOM become-PST-DCL)
RMC output	(seal-and whale-hunting-GEN base-NOM become-PST-DCL)

Figure 4.4: *Clause segmentation results for a sample sentence*

	True cls	Suggested cls	Correct cls	Precision	Recall	$F_{\beta=1}$
ClSeg	110,875	11,914	60,063	54.15	54.17	54.16
RMC	110,875	25,258	13,609	55.66	12.68	20.65
SMP	110,875	2,930	2,587	88.29	2.33	4.55

Table 4.1: *Evaluation result for clause segmentation methods*

case ambiguity resolution task.

4.1.5 Case Decision Instance Extraction

Once we have segmented clauses, we extract case decision instances from the clauses. It is a relatively simple and straight-forward procedure because these clauses are almost free from noun phrase attachment ambiguity. Therefore we can attach noun phrases to predicates with little difficulty in most cases as shown in (68). There are, however, a few issues that we have to deal with, which can be seen in (69)-(71).

(68) a. namamelika/NNP kkeut/NNG+eseo/JKB namjjog/NNG+eulo/JKB
south-America/NNP edge/NNG+LOC south+INST
naelyeoga/VV+daga/EC
go down/VV+SUBCON
'While go down to the south from the southern edge of the south America'

b. (kkeut, -eseo, namjjog, -eulo, naelyeoga)
(edge, LOC, south, DIR, go down)

(69) a. gag/MM geonmul/NNG+e/JKB+neun/JX gigyesil/NNG+i/JKS
each/MM building/NNG+LOC+TOP machine-room/NNG+NOM
iss/VA+eumyeo/EC
exist-COCON
'Each building has a machine room and'

b. (geonmul, -e, gigyesil, -i, iss-)
(building, LOC, machine room, NOM, exist)

(70) a. Kim/NNP bagsa/NNG+ege/JKB+lo/JKB+man/JX dabjang/NNG+eul/JKO
Kim/NNP doctor/NNG+DAT+DIR+only reply/NNG+ACC
bonae/VV+eoss/EP+da/EF+./SF
send/VV+PST+DCL+./SF
'(I) sent a reply only to Dr Kim.'

b. (dabjang, -eul , bonae-)
(reply, ACC, send)

(71) a. yejeong/NNG+gwa/JKB dalli/MAG chille/NNP+gonggun/NNG+ui/JKG
plan/NNG+COM differently/MAG Chile/NNP+air
susonggi/NNG+ga/JKS deul/VV+eo/EC+o/VX+a/EC
force/NNG+GEN carrier/NNG+NOM come in-SUBCON
'Contrary to the plan, a Chilean air force carrier comes in'

b. (susonggi, -ga, deuleoo-)
(carrier, NOM, come in)

(72) a. namgeug/NNP+ui/JKG bom/NNG+do/JX
antarctic/NNP+GEN spring/NNG+also
munmyeong/NNG+segye/NNG+ui/JKG bom/NNG+gwa/JKB
civilised/NNG+world/NNG+GEN spring/NNG+COM
gat/VA+aseo/EC
same/VA+SUBCON
'As the spring of antarctic is the same as the spring of the civilised world'

b. (bom, -gwa, gat-)
(spring, COM, same)

In (69), noun phrase *geonmul-e-neun* 'building-LOC-TOP' contains one case particle and one auxiliary particle. In this case, it is safe to take the case particle *-e* LOC only since the auxiliary particle does not affect the case marking as we examined in Chapter 2.

In contrast to (68) and (69), *bagsa-ege-lo-man* 'doctor-DAT-DIR-only' has three particles and two of them are case particles. We discard this type of noun phrases.[2]

There are two case particles which need special treatments: *-ui* GENITIVE and *-gwa/-wa* COMITATIVE. *-ui* GENITIVE does not relate a nominal and a predicate. It relates two nominals. Thus we also discard noun phrases with *-ui* GENITIVE as we do in (70) and (71).

A noun phrase which contains *-gwa/-wa* COMITATIVE can be attached to either an adverb or a predicate. In (70), *yejeong-gwa* 'plan-COM' should be attached to the adverb *dalli* 'differently' to get a proper interpretation. On the other hand, *bom-gwa* 'spring-COM' is attached to *gat-* 'same'. This attachment decision is heavily dependent on the lexical features of adverbs and predicates. In our approach, we attach a noun phrase with *-gwa/-wa* COMITATIVE to a predicate only when it is adjacent to the predicate.

As we have seen in Chapter 2, the particle *-gwa/-wa* has another usage as CONNECTIVE which connects two noun phrases. This kind of noun phrase is not taken for the same reason as *-ui* GENITIVE was not taken.

Last but not least, we only take the stem part of a predicate wordform, and we take the last nominal component of a compound nominal.

[2] This phenomenon is called *case particle stacking*. See Sohn 1999, p. 343.

	Cls	ClSeg	RmcO	SmpO
True *CDI*s	48,950	48,950	48,950	48,950
Suggested *CDI*s	49,207	50,144	6,312	1,271
Correct *CDI*s	47,472	35,104	4,797	1,063
Precision	96.47	70.01	76.00	83.63
Recall	96.98	71.71	9.80	2.17
$F_{\beta=1}$	96.73	70.85	17.36	4.23
True *MCDI*s	63,739	63,739	63,739	63,739
Suggested *MCDI*s	65,144	65,080	7,973	1,982
Correct *MCDI*s	63,399	52,390	7,555	1,707
Precision	97.32	80.50	94.76	86.13
Recall	99.47	82.19	11.85	2.68
$F_{\beta=1}$	98.38	81.34	21.07	5.19

Table 4.2: *Evaluation result for case decision instance extraction*

Our case decision instance extraction procedure (CdiEx), which satisfies all the conditions and restrictions, is formalised in Procedure 1.

Table 4.2 shows the evaluation result for the case decision instance extraction method. This method was applied to the outputs of the clause segmentation methods. It was also applied to the true clauses from the treebanks to measure the performance of the procedure in isolation. Extracted case decision instances (*CDI*s) and minimal case decision instances (*MCDI*)[3] are compared with the true case marking instances retrieved from the treebanks.

It was pretty much expected that the case decision instance extraction procedure itself would show good performances on both *CDI*s (96.47 precision, 96.98 recall, 96.73 $F_{\beta=1}$), and *MCDI*s (97.32 precision, 99.47 recall, 93.38 $F_{\beta=1}$). Naturally, the performances of each clause segmentation methods directly affect the whole case decision instance extraction results. Consequently, the case decision instance extraction method hits the best figures when it is applied to CLSEG. It scores 70.01 precision, 71.71 recall, 70.85 $F_{\beta=1}$ for *CDI*s and 80.50 precision, 82.19 recall, and 81.34 $F_{\beta=1}$ for *MCDI*s.

4.2 Experimental Setup

This section presents the results of the training data construction and examines the human annotated test set. The performance bounds of the case ambiguity resolution task are also

[3]A minimal case decision instance consists of a predicate, a nominal, and a case particle.

Procedure 1 *Case Decision Instance Extraction Procedure (CdiEx)*

$CL \leftarrow \langle w_1, w_2, \ldots w_n \rangle$ {Input clause}
$pred \leftarrow$ stem of w_n
$NP \leftarrow$ [] {Noun phrase list}
$i \leftarrow 1$
while $i < n - 1$ **do**
  if w_i is a nominal wordform **then**
    $\langle N, J \rangle \leftarrow w_i$ {Split a wordform into a nominal part and a particle part}
    if J contains multiple case particles or no case particle **then**
      continue
    end if
    $j \leftarrow$ case particle in J
    if N is a compound nominal **then**
      $n \leftarrow$ last nominal of N
    end if
    if $j =$ GEN **then**
      continue
    end if
    if $j =$ COM and $i + 1 = n$ **then**
      append $\langle n, j \rangle$ to NP
    else
      append $\langle n, j \rangle$ to NP
    end if
  end if
  $i \leftarrow i + 1$
end while
return $CP + \langle pred \rangle$

Corpus	Words	Sentences	Clauses	*CDI*s	*MCDI*s
YSC-1	2,880,000	155,766	637,559	288,958	389,033
YSC-2	1,100,000	70,079	343,684	164,376	217,572
YSC-3	5,900,000	568,481	2,157,885	926,720	1,206,708
YSC-5	8,620,000	821,733	3,020,233	1,242,337	1,602,328
YSC-6	7,256,000	710,579	2,627,260	1,134,179	1,464,548
YSC-7	13,710,000	975,974	3,076,504	1,309,417	1,677,342
YSC-8	898,000	83,274	270,659	130,091	164,888
YSC-9	1,499,000	223,978	596,601	244,359	304,738
SJC-1	10,000,000	675,653	2,929,563	1,349,500	1,770,437
SJC-2	7,000,000	508,616	2,140,134	1,036,825	1,373,980
SJC-3	2,000,000	173,680	745,049	337,250	433,059
Total	60,863,000	4,967,813	18,545,131	8,164,012	10,604,633

Table 4.3: *The result of the training data construction*

suggested.

4.2.1 The Training Set

We applied the whole training data construction process described in Section 4.1 to a 60,900,000-word corpus which is originated from 11 sub-corpora. The result is summarised in Table 4.3.

As a whole, 18,545,131 clauses are segmented out and 8,164,012 *CDI*s and 10,604,633 *MCDI*s are extracted from the clauses. The training data was further divided into ten sub-training sets that contain approximately equal numer of *CDI*s. These sub-sets are used in measuring the performance of the system in regards to the number of training examples.

We counted the unique number of features with regards to the number of *MCDI*s using the ten sub-sets. Table 4.4 shows the counts for the single features n, and v and the combined features $\langle j, v, n\rangle$, $\langle v, n\rangle$, $\langle j, v\rangle$, and $\langle j, n\rangle$. As shown in the table, the numbers of unique features keep increasing.

4.2.2 The Test Set

To evaluate the performance of our case ambiguity resolution system, we apply the system to the test set and compare the output with multiple human annotations. The test set

Feature	0.8M		1.6M		2.4M		3.2M		4.0M	
	Num	Inc	Num	Inc	Num	Inc	Num	Inc	Num	Inc
n	40,439	0	55,540	15,101	64,809	9,269	72,057	7,248	78,814	6,757
v	16,717	0	22,571	5,854	26,680	4,109	29,796	3,116	33,193	3,397
j, *v*, *n*	563,022	0	1,003,066	440,044	1,379,150	376,084	1,723,481	344,331	2,039,595	316,114
v, *n*	495,409	0	863,983	368,574	1,171,942	307,959	1,451,044	279,102	1,704,226	253,182
j, *v*	58,275	0	81,310	23,035	97,058	15,748	109,350	12,292	120,648	11,298
j, *n*	94,106	0	138,622	44,516	168,935	30,313	193,286	24,351	214,949	2,1663

Feature	4.8M		5.6M		6.4M		7.2M		8.0M	
	Num	Inc	Num	Inc	Num	Inc	Num	Inc	Num	Inc
n	85,099	6,285	90,509	5,410	95,671	5,162	100,222	4,551	103,997	3,775
v	36,774	3,581	39,475	2,701	42,264	2,789	44,653	2,389	46,790	2,137
j, *v*, *n*	2,353,582	313,987	2,648,306	294,724	2,919,876	271,570	3,189,054	269,178	3,440,295	251,241
v, *n*	1,953,739	249,513	2,185,384	231,645	2,397,437	212,053	2,606,274	208,837	2,801,402	195,128
j, *v*	131,988	11,340	141,076	9,088	149,665	8,589	157,654	7,989	164,476	6,822
j, *n*	235,466	20,517	253,471	18,005	269,796	16,325	284,686	14,890	297,630	12,944

Table 4.4: *The increases of the unique features with regards to the number of training examples*

was extracted from the treebanks to allow annotators solely to concentrate on case decision tasks without worrying about the attachment ambiguities.

From the KAIST Treebank and the Sejong Treebank, 500 sentences that have at least one ambiguous instance per each sentence were randomly selected. Each sentence has approximately 1.6 ambiguous instances and the total number of ambiguous instances is 794.

We prepared two sets of annotation material. In the first set (full context), sentences were presented retaining their original form. In the second set (limited context), sentences were presented being edited only with limited contexts. Consider the following examples.

(73) Full context

Silche-neun ( ) **ijhyeoji-go** ***gagyeg sangseung-man*** ( ) munje-ga
substance-TOP ( ) be forgotten-COCON price rise-only ( ) problem-NOM
doe-eoss-da.
become-PST-DCL
'The substance is forgotten and the price rise alone becomes a problem.'

(74) Limited context

Silche ( ) ijhyeoji-da.
Substance ( ) be forgotten-DCL
'The substance is forgotten.'

Sangseung ( ) □-ga doe-da.
Rise ( ) □-NOM become-dcl
'The rise becomes (something).'

In (73), annotators are requested to choose the missing case particles in places marked

Measure	FullContext_1:FullContext_2	FullContext_1:FullContext_3	FullContext_2:FullContext_3	Average
Agreement	95.84	94.96	95.21	95.34
Kappa	0.92	0.90	0.91	0.91
Measure	LimContext_1:LimContext_2	LimContext_1:LimContext_3	LimContext_2:LimContext_3	Average
Agreement	83.88	85.39	82.62	83.96
Kappa	0.73	0.75	0.71	0.73

Table 4.5: *Pairwise agreement of the human annotations*

Annotation	All agree		Some agree		All disagree	
	Num	%	Num	%	Num	%
Full context	739	93.073	54	6.801	1	0.126
Limited context	609	76.700	173	21.789	12	1.514

Table 4.6: *Distribution of agreement patterns*

by pairs of brackets. Ambiguous instances and associated predicates are identified by the typefaces.[4] In (74), each clause is explicitly splitted out and auxiliary particles are removed from the ambiguous instances if there are any. Information from the neighbouring words is also reduced. For example, only the neighbouring case particle is provided without the nominal in the second clause of (74).[5]

Six human judges (three for each type) annotated the test material. The pairwise agreements of the annotation results are shown in Table 4.5. The pairwise agreements between the full context annotation results are very high. The average agreement is 95.34% and the average $Kappa$ is 0.91. This $Kappa$ value belongs to scales of 'almost perfect' (Landis and Koch, 1977) and 'definite conclusions' (Krippendorff, 1980). On the other hand, the pairwise agreements between the limited context annotation results are lower. The average agreement is 83.96% and the average $Kappa$ is 0.73. This $Kappa$ value is still in 'substantial' and 'tentative conclusions' scales. From the pairwise agreements, it is confirmed that the full context plays an important role in case decision task for human judges. The case decisions given by the human judges with only limited contextual information tend to be arbitrarily distributed and this tendency is reflected on the agreement measures. Table 4.6 also supports this fact. As a whole, all human judges rarely disagreed.

We also measured the pairwise agreement across the two types of test material as shown in Table 4.9. Surprisingly, the average pairwise agreement measures 84.95% and 0.74 are higher than those of the limited context annotation results. However, if we examine the figures in the table, we discover that a particular annotation result LimContext_1 has unusu-

[4] In practice, we used different colours to display noun phrase attachments.

[5] The full context annotation material is provided in Appendix D.

(a) X: FullContext$_1$, Y: FullContext$_2$ (Agree 95.84%, Kappa 0.92)

	NOM	ACC	LOC	DAT	INST	COM	Sum
NOM	515	4	0	0	0	0	519
ACC	3	90	1	0	2	0	96
LOC	5	4	130	0	1	0	140
DAT	3	0	0	0	0	0	3
INST	3	2	4	0	10	0	19
COM	1	0	0	0	0	16	17
Sum	530	100	135	0	13	16	794

(b) X: FullContext$_1$, Y: FullContext$_3$ (Agree 94.96%, Kappa 0.90)

	NOM	ACC	LOC	DAT	INST	COM	Sum
NOM	505	4	0	0	1	0	510
ACC	5	89	0	0	1	0	95
LOC	10	4	134	0	1	0	149
DAT	6	0	0	0	0	0	6
INST	4	3	1	0	10	0	18
COM	0	0	0	0	0	16	16
Sum	530	100	135	0	13	16	794

(c) X: FullContext$_2$, Y: FullContext$_3$ (Agree 95.21%, Kappa 0.91)

	NOM	ACC	LOC	DAT	INST	COM	Sum
NOM	501	4	2	1	1	1	510
ACC	5	88	2	0	0	0	95
LOC	7	3	135	0	4	0	149
DAT	4	0	0	2	0	0	6
INST	2	1	1	0	14	0	18
COM	0	0	0	0	0	16	16
Sum	519	96	140	3	19	17	794

Table 4.7: *Pairwise confusion matrices for full context human annotations*

(a) X: LimContext$_1$, Y: LimContext$_2$ (Agree 83.88%, Kappa 0.73)

	NOM	ACC	LOC	DAT	INST	COM	Sum
NOM	417	9	16	1	7	1	451
ACC	12	101	8	0	3	0	124
LOC	14	6	114	0	0	0	134
DAT	18	0	0	1	6	0	19
INST	19	2	10	0	15	0	46
COM	2	0	0	0	0	18	20
Sum	482	118	148	2	25	19	794

(b) X: LimContext$_1$, Y: LimContext$_3$ (Agree 85.39%, Kappa 0.75)

	NOM	ACC	LOC	DAT	INST	COM	Sum
NOM	433	13	18	1	9	2	476
ACC	12	95	4	0	1	0	112
LOC	17	8	121	0	2	0	148
DAT	9	0	0	1	0	1	11
INST	9	1	3	0	12	0	25
COM	2	1	2	0	1	16	22
Sum	482	118	148	2	25	19	794

(c) X: LimContext$_2$, Y: LimContext$_3$ (Agree 82.62%, Kappa 0.71)

	NOM	ACC	LOC	DAT	INST	COM	Sum
NOM	413	12	18	14	15	4	476
ACC	7	99	4	0	2	0	112
LOC	16	11	109	0	12	0	148
DAT	7	0	0	4	0	0	11
INST	7	2	1	0	15	0	25
COM	1	0	2	1	2	16	22
Sum	451	124	134	19	46	20	794

Table 4.8: *Pairwise confusion matrices for limited context human annotations*

Pair	Agreement	*Kappa*
FullContext$_1$:LimContext$_1$	88.54	0.80
FullContext$_1$:LimContext$_2$	81.36	0.67
FullContext$_1$:LimContext$_3$	86.27	0.75
FullContext$_2$:LimContext$_1$	87.15	0.77
FullContext$_2$:LimContext$_2$	81.86	0.69
FullContext$_2$:LimContext$_3$	86.02	0.75
FullContext$_3$:LimContext$_1$	86.27	0.75
FullContext$_3$:LimContext$_2$	81.86	0.69
FullContext$_3$:LimContext$_3$	85.26	0.74
Average	84.95	0.74

Table 4.9: *Pairwise agreement of human annotations across the context types*

ally high agreement with all three full context annotation results. If we remove the pairwise agreement measures involving the LimContext$_1$, the average agreement measures come down to 83.77% and 0.72 that are similar to the average pairwise agreements between the limited context annotation results.

In Tables 4.7 and 4.8, the pairwise confusion matrices are given. It is impossible to draw any concrete conclusion from these small number of matrices. At least, we can reason that the case particle alternation phenomenon described in Section 2.3.4 is reflected on the frequent confusions between the NOMINATIVE case particle *-i/-ga* and other case particles. According to the matrices, the DATIVE case particle *-ege* is the most confused case particle, and the COMITATIVE case particle *-gwa/-wa* is the least confused case particle.

4.2.3 Performance Bounds

4.2.3.1 Baselines

To draw the lower bound of the case ambiguity resolution system, we establish the following three baselines.

(75) Baselines

a. Always choose the NOMINATIVE case particle *-i/-ga*.

b. Choose the most probable case particle for the distance of the focus nominal from the predicate of the sentence (f=*-i/-ga* NOM, m=*-i/-ga* NOM, n=*-eul/-leul* ACC).

(a) Baseline performances on the full context annotations

Baseline	FullContext_1		FullContext_2		FullContext_3		AVG	
	Agr	K	Agr	K	Agr	K	Agr	K
Always choose NOM	66.75	0.00	65.37	0.00	64.23	0.00	65.45	0.00
f= NOM, m = NOM, n = ACC	45.47	0.10	43.95	0.08	42.95	0.08	44.12	0.09
f= NOM, m = LOC, n = ACC	42.82	0.11	41.81	0.11	40.68	0.10	41.78	0.10

(b) Baseline performances on the limited context annotations

Baseline	LimContext_1		LimContext_2		LimContext_3		AVG	
	Agr	K	Agr	K	Agr	K	Agr	K
Always choose NOM	66.25	0.00	56.80	0.00	59.95	0.00	60.70	0.00
f= NOM, m = NOM, n = ACC	46.22	0.11	40.93	0.07	40.30	0.05	42.49	0.08
f= NOM, m = LOC, n = ACC	43.32	0.12	38.80	0.08	38.29	0.07	40.13	0.09

Table 4.10: *Some possible baselines and their performances*

c. Choose a case particle reflecting the canonical word order SOV. (f=*-i/-ga* NOM, m=*-e* LOC, n=*-eul/-leul* ACC).

The baseline (75a) always chooses the most frequently used case particle *-i/-ga* NOMINATIVE. Naturally, the agreement measures in percentage for this baseline (65.45% and 60.70% in average) are the highest while the $Kappa$ values (0.00) are the lowest which is equivalent to the chance agreement. The baseline (75b) chooses the most probable case particle in regards to the distance of the focus nominal from the predicate. The average $Kappa$ values are improved up to 0.09 and 0.08. The canonical word order in Korean (SOV) is reflected in baseline (75c). The average $Kappa$ values reach 0.10 and 0.09.

We have also tried to set other baselines. However none of them outperformed the baseline 3 in terms of the average $Kappa$ values.

4.2.3.2 Upper Bounds

Establishing upper bounds are harder than establishing baselines. We don't expect our case ambiguity resolving system to perform better than a human with or without the full contextual information. Thus, the approximate upper bound would be the average pairwise agreement of 83.96% and 0.73 with the limited context annotations and 95.34% and 0.91 with the full context annotations.

4.3 Summary

This chapter presented the details of the individual procedures of the training data construction process based on simple language processing techniques and knowledge-lean data collection methods. We presented the evaluation result for the data collection methods tested on the treebanks.

The second part of the chapter concentrated on experimental setups. It showed the training and the test data construction results and analysed the human annotations. The upper and lower performance bounds were also suggested.

Chapter 5

Statistical Case Ambiguity Resolution in Korean

This chapter presents the experimental results for our approach of statistical case ambiguity resolution in Korean. Sections 5.1 and 5.2 report the experimental results for the discrete and the sequential case decision models in turn. Section 5.3 discusses the roles of each feature used in the models and compares the discrete and the sequential case decision models. We also present some theoretical implications of statistical case ambiguity resolution. Section 5.4 presents vagaries of the data we use for our experiments. Finally, this chapter is summarised in Section 5.5.

5.1 Discrete Case Decision Models

In this section, the experimental results for the discrete case decision model are presented. We start with the basic model and extend this model by incorporating more features into it.

5.1.1 The Basic Model

The basic discrete case decision model DCD_0 uses the minimal set of features: the focus nominal (n) and the predicate (v). As described in Chapter 3, we represent a case decision process as a joint probabilistic event. Thus, DCD_0 is formalised as (5.1).

Annotation	Measure	Training set size									
		0.8M	1.6M	2.4M	3.2M	4.0M	4.8M	5.6M	6.4M	7.2M	8.0M
Full context	Agree	68.01	68.43	68.64	70.57	70.57	70.65	71.24	71.62	71.07	72.42
	Kappa	0.47	0.48	0.48	0.51	0.51	0.51	0.52	0.52	0.52	0.53
Lim context	Agree	64.57	65.58	66.29	67.76	67.84	68.22	67.80	68.64	68.18	69.23
	Kappa	0.44	0.45	0.46	0.48	0.48	0.49	0.48	0.50	0.49	0.50

Table 5.1: *Average pairwise agreement and Kappa for DCD_0 evaluated against full and limited context annotations*

$$\begin{aligned} DCD_0 &= \arg\max_j P(v, j, n) \\ &= \arg\max_j P(v)P(j|v)P(n|v, j) \\ &= \arg\max_j P(j|v)P(n|v, j) \end{aligned} \tag{5.1}$$

As presented in Section 3.3.1, the probability is estimated from the counts obtained form the corpus. To smooth the counts, the back-off strategy which is also described in Section 3.3.1 is used.[1]

We have tried other variable orderings such as (v, n, j). However, the order (v, j, n), which is believed to be accordant with the linguistic causal relation between the three variables, obtained the best result. Table 5.1 shows the average pairwise agreements between the output of the system and the two sets of human annotations. Agreements were measured ten times while increasing the number of the training examples by 10% at each stage.

First of all, this model agrees more with the full context annotations than the limited context annotations. That is, although DCD_0 uses very limited features and contextual information, even less than the limited context annotators, the output of the system is much more similar to the full context annotations.[2]

The average pairwise agreements between DCD_0 and the full context annotations started off with 68.01% and 0.47 and ended up with 72.42% and 0.53. The performance improved along with the increase of the number of training examples. The agreements between the limited context annotations exhibit the same aspects. It is hard to predict how the model will behave if we provide more training data (Banko and Brill, 2001a,b).

As presented in Chapter 2, previous statistical approaches also used only n and v as the feature of the statistical case ambiguity resolution models. For comparison, we implemented

[1] All the discrete case decision models introduced in this section use the same back-off strategy.

[2] This tendency is maintained in all case decision models. From now on, we only concentrate on the agreements with full context annotations.

Annotation	Measure	Training set size									
		0.8M	1.6M	2.4M	3.2M	4.0M	4.8M	5.6M	6.4M	7.2M	8.0M
Full context	Agree	61.04	61.34	63.22	63.24	63.39	63.94	63.43	63.77	64.15	64.48
	Kappa	0.37	0.37	0.40	0.41	0.40	0.42	0.41	0.41	0.41	0.42
Lim context	Agree	59.40	60.12	62.01	62.05	61.63	62.17	62.01	62.47	62.68	62.80
	Kappa	0.36	0.37	0.40	0.40	0.40	0.41	0.41	0.41	0.41	0.41

Table 5.2: *Average pairwise agreement and Kappa for the word association model of Yoon et al. (1997) evaluated against full context and limited context annotations*

$P(j\|v)$	$P(n\|v,j)$	FullContext_1			FullContext_2			FullContext_3			Average (%)		
		Agr	Dsgr	Sum	Agr	Dsgr	Sum	Agr	Dsgr	Sum	Agr	Dsgr	Sum
Bigram	Trigram	82	23	105	80	25	105	82	23	105	77.46	22.54	100.00
Bigram	Bigram	371	107	478	366	112	478	361	117	478	76.57	23.43	100.00
Bigram	Unigram	29	19	48	29	19	48	30	18	48	61.11	38.89	100.00
Unigram	Bigram	80	51	131	79	52	131	78	53	131	60.31	39.69	100.00
Unigram	Unigram	9	7	16	8	8	16	7	9	16	50.00	50.00	100.00
Default		12	4	16	11	5	16	11	5	16	70.83	29.17	100.00
Sum		583	211	794	573	221	794	569	225	794	72.42	27.58	100.00

Table 5.3: *Decomposition of the output of DCD_0 according to the back-off stages*

a case ambiguity resolution system based on the word association model used in Yoon et al. (1997); Yoon (1998) and Chung (1999) shown below.

$$Assoc(v,n,j) = \alpha \times \overline{Assoc}(v,n,j) + (1-\alpha) \times \overline{Assoc}(v.j) \qquad (0.5 \le \alpha \le 1) \tag{5.2}$$

$$\overline{Assoc}(v,n,j) = P(n,j|v) \tag{5.3}$$

$$\overline{Assoc}(p,v) = P(j|v) \tag{5.4}$$

The evaluation result for this word association model is given in Table 5.2.[3] The performance of this model is far below that of DCD_0 even though it uses the same features. The performance difference between the two models could be attributed to the fact that our model is based on a sound probabilistic reasoning of the case decision process.

As stated in Chapter 3, we use a simple back-off smoothing method to cope with the data sparseness problem. To assess the effectiveness of the back-off smoothing, we decomposed the output of the system according to the back-off stages and compared the output with the full context annotations. The decomposition result is shown in Table 5.3.

We observe that a large number of responses returned by DCD_0 agreed with all the three full context annotations when the probability terms were backed-off to the 'Bigram-Bigram'

[3]The weight α was set to 0.999 as suggested in Chung (1999).

DCD_0		$FullContext_1$		$FullContext_2$		$FullContext_3$		Average	
		Agr No	Prec	Agr No	Prec	Agr No	Prec	Agr No	Prec
NOM	423	388	91.73	380	89.83	376	88.89	381.33	90.15
ACC	179	82	45.81	78	43.58	75	41.90	78.33	43.76
LOC	131	93	70.99	94	71.76	98	74.81	95.00	72.52
DAT	5	0	0.00	1	20.00	1	20.00	0.67	13.33
INST	42	10	23.81	10	23.81	9	21.43	9.67	23.02
COM	14	10	71.43	10	71.43	10	71.43	10.00	71.43
Sum	794	583	73.43	573	72.17	569	71.66	575.00	72.42

Table 5.4: *Precision measures for the decomposed output of DCD_0*

stage. For example, 478 test instances were processed at this stage and 366 instances agreed with $FullContext_2$. The agreement ratio is 76.57%. Other back-off stages also properly did their jobs. The 'Default' stage is activated when any of the two probability terms gets a value less than the frequency cut-off K.[4] This stage chooses the NOMINATIVE case particle as a default. As a whole, we witness that the back-off smoothing is quite effective for the model.

For a closer look at the output of the DCD_0, we decomposed the system output according to the six target case particles and measured precision and recall for each target case particle as shown in Table 5.4 and Table 5.5. These tables show that DCD_0 is good at picking up the NOMINATIVE, LOCATIVE, and COMITATIVE case particles. By contrast, DCD_0 is not good with the ACCUSATIVE, DATIVE, and INSTRUMENTAL case particles. In Table 5.5, we observe that the recall measures for the latter three case particles are quite high compared to the precision measures. The average recall for the ACCUSATIVE case particle is 80.76, which is the highest recall measure in this table, whereas the average precision for the case particle is 43.76.

For a further examination of the behaviour of DCD_0, we compare the confusion matrix for the pair ($FullContext_2$, DCD_0) with two confusion matrices for the pairs ($FullContext_2$, $FullContext_3$) and (FullContext-3, $LimContext_3 3$). The three matrices are shown in Table 5.6.[5]

In the confusion matrix for ($FullContext_2$, DCD_0), the confusions between the NOMINATIVE and three other case particles ACCUSATIVE, LOCATIVE, DATIVE and INSTRUMENTAL case particles are conspicuous.

The confusions between the NOMINATIVE case particle and the DATIVE case particle are also

[4] We used $K = 1$ throughout the experiments.

[5] Full set of confusion matrices are attached in Appendix E.

Annotation	Case particle		DCD_0	
			Agree No	Recall
$FullContext_1$	NOM	530	388	73.21
	ACC	100	82	82.00
	LOC	135	93	68.89
	DAT	0	0	·
	INST	13	10	76.92
	COM	16	10	62.50
$FullContext_2$	NOM	519	380	73.22
	ACC	96	78	81.25
	LOC	140	94	67.14
	DAT	3	1	33.33
	INST	19	10	52.63
	COM	17	10	58.82
$FullContext_3$	NOM	510	376	73.73
	ACC	95	75	78.95
	LOC	149	98	65.77
	DAT	6	1	16.67
	INST	18	9	50.00
	COM	16	10	62.50
Average	NOM	519.67	381.33	73.38
	ACC	97.00	78.33	80.76
	LOC	141.33	95.00	67.22
	DAT	3.00	0.67	22.22
	INST	16.67	9.67	58.00
	COM	16.33	10.00	61.22

Table 5.5: *Recall measures for the decomposed output of DCD_0*

(a) X: FullContext$_2$, Y: FullContext$_3$ (Agree 95.21%, Kappa 0.91)

	NOM	ACC	LOC	DAT	INST	COM	Sum
NOM	501	4	2	1	1	1	510
ACC	5	88	2	0	0	0	95
LOC	7	3	135	0	4	0	149
DAT	4	0	0	2	0	0	6
INST	2	1	1	0	14	0	18
COM	0	0	0	0	0	16	16
Sum	519	96	140	3	19	17	794

(b) X: FullContext$_2$, Y: LimContext$_3$ (Agree 86.02%, Kappa 0.75)

	NOM	ACC	LOC	DAT	INST	COM	Sum
NOM	454	5	9	1	5	2	476
ACC	19	84	7	0	2	0	112
LOC	20	6	119	0	3	0	148
DAT	9	0	0	2	0	0	11
INST	14	0	2	0	9	0	25
COM	3	1	3	0	0	15	22
Sum	519	96	140	3	19	17	794

(c) X: FullContext$_2$, Y: DCD$_0$ (Agree 72.17%, Kappa 0.53)

	NOM	ACC	LOC	DAT	INST	COM	Sum
NOM	380	6	26	2	5	4	423
ACC	86	78	12	0	3	0	179
LOC	29	4	94	0	1	3	131
DAT	3	1	0	1	0	0	5
INST	19	6	7	0	10	0	42
COM	2	1	1	0	0	10	14
Sum	519	96	140	3	19	17	794

Table 5.6: *Confusion matrices for the pairs (FullContext$_2$, FullContext$_3$), (FullContext$_2$, LimContext$_3$), and (FullContext$_2$, DCD$_0$)*

Annotation	Measure	Training set size									
		0.8M	1.6M	2.4M	3.2M	4.0M	4.8M	5.6M	6.4M	7.2M	8.0M
Full context	Agree	70.24	71.62	72.59	73.09	73.51	74.27	73.51	73.85	74.14	74.73
	Kappa	0.50	0.52	0.54	0.55	0.55	0.56	0.55	0.55	0.56	0.57
Lim context	Agree	68.93	70.45	72.17	72.08	72.50	72.17	71.87	72.04	72.12	73.47
	Kappa	0.50	0.52	0.55	0.55	0.56	0.55	0.54	0.55	0.54	0.57

Table 5.7: *Average pairwise agreement and Kappa for DCD_1 evaluated against full and limited context annotations*

salient in the pairs (FullContext_2, FullContext_3) and (FullContext_3, LimContext_3). These confusions seem to be related to the case particle alternations.

We can also see that the confusion patterns of (FullContext_2, FullContext_3) and (FullContext_2, DCD_0) are sharing a similar aspect except for the confusion between the NOMINATIVE case particle and the ACCUSATIVE case particle.

5.1.2 Extended Model 1

Now we introduce a new feature s, the list of the neighbouring case particles. To neutralise the effect of the word order variation, we use the sorted list.[6] This feature can be regarded as an approximation of the subcategorisation frame of the predicate. However, this feature is far from perfect. We can easily incorporate this feature into the basic model as shown in (5.5). The variable ordering (v, j, n, s) was also chosen following the causal relationships of the variables. The ordering (v, s, j, n) could be a reasonable choice. However, other features cannot depend on s as it can only be determined after the case slot is filled.

$$\begin{aligned} DCD_1 &= \arg\max_j P(v, j, n, s) \\ &= \arg\max_j P(v)P(j|v)P(n|v, j)P(s|v, n, j) \\ &= \arg\max_j P(j|v)P(n|v, j)P(s|v, n, j) \end{aligned} \tag{5.5}$$

Table 5.7 shows the evaluation result for the model. The best agreement measures 74.73% and 0.57 are obtained when the model was trained on the full training set. Overall, DCD_1 outperforms DCD_0 and this confirms that the introduction of the feature s made a difference. The improvement is also depicted in Table 5.8 and Table 5.9.

[6] We have also tried with an unsorted version of s. However, the sorted s worked better than the unsorted s.

DCD_1		$FullContext_1$		$FullContext_2$		$FullContext_3$		Average	
		Agr No	Prec	Agr No	Prec	Agr No	Prec	Agr No	Prec
NOM	429	397	92.54	393	91.61	390	90.91	393.33	91.69
ACC	148	81	54.73	76	51.35	75	50.68	77.33	52.25
LOC	145	98	67.59	101	69.66	105	72.41	101.33	69.89
DAT	14	0	0.00	2	14.29	3	21.43	1.67	11.90
INST	44	9	20.45	10	22.73	10	22.73	9.67	21.97
COM	14	10	71.43	10	71.43	10	71.43	10.00	71.43
Sum	794	595	74.94	592	74.56	593	74.69	593.33	74.73

Table 5.8: *Precision measures for the decomposed output of DCD_1*

When we compare Table 5.8 and Table 5.9 with Table 5.4 and Table 5.5, we find that the performance improvement was concentrated on the ACCUSATIVE case particle. The average precision for the ACCUSATIVE case particle went up from 43.76 to 52.25 while the average precision is slightly dropped. For the LOCATIVE and the DATIVE case particles, there were improvement of the reacall measures. For other case particles, no noticeable changes were found. The confusion matrix for the pair ($FullContext_2$, DCD_1) in Table 5.10 clearly demonstrates the effect of the feature *s*.

According to Table 5.10, DCD_1 picked up more NOMINATIVE case particles as answers from the confusions between the NOMINATIVE and the ACCUSATIVE case particles compared with DCD_0. It has brought the performance improvements on both case particles. However, the LOCATIVE, DATIVE, and INSTRUMENTAL case particles received little help from the feature *s*.

The improved performance of DCD_1 comes with a price. The use of a new feature also means the increase in the computation time. We can reduce the computation time if we can safely simplify DCD_1 by making an independence assumption. We presume that the link between *n* and *s* in the probability term $P(s|v,n,j)$ is relatively weak and make an assumption that the features *n* and *s* are mutually independent. As the result, we get the simplified version of DCD_1, DCD_{1s} as shown in (5.6).

$$\begin{aligned} DCD_{1s} &= \arg\max_j P(j|v)P(n|v,j)P(s|v,n,j) \\ &= \arg\max_j P(j|v)P(n|v,j)P(s|v,j) \end{aligned} \tag{5.6}$$

The performance of DCD_{1s} is almost as good as DCD_1 as displayed in Table 5.11. The overall agreement measures for the simplified model are slightly below the measures for the

Annotation	Case particle		DCD_1	
			Agree No	Recall
$FullContext_1$	NOM	530	397	74.91
	ACC	100	81	81.00
	LOC	135	98	72.59
	DAT	0	0	·
	INST	13	9	69.23
	COM	16	10	62.50
$FullContext_2$	NOM	519	393	75.72
	ACC	96	76	79.17
	LOC	140	101	72.14
	DAT	3	2	66.67
	INST	19	10	52.63
	COM	17	10	58.82
$FullContext_3$	NOM	510	390	76.47
	ACC	95	75	78.95
	LOC	149	105	70.47
	DAT	6	3	50.00
	INST	18	10	55.56
	COM	16	10	62.50
Average	NOM	519.67	393.33	75.69
	ACC	97.00	77.33	79.73
	LOC	141.33	101.33	71.70
	DAT	3.00	1.67	55.56
	INST	16.67	9.67	58.00
	COM	16.33	10.00	61.22

Table 5.9: *Recall measures for the decomposed output of DCD_1*

	NOM	ACC	LOC	DAT	INST	COM	Sum
NOM	393	8	20	1	3	4	429
ACC	58	76	11	0	3	0	148
LOC	32	6	101	0	3	3	145
DAT	11	1	0	2	0	0	14
INST	23	4	7	0	10	0	44
COM	2	1	1	0	0	10	14
Sum	519	96	140	3	19	17	794

Table 5.10: *Confusion matrix for the pair X: FullContext$_2$, Y: DCD$_1$ (Agree 74.56%, Kappa 0.57)*

Annotation	Measure	Training set size									
		0.8M	1.6M	2.4M	3.2M	4.0M	4.8M	5.6M	6.4M	7.2M	8.0M
Full context	Agree	69.73	71.33	72.29	72.29	73.26	73.22	73.47	74.22	73.26	74.73
	Kappa	0.49	0.51	0.53	0.53	0.55	0.54	0.55	0.56	0.54	0.57
Lim context	Agree	68.56	69.65	71.41	71.33	72.17	72.38	72.59	73.13	71.79	73.17
	Kappa	0.49	0.51	0.54	0.54	0.55	0.55	0.55	0.56	0.54	0.56

Table 5.11: *Average pairwise agreement and Kappa for DCD$_{1s}$ evaluated against full and limited context annotations*

original model. The peak scores 74.73% and 0.57 are preserved. Considering the reduction of the computation time (30%), a small amount of performance degradation would be acceptable in most situations.

5.1.3 Extended Model 2

The second new feature we use is the distance between the focus nominal and the predicate (d). As already noted, the Korean language has a very flexible word order. Consequently, the feature d could be useless. According to our observation, however, some case particles are strongly associated with particular positions in sentences. We know that SOV is the predominant word order in Korean. There are also some suggestions that a particular set of predicates has a particular word order preference.[7] However, we do not have any concrete empirical evidence.

We use three fixed values for the feature d. If a nominal is adjacent to a predicate, d gets the value 'n' and if a nominal is in the beginning of a sentence, d is assigned the value 'f'.

[7] For example, according to Yu (1997), adjectives conveying the meaning of possession/existence prefer the word order 'LOC-NOM' to the word order 'NOM-LOC' which is preferred by most predicates.

Annotation	Measure	Training set size									
		0.8M	1.6M	2.4M	3.2M	4.0M	4.8M	5.6M	6.4M	7.2M	8.0M
Full context	Agree	73.93	73.26	75.02	75.19	75.44	76.32	76.07	76.70	76.70	77.16
	Kappa	0.55	0.55	0.57	0.57	0.58	0.59	0.59	0.60	0.60	0.61
Lim context	Agree	70.86	69.94	72.80	73.17	73.17	73.30	73.68	74.01	73.97	74.35
	Kappa	0.52	0.51	0.56	0.56	0.56	0.56	0.57	0.57	0.57	0.58

Table 5.12: *Average pairwise agreement and Kappa for DCD_2 evaluated against full and limited context annotations*

Every position in between 'n' and 'f' gets the value 'm'.

We choose the variable ordering (v, j, n, d). We followed the same rationale for the variable ordering as we did with DCD_1. The model based on a joint probabilistic representation of the case decision with a new feature d is shown in (5.7).

$$\begin{aligned} DCD_2 &= \arg\max_j P(v, j, n, d) \\ &= \arg\max_j P(v)P(j|v)P(n|v, j)P(d|v, n, j) \\ &= \arg\max_j P(j|v)P(n|v, j)P(d|v, n, j) \end{aligned} \tag{5.7}$$

This model achieved an impressive result as shown in Table 5.12. DCD_2 performed very well even with the smallest training set. The agreement measures when the model was trained on the smallest training set are 73.93% and 0.55. These figures are well over the best figures of DCD_0. This model reached the peak performance with agreement measures 77.16% and 0.61. The overall performance is also better than that of DCD_1. Especially, we notice the big improvement of the $Kappa$ value. The $Kappa$ value exceeded 0.60 for the first time with this model.

Now we turn to the following tables which show the precision and recall for the output of the system measured against the full context annotations decomposed into the responses for the individual target case particles.

In Table 5.13 and Table 5.14, we see that both precision and recall measures for the four case particles NOMINATIVE, ACCUSATIVE, INSTRUMENTAL and COMITATIVE case particles have improved compared with Table 5.8 and Table 5.9. The improvements regarding precisions for the case particles INSTRUMENTAL and COMITATIVE draw our attention. Intuitively, these case particles tend to be closely related to predicates and placed adjacent to them. Therefore, the feature d was helpful for picking up these case particles. The recall measures for these particles have also been improved. Overall the feature d is more useful than s used

DCD_2		$FullContext_1$		$FullContext_2$		$FullContext_3$		Average	
		Agr No	Prec	Agr No	Prec	Agr No	Prec	Agr No	Prec
NOM	443	411	92.78	408	92.10	402	90.74	407.00	91.87
ACC	140	84	60.00	79	56.43	78	55.71	80.33	57.38
LOC	146	100	68.49	101	69.18	107	73.29	102.67	70.32
DAT	14	0	0.00	2	14.29	3	21.43	1.67	11.90
INST	40	10	25.00	12	30.00	11	27.50	11.00	27.50
COM	11	10	90.91	10	90.91	10	90.91	10	90.91
Sum	794	615	77.46	612	77.08	611	76.95	612.67	77.16

Table 5.13: *Precision measures for the decomposed output of DCD_2*

in DCD_1.

The effectiveness of the feature d is also illustrated in Table 5.15. This table tells us that the feature d had not much effect on the performances regarding the DATIVE case particle. The feature d had a slight positive effect on the INSTRUMENTAL case particle. However, it is still one of the most confused case particles. The case particle which got the most benefit from the feature d is the COMITATIVE case particle.

We also attempted to simplify the model DCD_2 to reduce the computation time. It seems that v has a stronger link with d than with n. However, since d is a feature directly associated with a case slot, discarding n could be harmful. Table 5.16 shows the performance of the simplified model (5.8).

$$\begin{aligned} DCD_{2s} &= \arg\max_j P(j|v)P(n|v,j)P(d|v,n,j) \\ &= \arg\max_j P(j|v)P(n|v,j)P(d|v,j) \end{aligned} \tag{5.8}$$

Unfortunately, the performance of DCD_{2s} dropped down as we expected. This model performed better than the basic model DCD_0 and managed to chase up DCD_1 and DCD_{1s}. However, the overall performance of the model is slightly under those of DCD_1 and DCD_{1s}.

5.2 Sequential Case Decision Model

The sequential case decision model which is based on a Markov chain tagging model is formalised as (5.9).[8]

[8]The full derivation is given in Chapter 3.

Annotation	Case particle		DCD_2	
			Agree No	Recall
$FullContext_1$	NOM	530	411	77.55
	ACC	100	84	84.00
	LOC	135	100	74.07
	DAT	0	0	·
	INST	13	10	76.92
	COM	16	10	62.50
$FullContext_2$	NOM	519	408	78.61
	ACC	96	79	82.29
	LOC	140	101	72.14
	DAT	3	2	66.67
	INST	19	12	63.16
	COM	17	10	58.82
$FullContext_3$	NOM	510	402	78.82
	ACC	95	78	82.11
	LOC	149	107	71.81
	DAT	6	3	50.00
	INST	18	11	61.11
	COM	16	10	62.50
Average	NOM	519.67	407.00	78.32
	ACC	97.00	80.33	82.82
	LOC	141.33	102.67	72.64
	DAT	3.00	1.67	55.56
	INST	16.67	11.00	66.00
	COM	16.33	10.00	61.22

Table 5.14: *Recall measures for the decomposed output of DCD_2*

	NOM	ACC	LOC	DAT	INST	COM	Sum
NOM	408	6	23	0	2	4	443
ACC	47	79	11	0	3	0	140
LOC	34	6	101	0	2	3	146
DAT	11	1	0	2	0	0	14
INST	19	3	5	1	12	0	40
COM	0	1	0	0	0	10	11
Sum	519	96	140	3	19	17	794

Table 5.15: *Confusion matrix for the pair X: FullContext$_2$, Y: DCD$_2$ (Agree 77.08%, Kappa 0.60)*

Annotation	Measure	Training set size									
		0.8M	1.6M	2.4M	3.2M	4.0M	4.8M	5.6M	6.4M	7.2M	8.0M
Full context	Agree	70.99	71.62	72.04	73.26	73.05	74.01	73.51	73.68	74.39	74.39
	Kappa	0.51	0.52	0.53	0.55	0.54	0.56	0.55	0.55	0.56	0.56
Lim context	Agree	67.80	68.93	70.19	71.54	71.41	71.91	71.66	71.79	72.25	72.92
	Kappa	0.48	0.50	0.52	0.54	0.53	0.54	0.54	0.54	0.55	0.55

Table 5.16: *Average pairwise agreement and Kappa for DCD$_{2s}$ evaluated against full and limited context annotations*

$$SCD = \arg\max_{j_{1,n}} \prod_i P(n_i|j_i, v)P(j_i|j_{i-1}) \tag{5.9}$$

In contrast to the discrete case decision models in which each case decision in a sentence is performed in isolation, the case decision process is understood as a sequential event in the sequential case decision model. The task is to determine the most probable case particle sequence given a sequence of nominals and a predicate. In our model, the sequence works backwards from the predicate. The first case decision is made using DCD_0.

We implemented a case ambiguity resolution system based on the sequential case decision model adopting the conventional design of a Markov chain part-of-speech tagger.[9]

According to Table 5.17, the overall performance of SCD is better than that of DCD$_1$ which uses s as an additional feature. However, it is worse than DCD$_2$ which uses d. The best agreement measures are 76.45% and 0.60. The precision and recall measures for the decomposed output of the systems are shown in Table 5.18 and Table 5.19.

Table 5.18 tells us that the overall picture is not much different from DCD$_1$ and DCD$_2$.

[9]We borrowed the code from the HMM module in the Natural Language Toolkit (Bird and Loper, 2004) which is available at `http://nltk.sourceforge.net`.

Annotation	Measure	Training set size									
		0.8M	1.6M	2.4M	3.2M	4.0M	4.8M	5.6M	6.4M	7.2M	8.0M
Full context	Agree	72.17	72.88	74.01	74.22	75.48	75.86	75.61	74.64	75.61	76.45
	Kappa	0.53	0.54	0.56	0.56	0.58	0.59	0.58	0.57	0.58	0.60
Lim context	Agree	70.99	71.28	72.67	73.22	72.92	74.27	73.43	72.84	73.43	74.94
	Kappa	0.53	0.53	0.56	0.57	0.56	0.58	0.57	0.56	0.57	0.59

Table 5.17: *Average pairwise agreement and Kappa for SCD evaluated against full and limited context annotations*

SCD		FullContext_1		FullContext_2		FullContext_3		Average	
		Agr	Prec	Agr	Prec	Agr	Prec	Agr	Prec
NOM	432	401	92.82	399	92.36	395	91.44	398.33	92.21
ACC	133	83	62.41	78	58.65	78	58.65	79.67	59.90
LOC	161	105	65.22	107	66.46	113	70.19	108.33	67.29
DAT	14	0	0.00	2	14.29	3	21.43	1.67	11.90
INST	42	9	21.43	9	21.43	9	9.00	21.43	
COM	12	10	83.33	10	83.33	10	83.33	10.00	83.33
Sum	794	608	76.57	605	76.20	608	76.57	607.00	76.45

Table 5.18: *Precision measures for the decomposed output of SCD*

The precision measures for the case particles NOMINATIVE, ACCUSATIVE, and INSTRUMENTAL went up from 90.15, 43.76, and 71.43 to 92.21, 59.90, and 83.33 compared with DCD_0. The precision for the NOMINATIVE case particle and the ACCUSATIVE case particle are the best among all the case decision models. On the other hand, the precision measures for the case particles LOCATIVE, DATIVE, and INSTRUMENTAL went down from 72.52, 13.33, 23.02 to 67.29, 11.90, and 21.43. In tabreftbl:scd-rec, we see that recall measures have been also went up except for the INSTRUMENTAL case particle.

Table 5.20 is the confusion matrix for the pair (FullContext_2, SCD). The performance improvements regarding the NOMINATIVE case particle and the ACCUSATIVE case particle are also confirmed in the confusion matrix. However, the frequent confusions between the NOMINATIVE case particle and the case particles LOCATIVE, DATIVE, and INSTRUMENTAL largely remained unresolved.

In summary, the sequential case decision model SCD was quite effective on the resolution of the confusions between the NOMINATIVE case particle and the two case particles ACCUSATIVE and INSTRUMENTAL. However, the overall performance was below that of the discrete case decision model DCD_2

Annotation	Case particle		SCD	
			Agree No	Recall
FullContext_1	NOM	530	401	75.66
	ACC	100	83	83.00
	LOC	135	105	77.78
	DAT	0	0	·
	INST	13	9	69.23
	COM	16	10	62.50
FullContext_2	NOM	519	399	76.88
	ACC	96	78	81.25
	LOC	140	107	76.43
	DAT	3	2	66.67
	INST	19	9	47.37
	COM	17	10	58.82
FullContext_3	NOM	510	395	77.45
	ACC	95	78	82.11
	LOC	149	113	75.84
	DAT	6	3	50.00
	INST	18	9	50.00
	COM	16	10	62.50
Average	NOM	519.67	398.33	76.65
	ACC	97.00	79.67	82.13
	LOC	141.33	108.33	76.65
	DAT	3.00	1.67	55.56
	INST	16.67	9.00	54.00
	COM	16.33	10.00	61.22

Table 5.19: *Recall measures for the decomposed output of SCD*

	NOM	ACC	LOC	DAT	INST	COM	Sum
NOM	399	8	18	0	3	4	432
ACC	42	78	9	0	4	0	133
LOC	44	4	107	0	3	3	161
DAT	11	1	0	2	0	0	14
INST	22	4	6	1	9	0	42
COM	1	1	0	0	0	10	12
Sum	519	96	140	3	19	17	794

Table 5.20: *Confusion matrix for the pair X: FullContext$_2$, Y: SCD (Agree 76.20%, Kappa 0.59)*

5.3 Discussion

This section discusses the roles of the features used in the case decision models and compares the discrete case decision model and the sequential case decision model. Some theorethical implications of statistical case ambiguity resolution are also presented.

5.3.1 The Roles of v, n, s, and d in Statistical Case Ambiguity Resolution

The biggest role players in the statistical case ambiguity resolution task are undoubtedly the predicate (v) and the focus nominal (n). After all, a case is the marking of the relationship between a predicate and a nominal. If we follow the explanation provided by a particular theory like GB theory, a case is assigned by a predicate to a nominal either directly or indirectly and the case is (optionally) marked by a case marker. Therefore, the pair of a predicate and a nominal is expected to give enough information about the case involved with the two words when dealing with an ambiguous instance. Sentences in (76) are examples in which all the full context annotations and DCD$_0$ have agreed on the case decisions. Recall that DCD$_0$ is the model which uses only v and n as its features for case ambiguity resolution.[10]

(76) a. *Na*-neun(-**ga**) i gos-eulo isa-ggaji ha-yeoss-da.
I-TOP(-**NOM**) this place-DIR move in-even do-PST-DCL
'I even moved into this place.'

b. Je-ga han *malsseum*-∅-(**eul**) deuli-gess-seubnida.
I-NOM one speech-∅(-**ACC**) give-FTR-DCL
(lit.) 'I will speak.'

[10] Example sentences were taken form the test set. Long sentences were shortened for brevity.

c. Jumin-deul-eun maeil *achim*-∅ buntong-eul teotteul-nda
Resident-PL-TOP everyday *morning*-∅(-**LOC**) anger-ACC burst out-DCL
'The residents burst out their anger everyday morning.'

d. Jagpum-ui gyeogjo-do eonue *jeongdo*-neun(-**lo**)
Work of art-GEN character-also certain *degree*-TOP(-**MAN**)
yujidoe-eo iss-da.
be maintained-AUXCON exist-DCL
'The character of the work is also maintained in some degree.'

e. Seoul gonggi-wa-neun daleu-n *geos*-∅(-**gwa**) gat-ass-da.
Seoul air-COM-TOP different-ADN *thing*-∅(-**COM**) same-PST-DCL
'It seemed to be different from the air of Seoul.'

To resolve a case ambiguity, DCD_0 tries to pick up the most frequently used case particle together with *v* and *n*. For example, to process the ambiguous instance in (76a), DCD_0 looks up the counts from the corpus and returns the NOMINATIVE case particle as an answer since it is the most frequently used case particle with the focus nominal *na* 'I' and the predicate *ha-* 'do'.[11] However, not all *v* and *n* pairs exist in the training data and the model has to back-off to use less specific information. Although our simple back-off method works fine in many situations,[12] it cannot cover every ambiguous instance. The following examples contain ambiguous instances in which all the full context annotations agreed on the NOMINATIVE *-i/ga* while DCD_0 returned other case particles.[13]

(77) a. Sohwagi-ga jungyohada-go *gwangyeja-neun*(-**ga**)
Fire extinguisher-NOM important-QUOT *person concerned*-TOP(-**NOM**)
ib-eul moeu-nda.
mouth-ACC gather-DCL
'All people concerned say that fire extinguishers are important.' (*-eul/-leul* ACCUSATIVE)

b. i *geos*-eun(-**i**) jeongbu-ui jeongchaeg-gwa jeongmyeon-eulo
This *thing*-TOP(-**NOM**) government-GEN policy-COM front side-MAN
wibaedoe-nda
run counter to-DCL
'This matter runs directly counter to the policy of government.' (*-e* LOCATIVE)

c. Najung-e-neun *hunjang-nim*-ggaji(-**ga**) sonsu galeuchi-eo
Later-LOC-TOP *teacher*-HON-even(-**NOM**) personally teach-AUXCON
ju-si-eoss-da.
give-HON-PST-DCL
'Later, even the teacher personally taught.' (*-ege* DATIVE)

[11] In practice, the nominal *na* 'I' is attenuated as **NP**.

[12] See Table 5.3 in Section 5.1.1.

[13] DCD_0's responses are shown with the translations.

d. Dotoli namu-ui *teugseong-eun*(-**i**) wanjeonhi seolmyongdoe-nda
Acorn tree-GEN *characteristic*-TOP(-**NOM**) completely be explained-DCL
'The acorn tree's characteristic is completely explained.' (*-eulo/-lo* INSTRUMENTAL)

The sentence (77a) is a transitive sentence and there is a nominal *ib-eul* 'mouth-ACC' marked as an accusative. Therefore it is quite obvious that the ambiguous case should be resolved as the NOMINATIVE case particle *-i/-ga*. However, DCD_0 responded with the ACCUSATIVE case particle *-eul/-leul* because it is the most frequently used case particle with the predicate of the sentence *moeu-* 'gather'. It is natural for DCD_0 to respond like this since it does not use any contextual information. Sentences (77b), (77c) and (77d) are in similar situations.

The limitation of the model DCD_0 can be partly overcome by using additional features *s* and *d*. For instance, facing the ambiguous instance *gwangyej-neun* 'person concerned-TOP' in (77a), both DCD_1 and DCD_2 responded with the ultimate choice the NOMINATIVE case particle *-i/-ga*. It was possible for DCD_1 to pick the right answer since the existing ACCUSATIVE case particle provided vital information being used as the feature *s*. DCD_2 was also able to return the NOMINATIVE case particle *-i/-ga* using the feature *d*. The distance between the ambiguous nominal *gwangyeja* and the predicate *moeu-* 'gather' is 'f' and the most frequently used case particle in this position with the predicate is the NOMINATIVE case particle. Even with *s* and *d*, both DCD_1 and DCD_2 could not return the right answer for many test instances. The following examples are such test instances.

(78) a. *Kkongteu*-neun(-**ga**) [sahoehag-i hagmun jung-ui hagmun-i-lago]
Comte-TOP(-**NOM**) [sociology-NOM science among-GEN science-COP-QUOT]
bo-ass-da.
see-PST-DCL
'Comte saw sociology as an ultimate science.' (*-eul/-leul* ACCUSATIVE)

b. Hanbando-eseo-do *gaecheogsaeop*-eun(-**i**) iss-eoss-da.
The Korean Peninsula-LOC-also reclamation work(-**NOM**) exist-PST-DCL
'There was a reclamation work also in the Korean Peninsula.' (*-e* LOCATIVE)

c. Taipingha-neun songalag *nolim*-ggaji(-**i**) gyesandoe-nda
Type-ADN finger *move*-even(-**NOM**) be counted-DCL
'Even the moving of the fingers that are typing is counted.' (*-eulo/-lo* INSTRUMENTAL)

When dealing with the sentence (78a), DCD_1 cannot use the feature *s* since there is no surrounding case particle in the sentence. Consequently the ACCUSATIVE case particle *-eul/-leul* is chosen. DCD_2 is not successful with this sentence either. The value of *d* DCD_2 uses

given in the sentence is 'n'. The most frequently used case particle in 'n' position with the predicate *bo-* 'see' is the ACCUSATIVE case particle *-eul/-leul*.

We might get an optimal result if we could somehow incorporate all the features in a single model. However, it is not easy to do so with the current statistical modelling method based on a joint probabilistic reasoning. The features *s* and *d* are not compatible with each other and the causal relationships between the two features cannot be established. To combine *s* and *d*, we need an alternative learning method in which arbitrary and sometimes overlapping features can be used together such as log-linear models (Abney, 1997). It would be also possible to break up the feature *s* into a set of smaller pieces.

In summary, when *v* and *n* were seen in the training data, DCD_0 generally did a good job. However, in our experiments, DCD_0 left a large number of confusions between the NOMINATIVE case particle *-i/-ga* and other case particles. The features *s* and *d* considerably improved the performance of the case ambiguity resolution system. There were, however, many test instances that did not have any neighbouring case particles. DCD_1 cannot be applied to these test instances. DCD_2 did not suffer from the same problem as DCD_1, and it achieved a better result. Although the feature *d* was effective in many test instances, it was not robust enough to cope with the relatively free word order of the Korean language.

5.3.2 Comparison of the Discrete Case Decision Model and the Sequential Case Decision Model

As reported in Section 5.2, the performance of our sequential case decision model SCD was better than DCD_0 and DCD_1, but worse than DCD_2. We still believe that the underlying idea of the sequential case decision is sound and correct. A similar model has worked in other free word order languages in a similar task (Brants et al., 1997; Skut et al., 1997). The difference is the availability of the fully annotated training material and richer representation scheme which can use the information provided by the training material. We had to rely only on the unannotated training material. Consequently, we were also bound to use a very simple representation scheme for our sequential case decision model SCD. As the result, we could not find any big difference between the sequential case decision model and the discrete case decision model in terms of their performances and behaviours.

(79) a. *Na*-neun(-**ga**) i gos-eulo isa-ggaji ha-yeoss-da.
I-TOP(-**NOM**) this place-DIR move in-even do-PST-DCL
'I even moved into this place.'

b. Taipingha-neun songalag *nolim*-ggaji(-**i**) gyesandoe-nda
Type-ADN finger *move*-even(-**NOM**) be counted-DCL

'Even the moving of the fingers that are typing is counted.' (*-eulo/-lo* INSTRUMENTAL)

c. *Allegsandeo*-neun(-**ga**) geulis-leul malbalgub mit-e jisbalb-ass-da.
Alexander-TOP(-**NOM**) Greece-ACC horse hoof below-LOC tread down-PST-dcl
'Alexander trod down Greece under the hooves of horses.'

We hoped that SCD would perform better for the test sentences in which two or more ambiguous nominals exist. Indeed, it worked well with some of these sentences including (79a). However, almost every test sentence that SCD successfully disambiguated were also successfully dealt with by DCD_1 and/or DCD_2. Sentences like (79b) which were hard for DCD_1 and DCD_2 were also hard for SCD. The test sentences that are correctly disambiguated only by SCD were very rare. (79c) is one of such sentences.

5.3.3 Theoretical Implications of Statistical Case Ambiguity Resolution

With regards to our experiments on statistical case ambiguity resolution in Korean, the following theoretical implications have emerged:

First, the fact was revealed that case ambiguity is closely related to the obliqueness hierarchies. This fact was not explicitly noted in the theoretical work we have surveyed. If a case particle is less oblique, it is very likely that this particle can be deleted or unrealised and vice versa. Theoretical research is called for which can offer an integrated view on case particle deletion and unmarked case.

Second, the case particle alternation phenomenon naturally affects the task of statistical case ambiguity resolution. If an extended and comprehensive descriptive work on the case particle alternations in Korean comparable to Levin (1993) is provided, it can be used as a base material for relevant computational work (c.f. Lapata 1999). These activities will positively contribute to statistical case ambiguity resolution.

Third, word order restriction and preference also affect statistical case ambiguity resolution task. In theoretical work, mainly the underlying mechanism and the hard constraints on word order variation have been studied. If these studies can be extended to offer an inventory of word order variation with regards to the type of predicates together with information on soft word order preference, it could be very useful for statistical case ambiguity resolution.

Fourth, many test instances with ambiguous nominals that have special adverbial or modal meanings were successfully disambiguated reflecting the work of Chung (1998). If we can quantify the suggested properties of the relevant nominals, statistical ambiguity resolution

Case particle	Training set		FullContext_1		FullContext_2		FullContext_3	
	Count	Ratio (%)	Count	Ratio (%)	Count	Ratio (%)	Count	Ratio (%)
-ga/-i NOMINATIVE	2,838,454	29.25	530	66.75	519	65.37	510	64.23
-eul/-leul ACCUSATIVE	3,709,779	38.23	100	12.59	96	12.09	95	11.96
-e LOCATIVE	1,787,510	18.42	135	17.00	140	17.63	149	18.77
-ege DATIVE	202,730	2.09	0	0.00	3	0.38	6	0.76
-eulo/-lo INSTRUMENTAL	1,102,235	11.36	13	1.64	19	2.39	18	2.27
-gwa/-wa COMITATIVE	64,021	0.66	16	2.02	17	2.14	16	2.02
Sum	9,704,729	100.00	794	100.00	794	100.00	794	100.00

Table 5.21: *Distributiont of target case particles in training set and full context annotations*

will benefit greatly from the result.

5.4 The Vagaries of the Data

This section briefly looks into a few prominent vagaries of the data observed while performing the experiments.

5.4.1 Unbalanced Distribution of Case Particles and the Scarcity of the DATIVE Case Particle

As shown in Table 5.21, the distribution of the target case particles are highly unbalanced in both training and test set. It is, of course, wrong to expect that the case particles are evenly distributed in a naturally occurring text. The real distribution of the case particles which can be obtained from fully annotated data will be quite different from the distribution in unannotated data. For example, we know that the NOMINATIVE case particle is more frequent than the ACCUSATIVE case particle whereas the situation in unannotated data is the other way around. That means the NOMINATIVE case particle has a strong tendency to be deleted or unrealised compared to the ACCUSATIVE case particle. We conjecture that every target case particle has different deletion and unrealisation tendencies.

It is also notable that the DATIVE case particle is extremely rare in the test set compared to other target case particles. In FullContext_1 annotation, the DATIVE case particle does not exist at all. The DATIVE case particle is also very rare in the training set although the COMITATIVE case particle is still rarer. The scarcity of the DATIVE case particle can be explained from the following facts. First, as noted in Section 2.2.1.4, particle *-ege* is used only with animate nominals while *-e* is used with inanimates. The animate nominals are quite rare compared to the inanimates. It is also true that when the animate nominals are proper

Annotation	Measure	DCD_0		DCD_1		DCD_2		SCD	
		Treebank	ClSeg	Treebank	ClSeg	Treebank	ClSeg	Treebank	ClSeg
Full context	Agree	62.59	62.30	63.90	63.56	67.51	68.09	66.56	65.83
	Kappa	0.39	0.39	0.39	0.39	0.46	0.46	0.43	0.44
Lim context	Agree	59.87	60.12	63.73	63.56	65.95	66.54	66.92	65.66
	Kappa	0.37	0.38	0.41	0.42	0.46	0.46	0.46	0.46

Table 5.22: *Pairwise agreement and Kappa for statistical models trained on data set constructed from the treebank-extracted clauses and segmented clauses*

nouns (e.g. names), it is probable that the part-of-speech tagger misanalysed or failed to analyse the wordforms containing the proper nouns. Second, as presented in Section 2.3.4, there exist case particle alternations in Korean. Particle *-ege* DATIVE is interchangeable with *-i/-ga* NOMINATIVE or *-eul/-leul* ACCUSATIVE case particles. In fact, all the focus nominals annotated as *-ege* DATIVE in FullContext_2 and/or FullContext_3 are annotated as *-i/ga* NOMINATIVE in FullContext_1. The following example shows the focus nominals annotated as *-e* DATIVE in FullContext_2 and/or FullContext_3 whereas it was annotated as *-i/ga* NOMINATIVE in FullContext_1.

(80) Seujeukki bagsa-neun *dongyangin-deul-eun* jaa-leul chimjamsiki-lyeoneun
Suzuki doctor-TOP easterner-PL-TOP ego-ACC calm down-ADN
gyeonghyang-i iss-go, *seoyangin-deul-eun* jaa-leul gangjoha-lyeoneun
tendency-NOM exist-COCON, westerner-PL-TOP ego-ACC emphasise-ADN
gyeonghyang-i iss-dago jijeogha-nda.
tendency-NOM exist-QUOT point out-DCL.
'Dr Suzuki points out that the eastern people have tendency of calming down their egos and the western people have tendency of emphasising their egos.'

The reason why the NOMINATIVE case particle *-i/-ga* is so predominant in the test set could be also explained by the case particle alternations. However, we do not have any concrete evidence for this claim for the moment.[14]

5.4.2 The Effect of the Knowledge-Lean Clause Segmentation

As laid out in Section 4.1, we did not use any high level language processing tools except for a part-of-speech tagger in constructing the training data. Although we evaluated the performance of our knowledge-lean data collection method in Section 4.1.4 and Section 4.1.5, it is difficult to predict the effect of the data collection method on the actual performances

[14] In order to investigate the case particle alternation prevalence, a larger human annotation experiment is required. It should also be noted that the test set is extracted from the small-size treebanks which consist of texts from limited subjects and genres.

of the statistical models. To see the effect of the clause segmentation method, we trained our statistical learners on the training set constructed from the clauses in the treebanks and another training set constructed from the clauses segmented by our method in the same treebanks even though the sizes of the training sets are too small to make a general claim. The pairwise agreements and Kappa measures between the outputs of the learners and the human annotations are shown in Table 5.22.

Surprisingly, the performance differences between the leaners trained on the treebank clauses and the segmented clauses are not that serious. There are some odd situations where the learner trained on the segmented clauses slightly outperforms the learner trained on the treebank clauses. Again, the sizes of the training sets are too small to make a general claim and it is hard to see what is going on under the surface. We saw that the size of the training data has a positive effect on the performances of the statistical models. However, it is not possible to explore the effect of the size of the training data in relation with the data collection method until a large-size treebank of Korean is available.

5.4.3 Odd Corpus Segment

As described in Section 3.2, the source material of the 60,863,000-word corpus we use came from diverse texts of various subjects and genres. It is natural to expect that different texts in different subjects and genres show different tendencies of case particle deletion and unrealisation. For example, texts from the primary school textbooks have more sentences with explicit case markings than other texts do while texts containing verbal communications have more sentences with implicit case markings. To neutralise the effect of the subject and genre differences, we shuffled the training material. Then we constructed 10 training sets while increasing the size of the training examples by 0.8M instances.[15]

We expected that the performances of the statistical models would increase while the size of the training set increased. Overall, this expectation was correct. However, performance dips in 6.4M and 7.2M points were observed. We do not have a satisfactory explanation regarding this matter. We can only conjecture that source texts included in these particular subsets affected the statistical models in some way.

[15] Technically, we divided the whole training set into 10 subsets and used them incrementally to save the storage space.

5.4.4 Data Sparseness and the Performance of the Sequential Case Decision Model

We expected that the sequential case decision model (SCD) which takes account of previous case decision history[16] performed well. However, the performance of SCD was below that of DCD_2 which is a discrete case decision model. We think that this is due to the limitations of the training data. Since our training data is not annotated, the case particle sequences that do not have any deleted or unrealised case particles are very rare. Furthermore, the linguistic characteristics of the Korean language means that argument drop is quite frequent, contributing to the data sparseness, which leads to the disappointing performance of SCD.

5.5 Summary

In this chapter, we presented the results of the statistical case ambiguity resolution experiments. Analyses of the evaluation results are also given for individual case decision models. The discrete case decision model using the features v, n, and d was the best performer. Any significant difference between the discrete case decision model and the sequential case decision model is not found. We discussed the theoretical implications of case ambiguity resolution. Finally we also looked at a few vagaries of the data revealed in our experiments.

[16] Note that the case decision sequence works backwards from the predicate in our model.

Chapter 6

Conclusion

This chapter concludes this thesis by presenting the results and contribution of the current work. We also suggest some possible future research directions.

6.1 Results and Contributions

The aim of this thesis is to tackle the case ambiguity problem in Korean with statistical methods. We obtained the following results from our work.

First, through an examination of the relevant theoretical work, we clearly identified the syntactic and lexical semantic causes for the case ambiguity problem in Korean. We were also able to precisely define the task and the target case particles.

Second, we provided a clear specification of our knowledge-lean training data construction method. The effectiveness of the data collection method was indirectly measured by applying the method to two treebanks of Korean consisting of 25,258 syntactically analysed sentences in total. It turned out that even though our data collection method is based on a set of very simple heuristic rules, the method could extract the training data consisting of the case decision instances from an unannotated material of reasonably good quality.

Third, we suggested two case decision models for the task of case ambiguity resolution: discrete case decision model and sequential case decision model. In the discrete case decision model, each case ambiguity in a sentence was treated in isolation. In the sequential case decision model, every case decision in a sentence is treated in the context of a series of case decisions that take place in the sentence. The discrete case decision model is based on a simple joint probabilistic representation of the case decision process. We incorporate the two new features, the list of neighbouring case particles and the distance between the focus

nominal and the predicate, which have never been used before into the discrete case decision model. We found that the two new features improved the performance of our case ambiguity resolution system. For the sequential case decision model, we adopted the well-known Markov chain tagging model. Due to the limitations of the representation scheme and the training set extracted from an unannotated material, we could not achieve any considerable performance improvement with the sequential case decision model. The overall performance of the best discrete case decision model was superior to the sequential case decision model.

We tried to bring forward the issues that previous approaches were not concerned about while pursuing the aim of the thesis. The contributions of the current work to the statistical case ambiguity resolution in Korean are as follows:

First, we clearly identified the case ambiguity problem in Korean and established the target case particles while paying cautious attention to the linguistic details by consulting the relevant theoretical work. The existing work approached this problem mostly from the computer science perspective taking very simplistic views of the linguistic facts. Thus only two or three case particles were considered as target case particles without proper justifications. We examined the theoretical work and identified the target case particles involved in two linguistic phenomena, case particle deletion and case particle unrealisation, that cause case ambiguity in Korean.

Second, we presented a fully reproducible data collection method, where existing work leaves many details unstated. We also attempted to measure the effect of our knowledge-lean data collection method. Due to the lack of sufficient syntactically annotated material, we were unable to draw a general conclusion regarding the matter. At least, we confirmed that the effect of the knowledge-lean data collection method is not very serious in a small-size training set.

Third, we exploited two new features, the list of neighbouring case particles and the distance between the focus nominal and the predicate, that have not been used before yet are easily obtainable from an unannotated training material using our simple data collection method. We achieved quite good results without using any external language resources such as a thesaurus which existing work extensively used. However, direct comparisons between our results and the results of previous work were not possible.

Fourth, we constructed our statistical case ambiguity resolution models based on sound probabilistic reasoning by considering the case decision operation as a joint probabilistic event. Even though previous statistical approaches used statistical information obtained from corpora, their models were not exactly probabilistic and not easy to extend. By con-

trast, we started from a simple joint probabilistic view of the case decision and factored out the variables following the linguistic causal relations involved in the case decision process. According to our experiments, considering the linguistic causal relations has a positive effect on the performances of the statistical models.

Fifth, we evaluated our statistical models on a test set annotated by six human judges. We constructed two training sets that have different ranges of contextual information. We provided agreement percentage and $Kappa$ statistic as well as precision and recall measures evaluated for the outputs of the models decomposed into six target case particles. Our test set also has a much wider coverage than the test sets used in most existing approaches, that typically included a limited number of test instances for a small set of predicates in the test sets. Although not definitive, the multiple human annotations confirmed the value of our approach.

6.2 Limitations and Future Work

Despite the positive contributions presented in the previous section, our approach still has its own limitations and requirements for future work that can be summarised as follows:

First, since we are using unannotated material, the training set contains a considerable amount of noisy data affecting the performances of the statistical models even though we tried to compensate for the noise by using a very large training set. We might use filtering techniques for the feature values such as *hypothesis testing*. What is more serious is the data sparseness problem which has a negative effect on the performance of the sequential case decision model which is thought to be more suitable for case ambiguity resolution than the discrete case decision models. This limitation which is bound to the unannotated training material could be overcome by using fully annotated resources. We can construct a relatively small amount of fully annotated traning material and use a co-training learning method which can maximise the use of unannotaed resources with a small annotated training set.

Second, although our models are based on a simple joint probabilistic model and are fairly easy to update with new features, it is still hard to reflect alternative feature representation schemes. For example, the feature s can be decomposed into a set of binary features which indicate the existence of a particular set of case particles. In future work, the learning methods that can handle arbitrary, overlapping features such as log-linear models would be appropriate.

Third, as the test instances we used for the evaluation of the statistical models were ex-

tracted from small-size treebanks that contain a limited variety of texts, there can be a question regarding the representativeness of the test set. It would be also beneficial if we could use a larger test set yet it would require a considerable amount of effort and time.

Fourth, we looked at some theoretical issues related to the case ambiguity problem in Korean and discovered a few linguistic clues that can be used for case ambiguity resolution. If we can successfully incorporate such linguistic information, we believe that we could improve the performance of the case ambiguity resolution system.

Appendix A

The Romanisation of Korean

Through out the thesis, we follow the Romanisation of Korean Standard officially released by the Korean Ministry of Culture and Tourism[1] Specifically, we use the Romanisation method recommended in Chapter 3, Clause 8 in the standard considering an easy reverse translation.

A.1 Consonants

ㄱ	ㄲ	ㄴ	ㄷ	ㄸ	ㄹ	ㅁ	ㅂ	ㅃ	ㅅ	ㅆ
g	kk	n	d	tt	l	m	b	pp	s	ss
ㅇ	ㅈ	ㅉ	ㅊ	ㅋ	ㅌ	ㅍ	ㅎ			
ng	j	jj	ch	k	t	p	h			

A.2 Vowels

ㅏ	ㅑ	ㅓ	ㅕ	ㅗ	ㅛ	ㅜ	ㅠ	ㅡ	ㅣ	ㅢ
a	ya	eo	yeo	o	yo	u	yu	eu	i	ui
ㅐ	ㅔ	ㅚ	ㅟ	ㅒ	ㅖ	ㅘ	ㅙ	ㅝ	ㅞ	
ae	e	oe	wi	yae	ye	wa	wae	wo	we	

[1] The Ministry of Culture and Tourism Notification No. 2000-8 (7 July, 2000)

Appendix B

The KAIST Part-Of-Speech and Phrasal Tagset

B.1 Part-Of-Speech Tags

Symbols

sp	,
sf	., !, ?
sl	opening quotation mark and bracket
sr	closing quotation mark and bracket
sd	dash
se	elipsis symbols
su	unitary symbols
sy	other symbols

Foreign words

f	foreign words

Nominals

ncpa	active predicative nouns
ncps	static predicative nouns
ncn	non-predicate nouns
nq	proper nouns
nbu	unitary bound nouns
nbn	non-unitary bound nouns

npp	personal pronouns
npd	demonstrative pronouns
nnc	cardinal numerals
nno	ordinal numerals

Predicates

pvd	demonstrative verbs
pvg	general verbs
pad	demonstrative adjectives
paa	attributive adjectives
px	auxiliary predicates

Modifiers

mmd	demonstrative adnominals
mma	attributive adnominals
mad	demonstrative adverbs
maj	conjunctive adverbs
mag	general adverbs

Interjections

ii	interjections

Particles

jcs	nominative case particles
jcc	complementative case particles
jcv	vocative case particles
jcj	conjunctive case particles
jcr	quotative case particles
jco	accusative case particles
jcm	genitive case particles
jca	adverbial case particles
jct	comitative case particles
jp	predicative case particles
jx	auxiliary particles

Endings

ep	non-terminal endings

ecc	coordinate conjunctive endings
ecs	subordinate conjunctive endings
ecx	auxiliary conjunctive endings
etn	nominalisers
etm	adnominalisers
ef	terminal endings

Affixes

xp	prefixes
xsm	adjectival derivational suffixes
xsv	verbal derivational suffixes
xsa	adverbial derivational suffixes

B.2 Phrasal Tags

S	sentence
NP	noun phrase
VP	verb phrase
ADJP	adjective phrase
MODP	adnominal phrase
ADVP	adverbial phrase
IP	interjectional phrase
AUXP	auxiliary predicate phase

Appendix C

The Sejong Part-Of-Speech and Phrasal Tagset

C.1 Part-Of-Speech Tags

Symbols

SP	„ :, /, ·
SF	., !, ?
SS	quotation marks, brackets, dash
SE	elipsis symbols
SO	~
SL	foreign words
SH	words in Chinese characters
SW	logical and mathematical symbols, currency symbols
SN	numbers

Nominals

NNG	general nouns
NNP	proper nouns
NNB	bound nouns
NP	pronouns
NR	numerals

Predicates

VV	verbs

VA	adjectives
VX	auxiliary predicates
VCP	positive copula
VCN	negative copula

Modifiers

MM	adnominals
MAG	general adverbs
MAJ	conjunctive adverbs

Interjections

IC	interjections

Particles

JKS	nominative case particles
JKC	complementative case particles
JKG	genitive case particles
JKO	accusative case particles
JKB	adverbial case particles
JKB	vocative case particles
JKQ	quotative case particles
JX	auxiliary particles
JC	conjunctive case particles

Endings

EP	non-terminal endings
EF	terminal endings
EC	conjunctive endings
ETN	nominalisers
ETM	adnominalisers

Affixes

XPN	nominal prefixes
XS	suffixes
XSN	nominal derivational suffixes
XSV	verbal derivational suffixes
XSA	adjectival derivational suffixes

XSB	adverbial derivational suffixes
XR	root

C.2 Phrasal Tags

S	sentence
Q	quoted sentence followed by quotation marks
NP	nominal phrase
VP	predicate phrase
VNP	positive copular phrase
AP	adverbial phrase
DP	anominal phrase
IP	interjectional phrase
X	pseudo phrase

C.3 Function Tags

SBJ	subject
OBJ	object
CMP	complement
MOD	adnominal modifier
AJT	adjunct
CNJ	conjuntive
INT	interjection
PRN	parenthetical

C.4 Others

L	opening quotation mark, bracket
R	closing quotation mark, bracket

Appendix D

The Test Set for Human Annotation

Note: Dependencies are marked by font shapes. In the actual annotation, colour printed material was used.

1. 집에서 발생한 안전사고의 **유형은** ( ) 낙상(21.6%), 미끄러짐(15.9%), 화상(6.9%)이 주류를 **이루었다.**
2. **그는** ( ) 객관주의에서 벗어나려면 작가의 세계관을 중시해야 되며, 반대로 주관주의를 탈피하려면 객관 현실의 반영 위에 작품을 구성해야 한다고 **말한다.**
3. 아까처럼 거 **아리랑이나** ( ) 한 **번** ( ) 더 **해보려무나.**
4. **다음날** ( ) 용훈이가 **나가고** 난 **새** ( ) **현우는** ( ) 몰래 용훈이의 바이올린을 **켜고** 있었다.
5. 이같이 석궁의 보급이 늘면서 각종 안전 **사고도** ( ) **빈발하고** 있다.
6. 칠하고 나면 **샤먼(무당)은** ( ) 빗살이 일곱 **개** ( ) **난** 빗으로 양팔과 양사타구니 그리고 가슴을 긁어 상처를 **낸다.**
7. 오히려 **나는** ( ) 그 사람을 위해 이곳으로 **이사까지** ( ) **했는데** – 결국은 떠나 버린 거요.
8. 중국 고대 문헌에 **혜성은** ( ) 악기에서 **태어나는** 것으로 혜성이 날아오는 방향과 그 꼬리의 장단(장단), 빛깔이 농담에 따라 큰 바람, 큰 가뭄, 큰 추위, 지진, 재질, 병란, 흉년을 몰아온다고 했다.
9. **현우는** ( ) **울고라도** 싶었다.
10. 이런 조직이 일본엔 10만 **명** ( ) **존재하는데도** **경찰은** ( ) 손을 **대지** 못한다.
11. 순복음의 신앙은 이처럼 성경이 말하는 신천신지와 영원한 나라를 소망하며 이 땅에 사는 **동안** ( ) 최선을 다하여 주를 **섬기겠다는** 종말론적 신앙이다.
12. **정부는** ( ) 하루 속히 그 진상을 조사, 공무원의 품위를 지키지 못한 책임을 묻고 아울러 통계의 조작·왜곡여부를 **가려내야** 한다.
13. 이들 공약이 지켜질 **경우** ( ) 나라경제가 어떻게 **될지** 걱정이 앞선다.
14. 학교에서 공부하는 자녀를 화면에 비치게 하여 장난을 하는가 감시할 **수도** ( ) **있고,**

또 의처증의 남편이 집에서 아내가 무엇을 하고 있는가 감시할 **수도** ( ) **있다.**

15. **선관위** ( ) 역시 그런 법의 비현실성 때문에 가급적 현실에 맞게 운용하려 **애쓰는** 것으로 알고 있다.
16. 그 가격이 실제 가격의 두배가 되는지, 10배가 되는지는 알 **수** ( ) **없는** 일이다.
17. 그래서 사회자의 특권을 이용하여 제가 한 **말씀** ( ) **드리겠습니다.**
18. 북의 잠수함 **침투사건은** ( ) 우리 군에 새로운 기회를 **던져주고** 있다.
19. 이런 호황에서 제품의 질을 운운할 **사람은** ( ) 공급자측이나 수요자측 어느 쪽에서도 **나서지** 않을 것이다.
20. 화재가 많은 겨울철, 가정의 소화기 한대가 소방차 **열대보다도** ( ) **중요하다고** **관계자는** ( ) 입을 **모은다.**
21. 오랜 세월의 우리 전통이 무너졌다 해서 서구의 역사와 문화가 낳은 전통을 그대로 우리의 전통을 삼을 **수는** ( ) **없고** 그것은 언제나 우리 민족집단의 공동한 사고방식인 문화와 융합, 변성됨으로써만 새로운 전통이 될 **수** ( ) **있는** 것이기 때문이다.
22. 개인의 관리 소홀로 분실된 **여권은** ( ) 곧바로 이들 위·변조단 수중에 들어가 불법 입국자들의 여행증명서로 **악용될** **수밖에** ( ) **없다.**
23. 흥이 돋기를 기다렸다가 우리가 가져온 유성기를 그들 앞에 틀어 놓자 처음에는 멈칫하고 구석으로 피하더니 한 **무관은** ( ) 전후 **체면도** ( ) **잊어버리고** 유성기의 서양 노래에 박자를 맞추어 **춤까지** ( ) **추었다.**
24. **군중들도** ( ) 일제히 팔을 들고서 만세를 **불렀다.**
25. 이쪽 계통의 어떤 회사가 사진 단 한 **장(누드)만** ( ) **찍자고** 제의했다.
26. **근소세인하문제는** ( ) 정부가 이미 검토작업에 **들어가** 있는데 신한국당은 그 공제범위를 정부와 비슷한 현행 20%에서 30%로, 국민회의와 **민주당은** ( ) 50%로 **확대하겠다고** 약속한 것이다.
27. **꽁트는** ( ) **사회도** ( ) 합리적으로 **재구성하고** 사회제도들이 저마다 고유의 기능을 수행할 **수** ( ) **있도록** 계획하고 수정하는 방안을 연구하는 사회학이 학문 중의 학문이라고 **보았다.**
28. 맘놓고 불러 볼 **수** ( ) **없는** 어머니이기 때문에 남몰래 가만히 돈으로나마 불러 보려는 나의 서글픈 부르짖음으로만 알아다오.
29. 깊이 감춰뒀던 약간의 돈과 여편네가 시집올 **때** ( ) **가져온** **가락지까지** ( ) 몽땅 **털렸는데** – 그게 이상하단 말이오.
30. 조선조의 태조(태조)는 소문난 격구 챔피언이었고, **세종대왕도** ( ) 왕위를 물리고 들어앉은 아버지 태종과 더불어 타구경기를 **즐겼다는** 기록이 실록에 나온다.
31. **세계경제는** ( ) 1920년대 초부터 흔들리기 **시작했고,** 1929년의 **대위기는** ( ) 대재난 속에서 지구경제의 상호의존관계를 **드러내** 보여주었다.
32. 인간 한계의 다른 도전으로 산소부족을 들 **수** ( ) **있다.**
33. 오랜만에 만난 인사가 끝나기 바쁘게 **어른들은** ( ) 곧 명절준비로 **분주해졌다.**
34. 또한 집단대응이 실현되는 **경우** ( ) 거의 동시적으로 감원·임금동결 사태가 **일어나**

파급영향이 증폭, 정치·사회 등 경제외적으로도 불안을 확산시킬 **수** ( ) **있는** 것이다.

35. 종인 어머니고 종의 자식인 너지만, 그래도 오막살이 속에서나마 맘놓고 어머니라 부를 **수** ( ) **있고,** 맘놓고 자식이라 부르고 사는 네가 아니냐.
36. 가만히 짐작을 하니 **어머니는** ( ) 아버지의 바지 안을 따고서 그 속에다 종이를 감춰 넣고 실로 **꿰매는** 모양이었다.
37. **우리는** ( ) 이 창조의 원리를 올바로 **인식하여야** 합니다.
38. **너** ( ) 배가 **고픈** 모양이로구나.
39. 손에 손에 든 **태극기는** ( ) 파도처럼 **나부끼고** "만세!" "만세!" 하고 외치는 함성 **소리는** ( ) 땅을 우끈우끈 **흔들었다.**
40. 얼마를 지나서 **음악은** ( ) 뚝 **끊어졌다.**
41. 공장지배인들이 이들로부터 뒷돈을 챙기는 **것은** ( ) **말할 것도** ( ) **없다.**
42. 또 사물이 굴절돼 제대로 보이지 않는 **제품도** ( ) 조사 대상의 **20%나** ( ) **됐었다.**
43. 그러나 야조프 등이 비행기에 올랐을 **때** ( ) **고르비는** ( ) 또다른 비행기에 **올랐으며** 이때를 놓치지 않고 루츠꼬이가 잽싸게 야조프와 클류치꼬프 등에게 수갑을 채웠다.
44. 우리의 ㅇ은 ㅅ다음이지만 북의 **ㅇ은** ( ) ㅎ다음 모음에서 **시작한다.**
45. **영재는** ( ) 넋없이 매남산을 **바라봤다.**
46. 미국과 러시아가 인류를 몇 **번이라도** ( ) **몰살시킬 수** ( ) **있을** 만큼의 핵보유고를 감축시키려는 노력을 하고 있음에도 불구하고 **핵무기는** ( ) 소형화되어 세계곳곳에 **분산되어** 있다.
47. 몇 년 또는 몇 **달씩** ( ) **걸린** 재판의 판결문이 엉망이라면 이 세상에 마음놓고 믿을 **수** ( ) **있는게** 무엇이 있겠는가.
48. 아닌게 아니라 죽당 선생의 말씀과 같이, 우리 나라 **백성은** ( ) 귀가 있어도 바른 말을 듣지 못하고, 입이 있어도 바른 소리를 **하지** 못하고 있다는 것이 깨달아졌다.
49. 소매의 곡선과 버선의 **곡선은** ( ) 기와집 추녀끝의 곡선과 **같고** 치마의 **주름은** ( ) 서까래와 **같아** 한인의 곡선미가 가옥에 나타난 것이 한국 기와집이고, 의복으로 표현된 것이 한복이라는 것이다.
50. 분명 악의가 엿보이기는 하지만 그가 태어나자마자 걷고 말할 **수** ( ) **있었다고** 하는 전설은 아마도 그같이 남다른 그의 재질에 근거하여 꾸며졌을 것이다.
51. 체코 **민주화** ( ) **이끈** 지식인 방한/
52. 전철 일산선을 이용하는 **주민들은** ( ) 매일 **아침** ( ) 발을 동동거리며 분통을 **터뜨린다.**
53. 야차같은 수양으로도 미친 **녀석** ( ) **같은** 김시습을 어떻게나 모셔 보려 애를 쓴 것은 무언가?
54. 게다가 각 정당이 거창하게 내건 **공약도** ( ) 언론 등의 분석으로 대부분 실현불가능한 껍데기 공약임이 **드러나고** 있으니 국민 **실망은** ( ) 더 **클 수밖에** ( ) **없다.**
55. 동회(동회)의 도움과 몇 번의 수소문 끝에 8년 전의 민요섭을 잘 아는 사람을 **하나** ( ) **찾아낸 남경사는** ( ) 그로부터 지금껏 들어온 것과는 전혀 다른 민요섭의

일면을 **들을 수** ( ) **있다.**

56. **그들은** ( ) 태초에 어떤 집에서 살았고, 무엇을 먹었으며, 살림을 어떻게 **꾸렸는가.**
57. 동화 전체의 강한 **지향성은** ( ) 아무래도 도덕과 윤리, 인간애 등에 **모아져** 있기 때문입니다.
58. 우선 시란 예술의 한 양식이며, 예술은 그 존재 의의가 의도나 사상만으로 규정될 **수** ( ) **없다.**
59. 기지부근에서 번식하는 **두종류는** ( ) 모두 알을 **두개씩** ( ) **낳아서** 대개는 **둘다** ( ) 부화되나 **성장과정에서** 상당수가 죽는다.
60. 그 노래소리를 들으면 **천하장사도** ( ) 홀리지 않을 **수** ( ) **없어** 이끌려가서는 피를 빨리고 혼을 박탈당한 **채** ( ) 무인도에 **버림받는다.**
61. **이때** ( ) **문제되는** 내적 자격이란 동인지의 주체가 될 사람들의 창작역량이라든가 잡지 편집능력, 문학활동의 정신내용을 이루는 이데올로기 등이다.
62. 따라서 의식적인 **선택은** ( ) 시간성 안에서 **행해짐으로** 인해서, 싸르트르에 있어서 인식과 **시간은** ( ) 긴밀한 관계를 **갖는다.**
63. **상투** ( ) **짠 젊은이는** ( ) 상투를 끊고 **일어서고,** 가난한 집 **소년들도** ( ) 행장을 꾸려 메고 공부를 하러 집을 **나섰다.**
64. **현우는** ( ) 갑자기 하늘이 노래지면서 뒤로 나동그라지려고 **했다.**
65. 새 청와대 **본관은** ( ) 옛 기맥을 되살린다는 뜻에서, 북악산정과 경복궁 - 광화문 머리로는 관악산을 잇는 축상에 **세워졌다.**
66. 매달린 사람이 틀림없는 아버지인 것을 살펴낸 **현우는** ( ) 그 자리에서 눈을 홉뜨고 **까무러쳤다.**
67. 인간의 **유전자연구는** ( ) 생명의 존엄성을 훼손한다는 윤리적인 면 때문에 어느정도 질서가 **잡혀** 있다.
68. 삼공 **벼슬** ( ) **준다** 한들, 이 강산을 놓을소냐.
69. 만의 하나라도 이번 **합의조차** ( ) 비등하는 여론에 밀린 일시적 당략으로 **끝내려** 했다가는 앞으로 정치권의 설 **땅은** ( ) 그만큼 **좁아질** 것이다.
70. 그러나 **원장은** ( ) 화들짝 놀라며 목소리를 **높였다.**
71. 이들 장소에서 파는 **태극기는** ( ) 가정용의 **경우** ( ) 가로 90㎝ 세로 60㎝ 등의 여러 가지 크기가 나와 있고, **깃대는** ( ) 2단과 3단형이 **선보이고** 있다.
72. 정보통신기술에 대한 **이해방법도** ( ) 시급히 **재고되어야** 한다.
73. 이 95 개조의 **반박문은** ( ) 처음엔 라틴어로 **씌여졌다.**
74. 당초 기준대로라면 삼성·현대·LG·대우그룹 등 4대 통신장비 **제조업체중** ( ) 2개 그룹이 신규사업권을 따낼 **수** ( ) **있게 돼있었으나** 이를 바꿔 그중 한 사를 비통신장비 제조업체의 몫으로 돌린 것이다.
75. **껄찬은** ( ) 국가적 체면을 위해서나, 군의 명예를 위해서나, **증폭만** ( ) **되어** 가는 국민적 의혹을 가라앉히기 위해서도 수사에 더욱 **밀도** ( ) **있는** 노력을 **기울여야** 하겠다.

76. **기념비는** ( ) **무너지고,** 국가는 ( ) 사라지고**, 문명은** ( ) **쇠퇴하여** 암흑기가 있은 **다음** ( ) 새로운 민족이 다른 문명을 **세운다.**
77. 신 인사제도를 도입한 우리 나라 기업체의 **경우도** ( ) 남직원은 기획, 입안 등 총괄적 관리를 맡는 일반직으로, **여성은** ( ) 서무, 경리직원으로 **나눠지고** 있다고 윤정숙씨(여성민우회 사무직 여성부장)는 말한다.
78. 물론 일부의 **학설은** ( ) 극단적인 편향을 기피하여 절충·중도적 성격의 이기설로 이루어진 **것도** ( ) **없지** 않았다.
79. 이간에 의하면 본성은 오상(인의예지신)으로 얘기되지만, 근본적으로는 우주의 근원인 태극으로서의 리(성즉리)이므로, 모든 사물이 다 태극으로 말미암아 생겨난 이상, 인성과 **물성은** ( ) 서로 **같다는** 것이다.
80. 유전자를 조작하거나 세포융합 조직배양 미생물 이용 등 바이오테크놀로지에 의한 **품종개량은** ( ) 식량 및 에너지부족 환경문제 해결에 결정적 역할을 **할** 것으로 기대되고 있다.
81. **미국도** ( ) 남미, 아시아, 아프리카 등에서 책을 통해서 미국의 이데올로기나 관습, 제도, 생활방식 등을 **유포하고** 있다.
82. 고문헌을 찾고, 떡을 직접 만들어 사진을 찍으며 준비를 했다고 **윤숙자교수는** ( ) **말한다.**
83. 자기를 속이고도 마음의 아픔을 느끼지 않는 사람은 무슨 짓이든 할 **수** ( ) **있는** 사람이다.
84. 한국의 **경우는** ( ) 전술한 것처럼 소비지출의 규모가 연령과 함께 증가하다가 45 54세에서 정점에 이르고 그 **이후** ( ) **축소되는** 모습을 **보이고** 있다.
85. 그럴수록 **현우는** ( ) 계속 **밤마다** ( ) **빌기를** 잊지 않았다.
86. 또 **23일** ( ) 대만의 총통선거가 **끝난** 후에도 군사훈련을 계속, 목조르기를 늦추지 않을 생각이다.
87. 강물이 더러워지고 개천물이 더러워지면 **잉어도** ( ) 못살 **뿐 아니라** ( ) **우리** 입으로 들어오는 물도 **더러워짐** ( ) **수** 밖에 없다는 진리를 뼈저리게 깨달아야 하겠습니다.
88. 배가 결리고 숨이 막혀서 한참 **동안** ( ) **현우는** ( ) 한쪽 손으로 배를 움켜쥔 **채** ( ) 방네 구석을 **기었다.**
89. 정서면에서나 경제적인 면에서 우리 고유의 캐릭터를 만드는 일이 시급하다고 **전문가들은** ( ) **아쉬워한다.**
90. 이러한 유통과정을 거쳐 최초에 투입되었던 화폐자본이 상품자본으로 그리고 증대된 화폐자본으로 다시 돌아오는 자본의 운동과정을 거치며출판산업도 **출판산업도** ( ) 자본축적을 **마련한다.**
91. 분쟁의 **핵심은** ( ) 영국이 최근 신임 홍콩 총독에 보수당 의장을 역임한 **바** ( ) **있는** 원로 정치인 크리스 패턴을 임명하면서 제기한 몇가지 중요한 제안에 **있다.**
92. 정당활동을 가장한 과도한 돈쓰기와 이른바 전략지구에 대한 당차원의 **과잉지원도** ( ) **자제해야** 한다.

93. 고티의 재판이 진행된 지난 2 개월 **동안** ( ) 미국의 TV **시청자들은** ( ) 감비노가의 암투, 배신, 범죄 등을 마치 영화 '대부'의 속편을 보듯 흥미진진하게 재판 과정을 **지켜봤다.**
94. **정부당국은** ( ) 이번 회담에서 이 점을 미국에 **설득력** ( ) **있게** 설명할 **수** ( ) **있어야** 한다.
95. 양귀비의 **신발도** ( ) 그에 못지않게 값나가는 신으로 **알려져** 있다.
96. 그러므로 소부 · 허유가 사실로 있었거나 없었거나, 자릉이 정말 광무의 배때기를 눌렀거나 아니 눌렀거나, 디오게네스가 과연 알렉산더 눈깔을 쏘아보았거나 말았거나, **그것은** ( ) 문제가 **아니다.**
97. 앙드레 **모르와는** ( ) 연애하는 여성들에게 이런 충고를 **하고** 있다.
98. 동양과 중국분위기의 **옷만** ( ) **파는** 의류회사가 성업중일 정도.
99. **창문** ( ) **열고** 자는 **습관** ( ) **바꿔야**
100. 문간에서 **표** ( ) **받고** 앉아 있던 젊은 사내가 내다 소리를 쳤다.
101. 연탄 몇 장이 없어 차가운 방에서 신음하는 병든 **노인들도** ( ) 한두명이 **아닙니다.**
102. 그러나 그동안에도 18·19세기에는 계몽철학 이후로 리성주의가 성행하던 시기가 있었고, 20세기에 들어서서는 또 생의 철학을 뒤따라 프로이트의 심리학이니 실존주의 사상이니 하는 새로운 경향이 생겨서 **이제는** ( ) 이성보다도 인간의 비합리적인 면, 즉 정의·의욕적인 면을 더 강조하는 풍조를 이루게 **되었다.**
103. **이** ( ) **같은** 인력난 때문에 일손을 구하지 못한 일부 아파트 단지내 중국음식점 등에서는 주인이 자신의 승용차로 음식을 배달하는 **경우도** ( ) **생겨나고** 있다.
104. 또 무르깁도로 공사장 등 현대건설 현장사무소 관리를 위임받은 유재성 **씨는** ( ) 이라크군이 크레인 불도저는 물론이고 승용차 **타이어까지** ( ) **빼갔지만** 속수무책이었다고 **말했다.**
105. 그러나 **이것은** ( ) 정부가 조장할 것을 공언해 온 은행의 자율경영 정책과 정면으로 **위배된다.**
106. 청년문화 속에서 새롭게 관심의 대상이 된 우리의 **전통문화도** ( ) 철저한 반성과 재구성을 통한 변혁을 거쳐야 비로소 현대문화 속에 참여할 **수** ( ) **있게 된다.**
107. 합참의장·국방장관이 5년간 브로커에게 질질 끌려 다닌 셈인 이씨 **사건은** ( ) 국민과 군의 권위에 엄청난 상처를 주었을 뿐더러 문민정부의 위상과 도덕성에도 큰 상처를 **주었으며** 특히 인사능력에 결정적 흠을 드러낸 것은 **가슴** ( ) **아픈** 일이다.
108. 조선조에서 이루어진 성리학의 역사적 응용과 이론탐구야말로 한국 성리학의 진면목이라 할 **수** ( ) **있다.**
109. 지난 **번** ( ) LA올림픽을 **유치할 때** ( ) IOC가 어찌나 **까다롭게** 굴었던지 미국의 **저넌리즘은** ( ) "IOC는 무기력한 귀족과 맥빠진 노인들의 집단이며, 현대로부터 유리된 완전한 궁정(궁정)살롱이다"라고 **비아냥거렸었다.**
110. **약도** ( ) **올랐고 호기심도** ( ) **일어나고** 하여 **그는** ( ) 부하를 데리고 **디오게네스** ( ) **있는** 곳을 찾아 **갔다.**

111. 아하스 페르츠가 어떤 길을 따라 이집트로 갔으며 그가 **처음** ( ) 발을 **디딘** 도시가 어디였는지는 잘 알 길이 없다.
112. 나무 위에 올라가 거울로 **밭** ( ) **매는** 순이의 얼굴을 비추는 돌이의 장난 정도가 아니라 러시아에서는 지름20m의 거울을 지상 3백 50 km의 우주 공간에 올려 지름 5 km의 지상을 보름달 수개의 밝기로 반사시키는 **데** ( ) **성공했다** 한다.
113. **상호도** ( ) **올라가고,** 밤마을 **종수도** ( ) **올라가서** 몸이 닳아 못 견디도록 부러운 편지를 전해 오고 있다.
114. 곧 **이들은** ( ) 부모나 어른에 대한 존경심 등 가족주의적 전통규범이나 가부장적 의식을 지키고 있으며, 권위주의적인 공동체의 질서에 **순응하고** 있다.
115. 그러나 만약 정부규제로 인해 기업의 생산비용이 높아지는 것이 그 한 요인이라면 **이는** ( ) 결코 **바람직하지** 않다.
116. **오늘날은** ( ) 옛날보다 물질적으로 풍요를 **누리고** 있을지도 몰라도 실제 생활에서 행복을 누리며 사는 **사람은** ( ) 별로 **많지** 않다.
117. 거기다가 늙은이가 다시 그렇게 확인하자 **남경사는** ( ) **처음부터** ( ) **궁금하던** 것을 묻기 **시작했다.**
118. 그런데 이러한 세부에 대한 **집착은** ( ) 소설 속에서 인물과 환경의 부조화상태를 통해 **나타난다.**
119. 해괴한 **것은** ( ) **범인은** ( ) 버젓이 거리를 **활보하고** 있는데도 배후조종 **세력은** ( ) 여전히 베일에 **가려져** 온 것이다.
120. 한국군에 대한 작전통제권이 미국에 이양된 **것은** ( ) 6.25전쟁 직후인 1950년 7월 **15일** ( ) **이루어졌다.**
121. 왜냐하면 다른 어떤 **족속도** ( ) 그와 같이 질투와 분노와 변덕의 신을 **거들떠보지** 않았는데 오직 우리 조상들만이 그를 받아들였기 때문이다.
122. 이리하여 **뒷날** ( ) 참으로 다시 오나라를 **쳐엎치는** 데에 성공을 했던 것인데, 말하자면 '와신상담'이란 장작 위에 눕고 쓸개를 맛보면서까지 장차의 원수를 갚기 위하여 이를 악물고 괴로움을 참아 나간다는 뜻이었다.
123. 하지만 많은 살생과 파괴가 따르는 전쟁이 싫어 **전쟁** ( ) **아닌** 다른 방법으로 승부를 가렸던 사례가 역사에 비일비재하다.
124. 방콕에 함께 머무르면서도 등을 돌릴 것 같던 한·일정상이 **2일** ( ) 결국 마주 **앉았다.**
125. 지난 1년의 **격동기도** ( ) **어려웠지만** 앞으로가 더 어려우리라는 것이 일반 시민이나 전문가들의 일치된 전망이다.
126. 인간이란 **초생물체는** ( ) 정신의 생명이라든가 신화의 생명, 사고의 생명, 의식의 생명 등 생명의 새로운 영역들을 **창조해냈다.**
127. 첫 아이 출산 **후** ( ) 임신을 **기다리던** **김씨는** ( ) 지난 **8월** ( ) 잡지 광고를 보고 임신진단시약을 구입, 소변검사를 한 **결과** ( ) 임신이 **아닌** 것으로 나타났다.
128. **케니는** ( ) 두 **발** ( ) **없음이** 슬픈 일이 아니라 조금 불편할 따름이라는 장애철학을 천진난만하게 구현해 **보였다.**

129. 가장 **인기** ( ) **있는** 색은 깨끗함이 돋보이는 흰색.
130. 질병이 인간 생활에 있어 기본적인 고통이요, 죽음의 전 단계라는 점에서 모든 **인간은** ( ) 병을 **두려워한다.**
131. 예컨대 도토리나무의 **특성들은** ( ) 그 나무가 도토리에서 발전되는 과정을 기술함으로써 완전히 **설명될 수** ( ) **있다.**
132. 이때 소련으로부터 도움을 받아 1924년 11월 **26일** ( ) 몽골 인민 공화국을 **선포하기에** 이르렀다.
133. 이 도핑으로 기록이 얼마나 향상되는가에 대한 조사 연구된 바로는 남자의 경우 트랙 경기에서 2-7 %, **중량경기는** ( ) 19-27 **%** ( ) **향상된다고** 한다.
134. **자금은** ( ) 이 사업의 실패로 5천만 원으로 **줄어들고** 말았다.
135. 1백20년 전인 **1866년** ( ) 아산만에 배를 대고 통상을 **강요했던** 프러시아 상인 오페르트가 당시 해미현감과 대령계급의 무관을 배 안에 초청하여 브랜디와 포도주로 향응을 베풀고 거나해지자 이 유성기를 틀어놓았던 것 같다.
136. **연인들은** ( ) 노래를 녹음해서 승용차안에서 함께 **듣는답니다.**
137. **아이들은** ( ) 모두 숨소리를 죽이고서 **눈들만** ( ) 커다랗게 **뜨고 두리번거렸다.**
138. 임부복 전문 업체인 몽실의 경우 2~3년전만 해도 레이스 등으로 귀여운 맛을 강조한 홈웨어풍의 임부복이 많이 나갔으나 요즘은 정장 개념으로 블라우스와 치마, 반바지, 긴 바지, 조끼 등을 맞춰입을 **수** ( ) **있는** 단품이 홈웨어의 2배 이상 나간다는 것이다.
139. 우리 나라 경제의 주력산업인 자동차, 조선, 기계 **등도** ( ) 치솟는 인건비로 해외경쟁에서 **고전하고** 있고 섬유, 신발, 완구류, 식기류 등 경공업 **제품들은** ( ) 이미 고임금의 중압을 이겨 내지 못해 대다수의 기업이 해외로 이전했거나 아니면 **폐업한** 것이다.
140. 어떻게 당신이 비상위원회를 설치할 **수** ( ) **있는가?**
141. 진 나라 **효무제는** ( ) 혜성이 나타나자 천제가 내린 축수배라 하여 잔치를 **베풀기까지** 했다.
142. **현우는** ( ) 아버지가 걱정되실까 봐 입을 **다물었다.**
143. 모든 음식이 **제철** ( ) **나는** 것으로 만드는 것이 **맛** ( ) **있듯, 떡도** ( ) 철이 **있어요.**
144. 건강한 **신앙인은** ( ) 자신의 욕망과 욕심을 절제하면서 진실로 사람답게 살려고 **노력한다.**
145. 이러한 모든 과정을 거친 스승이 그 제자가 올바른 깨달음을 얻었는지 아닌지 그 여부를 직관하여 판단할 **수** ( ) **없다면,** 오히려 이상한 일이다.
146. **나이** ( ) 육십이 **넘은** 고동 양반의 아버지였다.
147. **당신은** ( ) 태어날 **때부터** ( ) 죄가 무엇인지 **알고** 있었나요?
148. 나한이 개인적 자각인 **데** ( ) **대하여,** 보살은 사회적 자각에 입각한 것이니, 나한은 언제든지 개인 본위이고 개인 중심주의인 **데** ( ) **대하여,** 보살은 사회 본위이고 사회 중심주의인 것이다.
149. **박씨는** ( ) 지난 **11일** ( ) 아들을 **찾아가** 함께 밤을 **보냈다.**
150. 현대 **심리학은** ( ) 지식이 언어나 문자 이외에 다른 방법으로 충분히 전달될 **수** ( ) **있**

다고 말한다.

151. 그런데 그들 세 사람의 현명을 더욱 의심케하는 것은 야훼께서 아들의 궁색한 산실로 빌어쓰던 마굿간에 이를 **때까지** ( ) 그들이 **보여준** 지각없는 언동이었다.
152. 그것을 **나는** ( ) 도시에 사는 사람 즉, 시민들의 생활 속에 흐르고 있는 질서감 혹은 생활 질서 속에서 **찾아보고** 싶다.
153. 킹조지섬은 남극이지만 비교적 북쪽이어서 **바다만** ( ) **얼지** 않는다면 상당수의 새들은 볼 **수** ( ) **있다.**
154. 우선 보건복지부산하에 안전본부를 두는 것으로 출발하지만 법령정비·검사제도개편 등의 준비를 거쳐 **97년중** ( ) 독립된 외청으로 **발족시킨다는** 것이 복지부의 계획이다.
155. **갈브레이스는** ( ) 당시 관계자의 말을 인용, 투기 열풍을 다음과 같이 **묘사하였다.**
156. 어떻든 **그는** ( ) 범행 **후** ( ) **한때** ( ) 특정세력의 반짝 비호를 **받았으나** 평생 ( ) 머리를 **짓누르는** 죄악의 쇠사슬에 얽힌 **채** ( ) 추격과 테러공포 속에 **지낸** 것은 당연한 업보다.
157. 닭고기와 닭 가공 식품을 산매 가격보다 **30%정도** ( ) **싼** 값에 살 **수** ( ) **있는** **할인매장도** ( ) **운영된다.**
158. 이 **말은** ( ) 자기 자신을 심리적으로 압박하는 국가의 명예나 자신의 명예, 그리고 **승부욕** ( ) **같은** 것을 전혀 염두에 두지 않고 뛴 것이 바로 금메달을 타게 한 요인이 되었다는 것이 **된다.**
159. 둘째, 무엇보다도 확실한 성령 세례의 **증거는** ( ) 강력한 복음 전파에 **있다.**
160. 마치 버스기사·업자·행정당국이 짜고 하는 **것** ( ) **같은** 요금인상과정의 되풀이에 **시민들은** ( ) 지치고 부아가 **치민다.**
161. **농협도** ( ) 전국 1천곳에서 국군의 날인 10월 1일 전에 판매를 **시작한다.**
162. 방안을 휘둘러본 **남경사는** ( ) 곧 민요섭이 가지고 있던 물건들을 **찾았다.**
163. 페어 **플레이는** ( ) 인간의 성실과 관용의 정신, 기회균등을 존중하는 정신의 **나타남이다.**
164. 그것을 **맹자는** ( ) 인의예지의 성이라 **하였다.**
165. 이 시기에 **일제는** ( ) 파시즘을 강화하여 각종 사회운동을 **핍박하는데,** 그러한 사상적 **탄압은** ( ) 문학과 예술에도 직접적인 영향을 미쳐 1935년 **5월** ( ) 카프의 해산을 **가져오기에** 이른다.
166. 개벽사상에 따르면 과거 선천의 **시대는** ( ) 운이 **다하였기** 때문에 이제 새로운 후천의 시대가 도래한다.
167. **그때** ( ) **필자는** ( ) 분명히 **저것은** ( ) 정성이 **아니라** 고집이라는 사실을 **깨닫게** 되었다.
168. 서양언어에 대한 **지식은** ( ) 가장 많은 노력을 기울이는 중요한 학습의 대상이요, 무슨 일에서나 탁월한 능력의 조건으로 **비쳐졌다.**
169. 그런 의미에서 방콕의 정상회담에서 배타적 경제수역(EEZ)의 경계선 획정교섭을 독

도문제와 분리하기로 한 **것은** ( ) 현실적인 해결책으로 **볼 수** ( ) **있다.**

170. 사림파의 이 학통관 즉 **도통관은** ( ) **실제** ( ) 학문의 전수관계나 학문의 업적만으로 설정되고 인정되는 것이 **아니다.**
171. **이는** ( ) 양국의 오랜 우호관계로 볼 **때** ( ) 너무나 **당연한** 것이다.
172. 곧 효의 **규범도** ( ) 전통적 의미와 실천형식들이 **상당부분** ( ) **제거되면서** 새로운 도덕적 의미와 양식의 창조를 통하여 전통문화를 건전하게 계승하는 것은 청년문화가 추구해야 할 과제이다.
173. 세침흡인 **검사는** ( ) 스웨덴을 비롯, 유럽 - 미국 등에선 이미 **50년대부터** ( ) 유방암, 전립선 암 등의 진단에 흔히 사용되는 간편하고도 정확도가 높은 종양 진단법으로 **평가** ( ) **받고** 있다.
174. 그래서 고대 올림픽에서 **여성은** ( ) 선수는커녕 관중으로서도 경기장에 들어가는 것을 **금지** ( ) **당했었다.**
175. **그는** ( ) 또 그러한 미국의 협상노력이 실패하게 된 **것은** ( ) 광주시민들 안에 무기를 버리길 거부한 일파가 **존재했기** 때문에 이른바 이들 과격파에 유혈사태의 책임을 **물었다.**
176. 갑자기 눈에 불을 켠 **헌병들은** ( ) 총부리로 마을 사람들을 돌려 세워 골목길로 몰아 **내려가고,** 나머지 **헌병들은** ( ) 거미같이 사방으로 흩어지면서 마을 안팎을 샅샅이 뒤지기 **시작했다.**
177. 그러다가 **그녀는** ( ) 문득 심상찮은 느낌이 들었는지 의심스레 **물었다.**
178. 꼬집어 말할 **수는** ( ) **없어도** 아들과 민요섭의 그같은 접근은 피할 **수** ( ) **있는한** 피해야 할 악연(악연)이란 느낌이 들었던 것이다.
179. 1962 년 선거를 거쳐 1963 년 2 **월** ( ) **등장한** 보쉬정권하에서 **미국은** ( ) 군부. 경찰 자본가집단과 긴밀한 유대를 가지면서 노동운동의 조직화를 철저히 억압토록 **했다.**
180. 그러노라니 흥분에 흥분이 겹쌓여서 **현우는** ( ) 가슴이 부풀어오르는 것을 억제할 수가 **없을** 지경이었다.
181. 그리고 일본의 젊은 **신입사원들은** ( ) 면담과정에서 선 수행 경력을 **질문** ( ) **받는다고** 한다.
182. 선수없이 **국기만** ( ) **들고** 나왔지만 올림픽이라는 마법의 장에서는 미국이나 소련과 대등한 것이다.
183. 고대는 물론이고 중세기를 통하여 책을 읽는 **사람들은** ( ) 반드시 크게 소리를 내어 **읽었다.**
184. 마케도니아의 한 절반 야만인의 자식인 **알렉산더는** ( ) 천하를 정복할 적에 당시 문화의 동산인 그리이스를 말발굽 밑에 두루 **짓밟았다.**
185. 일본인에게 일본도로 죽는다는 **것은** ( ) 총이나 독약으로 죽는 것과는 전혀 다른 분위기와 충격이 **있는** 것이다.
186. 그의 시장경제를 위한 급진적인 가격 **자유화는** ( ) 가게 앞에 서있는 긴 줄을 줄이는 데는 성공했지만 국민의 생활 수준을 더욱 **떨어뜨렸다.**

187. 일행 중에 야나예프와 **뿌고는** ( ) **빠진** 것으로 밝혀졌다.
188. 보도사진을 보니 이 **소녀는** ( ) 골인하자마자 **쓰러지고** 있다.
189. 그러나 페로 **자신은** ( ) 아직까지 명확한 태도 표명을 **하지** 않고 있다.
190. 이 문제로 고민하는 **사람은** ( ) 우선 서점에 **가라.**
191. 나무라지 않는 게 아니라, 나중엔 **훈장님까지** ( ) 타작마당에 나가서 줄을 그렇게 드려서야 쓰느냐고 손수 대들어서 **가르쳐** 주시기까지 했다.
192. 그가 내세운 본격소설은 우리 문학사적 맥락에서 살아잇는 개념이었으며 당대에 그것을 주장한 **것은** ( ) 오히려 임화의 올바른 역사의식을 **뜻하는** 것이었다.
193. **요금은** ( ) **해마다** ( ) **오르면서도** 그때마다 운휴위협에까지 시달려야 하는 짜증나는 **현실은** ( ) 행정이 책임지고 **바로잡아야** 한다.
194. 50년대 **중반** ( ) 프랑스가 인도차이나에서 **물러난 후** ( ) 이제 식민모국이었던 유럽이 경제협력파트너로 다시 **돌아오고** 있다.
195. 문득 몸을 일으켜 휘적휘적 떠나가며 남긴 사내의 그같은 **말은** ( ) 이미 거부할 **수** ( ) **없는** 명령과도 **같았다.**
196. 현재 직장생활을 하고 있는 **여성들은** ( ) 집안 식구들의 이해가 **깊다는 것도** ( ) 이 **조사결과** ( ) **드러났다.**
197. 명분론적 합리주의의 **사고는** ( ) 달리 말해 객관적 경험사실과 관계없는 순전한 규범적 합리주의에 **불과함을** 알아야 한다.
198. 수십 명이 죽거나 다치고 수백 명이 형무소로 **계속** ( ) **빨려들어가서** 조직 자체가 위태로워진다.
199. **당장** ( ) 국제경기로 **가꿀 수** ( ) **있는** 전통스포츠로 겨루어 승부를 내는 씨름을 들 **수** ( ) **있고,** 굴러서 보다 높이 날기를 겨루는 그네, 반동으로 보다 높이 오르기를 겨루는 널뛰기, 인원수를 제한해 잡아당기는 줄다리기도 국제 스포츠화할 수가 있을 것이다.
200. 이것으로 그의 반계급 민족문학론이 어느 **정도** ( ) **변명될 수** ( ) **있었던** 것이다.
201. 영 **불** ( ) **앞세운** 속전 가능성
202. 변혁에 필요한 합목적적 차원인 이러한 **의식은** ( ) 물질적 삶의 모순으로부터 **설명되어야** 한다는 것이다.
203. 제3의 **가설도** ( ) **들어봐야** 할 것이다.
204. 그 **격과** ( ) 실재할 **리도** ( ) **없고** 실재할 **수도** ( ) **없는** 평균치적 인간이라는 것을 과학적이라고 하는 조작에 의해서 **탄생시킨다.**
205. 이렇게 볼 **때** ( ) 조어도 사태는 동북아의 양대 강국인 중국과 일본의 민족주의의 경연장이며 이들이 벌일 패권경쟁의 서곡이라고 **할 수** ( ) **있다.**
206. 억지로 **술** ( ) **권하지** 맙시다
207. 한 집에서 화재감지기가 작동하면 경비실과 실외의 소화전에서 동시에 경보음이 나므로 자칫 엉뚱한 소동이 일어날 **수도** ( ) **있다.**
208. 그 **동안** ( ) 국내외에서 틈틈이 **소개됐던** 요오드란, 유정란, 영양란 등 특수 달걀(란)

과 난유, 닭고기 소시지, 치킨 버거와 같은 닭고기 **가공품도** ( ) 모두 **모였다.**

209. 이것은 인류가 원시인이었을 **때** ( ) 야수를 만나거나 **하면** 나무에 기어오르기 쉽게 하기 위한 조건반사 작용이라 한다.
210. 기능미화중앙협 **작년** ( ) **이어** 두번째…
211. 야훼의 말씀과 율법에 대해, 선지자의 가르침과 예언에 대해, 열왕(열왕)들과 판관들의 신앙과 행적에 대해, 모든 율법학자들의 주석과 해설에 대해, 여러 믿음에 찬 노래들과 묵시문학에 대해, 성전과 회당에서 이루어지는 모든 제례와 의식에 대해, 그들 민족의 삶을 지배하는 여러가지 규율과 관습에 대해, 그보다 더 많은 것을 배우고 기억하는 **젊은이는** ( ) **아무도** ( ) **없었다.**
212. 그 **동안** ( ) 가입문제를 놓고 정치권, 재계, 학계 사이에서 **실익은** ( ) **적고 부담만** ( ) **크다는** 이유에서 현 단계에서의 가입에 **반박도** ( ) **컸다.**
213. 조일알미늄 주의 **주가는** ( ) 1만5천 원대로 **한차례** ( ) **내려가더니** 다시 맹반발, 2만5천 **원대까지** ( ) **튀어올랐다.**
214. 이런 점에서 한국성리학은 비록 다른 분야의 **연구도** ( ) 중국 등에 견주어 **뒤지지** 않지만, 사단칠정론을 중심으로 한 심성설 위주의 탐구를 그 특색으로 꼽지 않을 **수** ( ) **없는** 것이다.
215. 그리하여 르네상스 시대의 인간이란 정감·욕망·세속적 의욕을 강조하는 '정의적' 혹은 '의욕적' 인간이라고 부를 **수** ( ) **있다.**
216. **너** ( ) **보고** 싶다고, **어머닌** ( ) 돌아가실 **때까지** ( ) **너만** ( ) **부르고** 계셨어.
217. **우리는** ( ) 일선 금융기관에서 금융실명제 위반 유혹에 빠지기 쉽다는 것을 **알고** 있다.
218. **역사란** ( ) 기억하기 위해 **존재한다.**
219. 그러나 최근 수입실적의 내용을 살펴보면 **낙관만** ( ) **하기에는** 아직 이른 실정이다.
220. 물론 그 사이에 냉전체제가 미국의 승리로 끝나는 지각변동이 발생, 한반도 정세가 크게 변한 **것도** ( ) **있다.**
221. **우리나라도** ( ) **예외는** ( ) **아니었다.**
222. **실체는** ( ) **잊혀지고** 가격 **상승만** ( ) 문제가 **되었다.**
223. 이 점으로 볼 **때** ( ) 성령 세례의 표적 중의 **하나는** ( ) 방언이라 **할 수** ( ) **있다.**
224. **청년은** ( ) 친절하게도 당직으로 보이는 직원 서넛만이 커다란 석유난로가에서 잡담을 나누고 있는 **방까지** ( ) **안내해** 주었다.
225. 이런 활동을 할 **수** ( ) **없는** 원외출마예정자로서는 현저히 불리할 **수밖에** ( ) **없고,** 선거운동의 기회균등의 원칙에 어긋남은 더 말할 **것도** ( ) **없다.**
226. 그렇게들 많은 벼슬아치들이 선정을 베풀었다면 왜 우리 조상들은 그토록 지지리도 못 살고 남부여대 피난하는 **데** ( ) 이골이 **났으며** 끝내는 **나라마저** ( ) **빼앗기고** 갖은 수모를 다 당해야 했단 말인가.
227. 1517 **년** ( ) 마르틴루터에 의해 **시작된 종교개혁운동은** ( ) 전유럽의 유럽의 정신세계를 **뒤흔들어** 놓았다.

228. 어느 한 **외국기자는** ( ) 이 뽐을 두고 동방에서 용 한마리가 탄생했다고 **했다.**
229. 오늘날 우리 주변에는 **자기** ( ) 혼자 **예수** ( ) 잘 **믿는** 사람들이 많습니다.
230. 반면에 혜성이 나타나면 불길한 조짐으로 받아들였던 **것** ( ) 또한 동서가 **다르지** 않았다.
231. 이러한 과정에서 한국 **교회는** ( ) 6.25를 **맞이하게** 되었다.
232. 어떤 **사람** ( ) **같아** 보였느냐구요?
233. **지금까지** ( ) 두어 달 **이상은** ( ) **살고** 있으나 이번 겨울을 넘길지는 의문이다.
234. 스포츠 공시학을 위하여 이 접근방법에 대한 **검증은** ( ) 추천할 가치가 매우 **높지만** 이론의 추상성에 근거하여 실증적 **문제제기는** ( ) 매우 **어렵다.**
235. 김남천이 리얼리즘에 대한 자세한 **개념규정** ( ) **없이** 이렇게 강조했을 **때** ( ) 그 바탕에는 엥겔스의 발자크론이 **놓여** 있었다.
236. 한편 소비지출에 대한 식료품 **비중은** ( ) **한국은** ( ) 30대 중반에서 가장 **높고 여타국가는** ( ) 30대 후반에서 40대 중반까지가 **높게** 나타나 대조적인 모습을 보인다.
237. 다행히 아들의 **대답은** ( ) 그런 부친의 불길한 예감을 한꺼번에 떨쳐 버릴 **수** ( ) **있게** 할 **만큼은** ( ) 안 **되어도** 어느 **정도** ( ) 덜어 **주기에는 넉넉했다.**
238. 박씨의 **작품은** ( ) 우아하면서도 점잖은 우리 전통 한복을 변형 없이 그대로 전수하고 있다는 평가를 **받는다.**
239. 곡식을 진 사람, 나무를 진 사람, 짚신 꾸러미를 멘 사람, 송아지를 모는 사람, 마을의 **장꾼들은** ( ) 죽당 선생의 앞뒤에 서서 열을 **지었다.**
240. 아직은 어린애들의 장난으로 때때로는 헛탕치는 **합승차도** ( ) **있으나** 신기한 느낌이 한물 가시면 어린애들의 **장난도** ( ) 자취를 **감추리라.**
241. 우리 전통화장에서 **이** ( ) **같은** 화장을 '화냥년 화장'이라 천시했던 화장이다.
242. 앞서 한 말보다는 나았지만, 배교수의 말을 온전히 알아듣기에는 경찰로서의 **10년** ( ) **가까운** 세월이 여전히 수월찮은 장애로 남아 있었다.
243. **연락선** ( ) **타는** 데서 **짐** ( ) **나르는** 일이 많대.
244. 사람이 죽고 다치는 이 편싸움이 이렇게 수천 **년간** ( ) 지켜져 **내려온 데는** ( ) **이** ( ) **같은** 내부의 불만을 외부로 분출시켜 촌락 공동체의 단합과 결속을 노리는 저의가 **숨겨져** 있었던 것이다.
245. **현우는** ( ) 간신히 읍내 복판으로 **잡아들었다.**
246. 우리나라에서는 역사적으로 전자인 정주계의 이학이 크게 발달하고 후자인 육왕계의 **심학은** ( ) 별로 **발달하지** 않았다.
247. 그제서야 **남경사는** ( ) 의심을 **풀었다.**
248. 여기서 우리 시대의 청년문화에 나타나는 전통에 대한 평가의식을 몇 가지 태도로 탐색해 볼 **수** ( ) **있다.**
249. 스스로를 위해서는 양말 한 켤레 속옷 한 **장** ( ) 여분으로 **지니는** 법이 없었고, 또 **방학은** ( ) 항상 고아원에서 무료봉사를 하거나 나환자촌(촌)에서 **지낼** 정도였읍니다.
250. 미군정이 시작될 당시 한국문화가 놓였던 처지는 일본문화에 의해 상당한 **깊이까**

지 ( ) 전통문화가 변질당하고 **파괴당한** 상태였다.

251. 주위의 **지면은** ( ) 하얗게 **덮이고 빙벽은** ( ) 유난히 푸른 색을 **발한다.**
252. 어쩌면 **동화** ( ) **읽기의 재미는** ( ) 이 기적을 발견하는 데에 **있을지도** 모릅니다.
253. 첫째 자녀의 성비차이가 둘째 셋째 넷째로 갈수록 더욱 증폭, 남아비율이 1백 14.3명 2백 5.9명 2백 37.7명이나 된다는 **것은** ( ) 성감별을 통한 여아낙태가 그만큼 성행함을 **입증하는** 것이다.
254. 이처럼 서양에 있어서의 **인간은** ( ) 신과 자연, 혹은 리성과 정의, 이 양쪽을 마치 두 개의 여관집 모양으로 한때는 이 집, 한때는 저 집으로 정처없이 왔다 갔다 하는 인간이 **되었다.**
255. **야나예프는** ( ) 무척 **당황했다.**
256. **중개료는** ( ) **받지** 않는다.
257. 홍씨의 믿음처럼 **해마다** ( ) 4~5월이면 새싹이 **돋은** 소나무에서 지난 **가을** ( ) 5 개의 싹이 **나온** 데 이어 4일 **현재** ( ) 또 2 개의 조그만 싹이 모습을 **드러냈다.**
258. 부모가 모두 일에 쫓긴 탓인지 겨우 걸음마를 할 **때부터** ( ) 우리 신전 앞뜰을 **아장거리던** 그 **아이는** ( ) 대여섯 살이 되어 부모와 함께 이 도시를 떠날 **때까지** ( ) 줄곧 우리 신전 주위를 맴돌며 **자라났다.**
259. 튤립 열풍의 **종말은** ( ) 1637년 2월 4일, 불시에 **찾아왔다.**
260. **미화원협의회는** ( ) 작년 12월 **17일** ( ) 구두닦는 일을 하는 회원들의 권익 보호 등을 위해 **발족됐으며** 이번이 두 번째의 정기 총회였다.
261. 아파트의 **경우** ( ) ▲ 분양평수를 실제보다 크게 하거나 ▲ 교통 거리표시기준이 모호하고 ▲ 근거 없이 분양가격이 싼 것처럼 광고하는 것이 문제점으로 **지적됐다고 소보원은** ( ) **밝혔다.**
262. 이번 재판으로 12·12와 5·18은 물론이고 군부 집권과정의 모든 불법행위가 한점의 의혹없이 모두 밝혀지고 **처벌까지** ( ) **마무리지어져야** 한다.
263. 아하스 페르츠가 그렇게 말하자 **그녀는** ( ) 포옹을 풀고 희고 부드러운 손을 들어 그의 입을 **막았다.**
264. 숨을 한 **번** ( ) **들이마셔도** 서울 공기와는 다른 **것** ( ) **같았다.**
265. **그는** ( ) 공산주의자들이 흔히 그렇듯이 나에게 '어떻게 사는가', **'아파트는** ( ) **어떤가'** 등의 사적인 **질문은** ( ) 전혀 **하지** 않았다.
266. 낡았으나 단정한 검은색 정장(정장)이며, **목소리는** ( ) 되도록이면 **부드럽게** 그리고 **몸가짐은** ( ) 과장의 혐의가 들만큼 겸손하게 **가지는** 것 등에서 풍기는 독특한 분위기 때문이었다.
267. 우리의 전통문화는 우리의 생활과 정서 속에 살아 움직이고 있는 **동안** ( ) 우리 시대에 그 전통문화가 변혁되고 재창조될 **수** ( ) **있는 가능성은** ( ) **열려** 있다.
268. 사실 그렇게 되고보니 **교회는** ( ) 엉망이 **되고** 말았읍니다.
269. 현관문을 열고 들어가는 **순간** ( ) **그들은** ( ) 기겁을 하고 제자리에 **얼어붙었다.**
270. 한 **걸음** ( ) **물러나서** 그 일에 대하여 깊이 생각해 볼 **때야말로** ( ) 진정으로 비평이

시작되는 것이다.

271. **사람들은** ( ) 혹은 산에 올라가서, 혹은 동청에 모여서 천지가 떠나가라고 통곡을 **했다.**
272. 셋 중에서 가장 나이가 지긋하고 세상일에 경험이 많은 **발타자도** ( ) 식은땀을 흘리면서 **당시** ( ) 그런 일에 효험이 있다고 **믿기우던** 주문(주문)을 아는대로 **외어댔다.**
273. 이해관계, 유용성, **갈등** ( ) **같은** 개별적 **변인은** ( ) 무지와 **우연성** ( ) **같은** 가정을 **무력화 시킬** ( ) **수** 있다.
274. 쇼스포츠를 설명하기 위하여 다섯 가지 방법론이 **현재** ( ) **이용되고** 있다.
275. 그래서 이 책의 일차적 **목적은** ( ) 포퍼 철학의 중심 사상을 체계적으로 **정리해** 보고자 하는 **데** ( ) **있다.**
276. 그리고 조영감이 돈으로 우겨 간신히 허락받은 사립학교의 **전학마저** ( ) **거부하며** 전에는 민요섭과 숨어서 하던 일을 공공연히 드러내놓고 하기 시작했다.
277. 클린턴이 그같이 애매한 상태에 있을 **때** ( ) 지금 그와 후보 경쟁을 하고 있는 봅 캐리 **상원의원은** ( ) 월남의 정글에서 베트콩과 전쟁을 **하고** 있었다.
278. **사람** ( ) **쓰는 데** ( ) **있어서의 적성도** ( ) 외형적인 것만으로 적소를 **가릭** 게 아니라 심리적인 적성을 중요시하는 것은 쓰는 사람이나 쓰이는 사람을 위해서도 좋고, **효율도** ( ) **배가될** 것이 정한 이치다.
279. 이 점을 고려하면 사칠론과 서로 다 같은 심성론이라 하더라도, 이것은 사칠론보다 형이상학(이기론)의 폭이 한층 더 확대된 경우라 할 **수** ( ) **있다.**
280. **당시** ( ) 인도네시아 전역에는 약 1천 4백 명의 한국인 군속이 **있었으나,** 비밀 유지를 위해 한 **사람씩** ( ) **접촉해** 모인 정예 당원들이다.
281. 음악단에 돌아가서 예정된 위안 연주에 참가하는 것, 이것이 돌아가신 어머니의 영혼을 위로해 드리는 것이 되고, 또 걱정을 거두지 못하고 눈을 감으신 아버지에 대한 마음의 **갚음도** ( ) **되는** 것이라 생각한 때문이었다.
282. **그것은** ( ) 유입 **초기부터** ( ) 퇴계를 비롯한 정주계 학자들로부터 **이단시되어** 심한 배척을 받아 발전할 여건을 **맞지** 못하였다.
283. 그렇지만 이 녀석들아, 일본말 **한마디도** ( ) **모르는** 내가 무슨 신식 학교 **훈장** ( ) **될** 자격이 있단 말이니.
284. 그토록 여론의 지탄을 받으면서도 각 정당이 시정하지 못한 저질성명에 대해 마침내 **선관위까지** ( ) 나서게 **됐다.**
285. 거기에 포상이 걸리지 않은 다수의 여타 선수들의 사기에 끼치는 **영향도** ( ) 생각해 **봄직하다.**
286. 그러나 지난해 우리 **나라는** ( ) 88억2천만달러의 경상수지적자를 **기록했으나 대만은** ( ) **9월까지** ( ) **13억8천만달러나** ( ) 흑자가 **났다.**
287. 지난해 **10월** ( ) 대통령 관저가 **신축된** 데 이어, 대통령 집무실인 청와대 본관 건물이 착공 2년 2 개월만인 **4일** ( ) **준공됨으로써,** '새 청와대' 시대가 막을 올리게 됐다.
288. 아하스 **페르츠는** ( ) 그때껏 쓰던 빈정거림이나 비꼼의 말투에다 갑작스런 악의와 공

격성을 더하며 부친의 말을 **받았다.**

289. **페르손은** ( ) 심판석에 다가가 카운트가 잘못 됐다면서 자신의 '17'스코어를 '16'으로 **낮추었던** 것이다.
290. 그 **당시** ( ) 수많은 지식인들과 공무원들과 국영기업체 직원들과 초중고등학교 교사들이 신정을 설날로 받아들이고 **지키지** 않았습니까?
291. **지구는** ( ) 그 나름대로의 역사를 **가지고** 있다는 사실이 발견되었다.
292. 그리고 **다케시타는** ( ) 오부치 게이조(소연혜삼·54) 전 간사장과 하시모토 류타로(교본용태랑·54) 전 대장상을 심복으로 **두고** 있다.
293. 스페인 소년들의 최대의 꿈은 일류 투우사가 되는 일이요, 그들이 가난을 벗어날 **수** ( ) **있는** 유일한 활로가 소와 싸우는 일이다.
294. 아시아의 여러 **나라는** ( ) **고사하고** 우리 나라의 **전통스포츠만** ( ) **해도** 그것을 국제규격으로 가꾸면 25종이 넘고도 남는다.
295. **아이들은** ( ) 쏜살같이 뒷문으로 튀어나와 뒤편 밤나무 숲 속에 엎드려서 **눈만** ( ) **내놓고** 있었다.
296. 정필준 북경대 중외부녀 **문제연구중심상무주임은** ( ) 때문에 여성의 문맹률이나 학교에서의 **중도탈락률은** ( ) 남성보다 훨씬 **높다고 말한다.**
297. 중국이 자국영토인 대만의 독립을 저지하고자 무력을 사용하는 데 대해 제3자가 왈가왈부하는 것은 내정간섭이라는 **주장도** ( ) **일리는** ( ) **있다.**
298. 이 숨구멍을 못 뚫으면 **해표는** ( ) **질식사할 수** ( ) **있다.**
299. 새들이 많이 앉아 잇는 지역은 길이 150 내지 200미터에 폭이 40, 50미터이니 크릴의 무리로서는 큰 **무리는** ( ) **아니다.**
300. 아하스 **페르츠는** ( ) 문득 아이다운 **자랑까지** ( ) **느끼며** 배운 것을 **늘어** 놓았다.
301. 응원하는 관중에게도, 또 개최국의 국민에게도 주어지고, 그 경기에 큰 도움을 준 **기상(기상)도** ( ) 대상이 **된다.**
302. 국민의 입장에서는 지하철공사도 정부고, 철도청도 정부인데 이렇게 단단한 지역이기주의의 껍질 속에 들어앉아 국민의 **불편은** ( ) 나몰라라 **해도** 좋은 것인가.
303. 아니 그 **목사는** ( ) 오히려 **예배시간까지** ( ) **줄여가며** 작업을 **독촉하기도** 했어요.
304. 게다가 교통수요에 맞추느라 무리한 **운행까지** ( ) **하고** 있다.
305. 포퍼의 기준에 따른다면 형이상학적 **언명은** ( ) 비과학적 언명이라 하더라도 무의미한 **언명은** ( ) **아니다.**
306. 스즈끼 **박사는** ( ) **동양인들은** ( ) 자아를 침잠시키려는 경향이 **있고, 서양인들은** ( ) 자아를 강조하려는 경향이 **있다고 지적한다.**
307. 물론 전통사회에서는 청년층이 분리되지 않고 청년문화가 성립하지 않는다고 보는 **립장도** ( ) **있으나,** 우리의 전통사회에서도 청년층이 독자적인 사회이념과 가치관을 갖고 행동으로 실천한 경우를 쉽게 찾아볼 **수** ( ) **있다.**
308. 이것을 **도스토예프스키는** ( ) 시베리아 유형지에서 뼈에 사무치도록 **깨달았었다.**
309. 오후 2~4시 12.7%, 낮 12시~오후 **2시도** ( ) 10.0%를 **차지해** 안전사고가 오후시간대

에 집중적으로 발생하는 것으로 나타났다.

310. 이것으로 보면 **맹자는** ( ) 육체적인 면에서 오는 요구보다 정신면에서 오는 요구, 다시 말하면 유한한 국한된 요구보다 영원하고 보편적인 요구를 사람의 참된 성이라고 **본** 것임을 알 **수** ( ) **있다.**
311. 기원전 4 세기경의 고대 희랍올림픽에서는 어느 폴리스(도시국가)에서보다 많은 우승자를 내느냐가 나라체면 **뿐** ( ) **아니라** 국력을 가늠하는 기준이 되었기로 우승자에게 거액의 상금을 거는 풍조가 공공연하게 자행되었었다.
312. **이제까지** ( ) 이에 대해서 비판하는 의견이 **제출된** **까닭도** ( ) 여기에 **있는** 셈이다.
313. 외아들의 청인데다 민요섭이 아들을 가르칠만한 능력이 있다는 걸 믿고는 있었으나, **조영감은** ( ) 왠지 선뜻 마음이 **내키지** 않았다.
314. 거기에서는 말이 **필요** ( ) **없으며** 이성의 **추론** ( ) 역시 **쓸모** ( ) **없게** 된다.
315. 어떻든 이번 **사건은** ( ) 정부에 대해 재외공관직원과 상사주재원, 유학생, 교민들의 안전문제에 관한 경각심을 **일깨워주었다.**
316. 지난번 종합대책 **때** ( ) 내놓을 **것은** ( ) 다 **내놨으니까** 남은 **일은** ( ) 수출업체 **독려밖에** ( ) **없다고** 생각하는 모양이지만 우리가 보기에는 사태의 심각성에 대한 문제의식이 너무 안이한 것 같고 대응자세도 너무 소극적이다.
317. 노개위에서는 복수노조의 **경우** ( ) 노조전임자에 대한 임금을 노조에서 지불하고 또한 단일대표권이 확립돼야 한다는 전제 **아래** ( ) 사용자측의 동의를 **얻어내기까지** 했다.
318. 숨겨진 부분의 **이미지는** ( ) 내밀한 꿈 안에서 혼합되는 섹스, 나뭇잎, 거울, 책, 무덤의 이미지에 **관계된다.**
319. **친구들도** ( ) 스스로의 마음이 아닌 외적 조건에서 사이가 **멀어진다.**
320. 한양유통 구매부의 한 **관계자는** ( ) 현재의 관행으로는 소비자 **뿐** ( ) **아니라** 대형**유통업체도** ( ) '빈병'과 관련, 손해를 **보고** 있다고 **털어놓았다.**
321. 하지만 이런 **부분들은** ( ) **이후** ( ) 장편소설론의 장르적 성격이 **파악되면서,** 모랄, 뭉속론을 통해 세계관과 전형적 상황의 문제가 검토되는 하나의 계기를 **이룬다.**
322. 작년 **이후** ( ) 남북한이 고위급회담을 갖는 과정에서 평양을 방문했던 한국 대표단이 김정일을 한 **번도** ( ) **만나보지** 못한 것은 그가 의식적으로 외부 손님을 피하기 때문이다.
323. 이 물음에 대한 고자의 **답변은** ( ) **없으므로** 고자가 어떻게 말한 것인지는 모르나 **맹자는** ( ) 다음과 같은 논리로써 그 구별을 **세웠다.**
324. 장애인 **이용신청** ( ) **밀려**
325. 원래가 미장부인 그에게 지적인 **매력까지** ( ) **더해** 주고 짙고 차분한 음영이었다.
326. 누가 마라톤 종목에 출전했는지도 모르고 있었고, 어느 **누구도** ( ) 우승하리라고 **기대하지도** 않았을 뿐더러 이전에 들어본 **적도** ( ) **없는** 무명 선수였기 때문이다.
327. 내 나라를 찾는단 바람에 **사람들은** ( ) **줄** ( ) **당길** 적 기분으로 너도나도 차림을 하고 **나왔다.**

328. **이는** ( ) 인쇄술이 발달하면서 출판업계에도 특별히 호평을 **받았다.**
329. 이 두 **논변이야말로** ( ) 한국 성리학이 지닌 주지주의의 특징을 입증하는 실례임에 **틀림없다.**
330. 올림픽 **성화는** ( ) 1936년 베를린 대회 **때** ( ) **시작된** 것으로 그 역사가 **깊지** 않다.
331. 대만에 대한 중국의 군사적 **시위는** ( ) 동북아 전체의 안보라는 틀에서 큰 파장을 **몰고** 올 전망이다.
332. 선관위는 여러 차례의 **경고도** ( ) **먹히지** 않자 고발 등 강경조치를 취할 방침이라는데 선관위로서는 당연히 그런 무서운 면을 보여야 한다.
333. 그러나 깨달음의 **정점은** ( ) 끈질긴 노력과 고통을 수반하지 않고서는 **도달되지** 않는다.
334. **그는** ( ) 명실공히 일본의 최고 영웅으로서 우뚝 **섰던** 것이다.
335. 그 학생이 그 여자와 무슨 일이 있었다 해도 비난받을 **쪽은** ( ) 그 **학생은** ( ) **아니라고** 생각됩니다.
336. 제1차 한국관광진흥회 및 OTF 관광교역전으로 열리는 이번 **전시회는** ( ) 43 개 나라, 6백여 업체가 참가, **12일부터** ( ) 일반에 **공개된다.**
337. **그대는** ( ) 아주 오래전에 부모와 함께 이 도시에 온 적이 **없는가?**
338. 영, 불간의 지루한 백년전쟁 **후** ( ) 영국의 헨리 8세와 프랑스의 프랑스와 **1세는** ( ) 도버 해협의 카레 시에서 전후의 복잡한 난제를 두고 회담을 **하고** 있었다.
339. 남극 개미자리의 가장 큰 **군락지는** ( ) 남쉐틀란드군도의 디셉션섬에 **있었으나** 1967**년** ( ) 화산 **폭발시** ( ) 화산재로 **덮여서** 멸실된 것으로 알려졌다.
340. 인민회의가 채택한 개헌안 중 외국자본 유치를 위한 사유재산제 일부 **허용은** ( ) 구소련의 붕괴에 따른 경제 원조 중단으로 인한 심각한 경제 위기를 맞고 있는 쿠바가 경제 부흥을 위해 국가 독점 경제체제에 급진적인 변화를 꾀한다는 점에서 특히 **주목된다.**
341. 온 덕수궁 안이 발끈해졌을 것은 말할 **것도** ( ) **없는** 일이었다.
342. 그럼에도 **매년** ( ) 8백여만명의 어린이들이 홍역, 백일해, 파상풍, 소아마비, 설사병 등 5가지의 질병으로 목숨을 **잃고** 있다고 **보고서는** ( ) **밝혔다.**
343. 문장로의 집을 나올 무렵도 마찬가지였으며 – 풍기는 분위기로는 **지금도** ( ) 어지러운 남성편력에 빠져 있는 것 **같았다.**
344. 그러나 이들 캐릭터 상품 대부분이 외국산이란 **점은** ( ) 문제로 **지적된다.**
345. **그것은** ( ) 직관적으로 알 **수밖에** ( ) **없는** 신의 의지에 자신을 복종시키는 일종의 달콤한 폭정과 **같다.**
346. 항상 손님으로 장사진을 이루고 있는 이 **음식점은** ( ) 고기, **생선뿐** ( ) **아니라** 국수류, 샐러드, 과일에 이르는 모든 음식에 마늘을 넣어 맛을 **내고** 있다.
347. **선진국** ( ) **같으면** 시위가 아닌 도시게릴라로 규정, 진압을 위해 **군마저** ( ) **동원했을** 게 틀림없다.
348. 다음 전설을 알지 않고는 이 원매의 **중추시는** ( ) 진수를 **맛볼** 수가 없다.

349. 그러나 독서 생활을 오락만으로 채우면 **결국은** ( ) 상상력이 황폐하되고 중독 현상이 **일어나기** 쉬우므로 주의할 일이다.
350. 그러나 **그는** ( ) 곧이곧대로 그 말을 **받았다.**
351. 아버지가 이런 봉변을 당한 삼사 일 뒤, **현우는** ( ) 서당에서 동무들과 창가를 하다가 훈장님한테 처음으로 매를 맞아 본 일이 **있었다.**
352. 한편 한반도에서도 **개척사업은** ( ) **있었다.**
353. **5년전부터** ( ) **앓아오던** 갑상선 기능 항진이라는 병의 원인이 신장의 이상 때문이라는 **사실도** ( ) 이 **때** ( ) **알았다.**
354. 이런 **사실은** ( ) 만해의 정신을 지배한 또 하나의 사상으로서 민족주의가 있음을 감안해 보면 그 성격이 더욱 **선명해진다.**
355. 당시 23세였던 **클린턴은** ( ) 징집 대신에 아칸소대학 ROTC를 하겠다고 **했는데** 등록한 **사실은** ( ) **없으며** 몇달 후인 그 해 **10월** ( ) 친구들이 전사하는 소식 등을 듣고 입대를 **결심했으나** 징병 제도가 바뀌어 입대하지 않아도 됐다는 것이다.
356. **이것은** ( ) 그가 한글을 표현매체로 한 현대시와 산문사용의 훈련과정을 전혀 거치지 않았음을 **뜻한다.**
357. 본체와 화살 10 개를 합해 30만원이란 고가에도 불구하고 석궁이 인기를 끄는 것은 양궁과 사격의 재미를 동시에 맛볼 **수** ( ) **있을 뿐** ( ) **아니라** 다루기가 쉬우면서도 정확도가 총에 버금가는 매력때문이다.
358. 옛이야기에 관심이 많다는 **그는** ( ) 5살짜리 아들에게도 전래동화를 많이 들려준다고 **말한다.**
359. 둘 사이의 **거리는** ( ) 백 **큐핏도** ( ) **넘었지만** 아하스 페르츠에게는 그녀가 바로 앞에 서 있는 것처럼 느껴졌다.
360. 하지만 그같은 **짓은** ( ) 귀엽게만 **보고** 있을 **수** ( ) **없는** 것이 그때의 내 처지였다.
361. 이것은 인간의 자각의 내적 요소를 이해하는 것이니 인간의 자기 인식 · 자기 반성에 의하여 그의 자각 **상태는** ( ) 단계가 **있어서,** 그것을 객관적으로 표시할 때에 세간 · 출세간 · 출출세간이라 하고, 인격적으로 표시할 때에 범부 · 라한 · 보살이라고 한다.
362. **그는** ( ) 이집트나 인도산 면사로 짠 팬티, 와이셔츠를 **입는다.**
363. 어느 한 올림픽 구기종목에서 우리 한국팀이 4 강에 오르면 3천만 원씩, 결승에 오르면 4천만 원씩, 우승을 하면 5천만 **원씩** ( ) 포상을 **한다는** 보도에 접하고 보니 이 나귀의 당근이야기가 생각나는 것이다.
364. 특히 비만자가 다치지 않고 맘껏 몸을 놀려 살을 빼는 **데** ( ) **적당한** 운동으로 식이요법과 병행하면 좋은 효과를 볼 **수** ( ) **있다고 진박사는** ( ) **말한다.**
365. 사귄 지가 **얼마** ( ) 안 **돼나서……**
366. **현우는** ( ) 별안간 무서워지려 **했다.**
367. 그래서 뚱이, 하면 공공 공간에서 지킬 **것** ( ) 못 **지키는** 한국사람이란 뜻이 돼버린 것이다.
368. 한국문화의 주체가 희미해지는 때에 그러한 순수 한국적인 것을 찾는 것은 **의의** ( ) **있**

는 방향이라고 할 수가 있겠지만 앞으로의 한국문화 전통의 전개에 하등의 시대적 **의의도** ( ) **없는,** 하나의 세계 문화에 아무 새로움도 없는 것을 굉장한 것으로 착각하고 그것이 새로운 인간이나 생활의 원형으로 제시되는 류의 맹목적 복고의 전통긍정 **태도도** ( ) 비판과 경고를 **받아야** 한다.

369. **당시** ( ) 오페르트라는 고약하고 고집센 독일 사람이 무장선을 이끌고 와서 역시 고집센 대원군에게 통상을 **강요했을** 때 일이다.
370. **내** ( ) **쓰던 바이올린은** ( ) 널 **주마.**
371. 현우와 **어머니는** ( ) 진종일 소리 없이 **드러누워** 있었다.
372. **예금주** ( ) **모르게** 두 구좌에서 도합 5백만 원을 찾아간 것이다.
373. 사람과 사람 사이를 가장 빠르고 쉽게 가까워질 **수** ( ) **있도록** 하는 것으로 피보다 더한 **것은** ( ) **없기** 때문이다.
374. **태국선수단은** ( ) 미스 유니버스를 **앞세웠길래** 손이 절로 흔들어지고.
375. 그들이 재조직되어서 하나의 통일체가 되었을 **때** ( ) 처음으로 완전한 의미를 표현하고 이해하게 **된다.**
376. 이런 관점에서 보면 지금과 같은 **독직사건은** ( ) 언제나 일어날 **수** ( ) **있는** 개연성을 **갖고** 있다.
377. 권련토막이 손가락 집게 **사이만** ( ) **남고** 완전히 재로 변한 뒤에야 **늙은이는** ( ) 가슴 깊은 곳에서 우러나는 한숨과 함께 얘기를 **시작했다.**
378. 동시에 몇 분 뒤에 현지에 도착할 **수** ( ) **있는가를** 응답해 주는 것이다.
379. 즉 노동은 단순한 육체적인 **활동** ( ) **뿐만** ( ) **아니라** 두뇌의 정신작용에 의한 정신노동을 포함하는 복합적인 행위이다.
380. 재앙을 몰아온다는 혜성 **공포는** ( ) 다소 화학화된 **채** ( ) 현대인의 **마음까지도** ( ) **사로잡고** 있다고 보도되고 있다.
381. 또 헌금 **강요는** ( ) **어땠는지** 아십니까?
382. **자기는** ( ) 그럴 만한 자격이 **없다고** 잘라말했을 **뿐만** ( ) **아니라** 그 아이가 자기와 가까와지는 **것은** ( ) 그 아이를 위해서도 결코 **이롭지** 못하리라는 걸 넌지시 일러 주기까지 했다.
383. 한데도 도시국가간의 경기인 올림픽이 전쟁 때문에 중단된 적이 한 **번도** ( ) **없다는** 사실이 어떻게 설명될 **수** ( ) **있는** 것일까.
384. 이제 **북한은** ( ) 대남 적화망상에서 깨어나 체제유지와 함께 남북한이 공존하는 길이 무엇인가를 **깨달아야** 한다.
385. 그러한 **사람들은** ( ) 갓 태어난 자기 자식을 교살하는 어머니와 **같다는** 것이다.
386. 무슨 과인지는 모르지만 아직 항만청에 다니는 **건** ( ) **확실하오.**
387. 비단옷을 입을 **때는** ( ) 그 위에 박사(박사)를 **걸쳐** 밖에서 보이지 않게 한다는 **교훈도** ( ) '시경(시경)'에 **있다.**
388. 고티를 유죄로 묶는 **데** ( ) 결정적인 역할을 **한** 것은 그의 밑에서 감비노가의 부두목 노릇을 해 온 살바토레 그라바노(42)였다.

389. 이 **때는** ( ) 경비실에서 경고음 작동을 **일시** ( ) **중지시킨 뒤** ( ) 수신반에 불이 들어오는지를 확인하는 **방법도** ( ) **있다는** 게 한국소방안전협회 김종관 연구위원의 설명이다.
390. 중국과 대만해협의 긴장에 대한 문제, 한·미간의 대북정책 공조문제 **등도** ( ) **포함돼** 있다.
391. 남을 위해 옳은 **일** ( ) **한** 사람을 사회가 이처럼 푸대접하는 것이 성폭행현장을 목격하더라도 간섭하지 말도록 가르치는 것과 무엇이 다르겠는가.
392. 그러나 **아들은** ( ) **달랐다.**
393. 그렇다고 주머니에 돈이 있는 **것도** ( ) **아니었습니다.**
394. 아리스토파네스의 희비극을 비롯, 각종 문헌 속에 기록된 고대올림픽 **개막식은** ( ) 대충 **이러하다.**
395. 본질적으로 텍스트중심적 제도인 **문학은** ( ) 인쇄술의 발명과 상당히 인접한 관련을 **맺고** 있다.
396. 아버지를 간호하느라고 **현우는** ( ) 며칠을 두고서 서당에도 **나가지** 못했다.
397. 또한 전통주의적인 이념과 질서가 아무리 확고한 권위를 누리고 있어도 **청년층은** ( ) 언제나 모험적 탐구의욕과 개방적 수용자세를 **지녔다.**
398. 그러나 **우리는** ( ) 문학이 죽음을 맞이한 원인을 크게 외적 요인과 내적 요인의 두 가지로 나누어 생각해 볼 **수** ( ) **있다.**
399. **경총은** ( ) 적자나 1인당 매출액 감소 기업의 **경우** ( ) 임원은 물론 일반사원까지 임금동결, 총액임금 및 개인임금의 동결, 고정급화돼 있는 상여금의 실적급화, 전 회원사 임원의 내년 임금동결 등을 **결의했다는** 것이다.
400. 그래서 인간이 원숭이의 후손이라는 **주장도** ( ) 이때부터 인정을 **받게** 되었다.
401. 세계적 스포츠 스타들의 종합 심리테스트 결과를 보면 **장거리선수는** ( ) 대체로 소극적이고 자기 자신을 엄하게 다스리는 자학성(자학성)기질이 있어야 기록이 **오르는** 것으로 나타났다.
402. 그리고 그 작품의 격조도 어느 **정도는** ( ) **유지되어** 있다.
403. 그리하여 테도스 **따위는** ( ) 까맣게 **잊고** 숙부의 집으로, 그리고 거기서 기다리는 열두 살 소년의 세계로 냅다 뛰기 시작했다.
404. 폐지해 버린 **이후부터** ( ) 각 종족 내부에 이전에 없던 일들이 잇달아 **일어났던** 것이다.
405. 현우와 **어머니는** ( ) 번갈아 미음을 **떠넣었다.**
406. **그뿐** ( ) **아니라** 성리학이 특히 17세기 이후에는 예학에 의거하여 거의 교조적으로 계승되었음을 감안하면서 이 인물성 탐구와의 관계를 살펴야겠다.
407. **풀잎은** ( ) 먼지가 보얗게 **나풀거린다.**
408. 요즘 1만원권 위폐가 계속 발견되고, 현금지급기 파손·도난사고가 자주 일어나는 것을 보면 신용사회의 기초가 중대한 도전을 받고 있다는 우려를 금할 **수** ( ) **없다.**
409. 1945년 **10월** ( ) **유림들은** ( ) 성균관에서 전국유림회의를 개최하여 앞으로의 재건

방향을 **논의하였다.**

410. 더욱이 안보위주의 주변 4강외교에 치중해 오던 우리로서는 외교의 지평을 넓혀 세계화의 길을 더욱 다듬는다는 **의미도** ( ) **있다.**
411. 에스키모인들의 신방이 설동이라는 **것도** ( ) **그렇다.**
412. **젊때** ( ) **줄** ( ) **당기던** 날 밤에 그 **집** ( ) **가서** **옳식** ( ) 실컷 **얻어먹고** 왔어요.
413. **전설은** ( ) 그 무렵의 아하스 페르츠를 이렇게 **전한다.**
414. 그래야 기업의 자금에 대한 **과잉수요도** ( ) **줄이고,** **이는** ( ) 다시 경제안정과 맞물려 금리를 **내리게** 하는 선순환의 고리를 만들 **수** ( ) **있다.**
415. 요금을 1년에 세차례나 찔끔찔끔 올리는 **식도** ( ) **찬성하기** 어렵다.
416. **이스라엘은** ( ) 지난달 24일 4차 쌍무협상 벽두에 팔레스타인 문제에 관한 제안을 **내놓았다.**
417. 가격이 상승함에 따라 튤립 재배와 무관한 **사람들까지** ( ) 투기에 **참가하여** 많은 사람들이 갑자기 부자가 되었다.
418. 이처럼 한국의 **종교들은** ( ) 전통적으로 관용과 조화의 정신을 **지니고** 있다.
419. 환자에게 정신적 신체적 부담이 거의 없다고 **공전문의는** ( ) **말한다.**
420. 포퍼 철학의 특이한 점은 그것이 철학자들보다 오히려 분과 학문의 학자들에게 **오늘날** ( ) 여타의 철학에서는 찾아보기 어려운 강한 영향력을 **행사하고** 있다는 점이다.
421. 신이 그 인간의 고소극한을 시험하는 **데** ( ) 영광스럽게도 당신이 **선택된** 것이냐고 물었더니, "산소통에다 인간의 인내력을 포기한 많은 다른 등산가들을 신이 선택하지 않았을 뿐"이라고 하던 말이 기억에 남는다.
422. 산골 **마을은** ( ) **가을빛만** ( ) **짙어가고** 있다.
423. 서양의 그것과 흡사한 **전통구기(전통구기)도** ( ) **많았다.**
424. 지구와 달과의 등거리에 있는 안정된 중력권에 세워질 우주식민시에는 푸른 나무에 **새까지** ( ) **우는** 공원이며 신에게 기도할 교회며, 그리고 **묘지지역도** ( ) **구획해** 놓고 있다.
425. 도수 **물안경은** ( ) 시력 0.03에서 0.3까지 시력에 따라 23종이 **있다.**
426. 그리고 기득이 **삼촌도** ( ) **안녕하시구.**
427. '물미장(물미장)놀이'라 하여 육상 **5종경기도** ( ) **있었다.**
428. 이렇게 본다면 김남천이 가지고 있는 소설론의 장점과 **한계는** ( ) **뚜렷해진다.**
429. **동화는** ( ) 어정쩡한 **절충** ( ) **같은** 것을 가장 **싫어합니다.**
430. 또 긴 원피스 속에 짧은 바지를 입고 원피스 아랫 단추를 풀어 **조끼** ( ) **같은** 기분으로 겹겹 패션을 연출하는 것도 올해 여름의 경향이다.
431. **이들은** ( ) 비록 전통적 가치와 생활관습을 소극적으로 허용하고 있지만, 적극적으로 서구적 가치관과 생활양식을 수용하여 자본주의화와 산업화를 실현하였으며, 6·25를 경험하여 반공의식이 **확립되어** 있다.
432. 그리고 대학들과 협조관계에 놓여 있던 라브라리수타치오나리 그리고 **자영서적상들도** ( ) **있었다.**

433. **플라톤부터** ( ) **시작하여** 지금까지 ( ) 철학은 ( ) 줄곧 언어가 본질적으로 수사적이라는 사실을 인정하지 않으려고 했다고 **니체는** ( ) **지적하였다.**
434. **1년전** ( ) 보리스 옐친이 러시아 사상 최초의 민선 대통령에 **당선된** 날인 동시에 러시아가 그 주권 독립을 선포한 이 날을 러시아 최고회의가 공휴 국경일로 정했기 때문이다.
435. 또 현지에서 활동하는 **동안** ( ) 여러 가지로 **도와준** 우넨 신문사 여러분의 **후의도** ( ) 잊을 **수** ( ) **없다.**
436. 전쟁속에서도 **학교는** ( ) 문을 **열었다.**
437. 그뒤 다시는 교회에 나가지 않았지만, 소년기가 거의 끝날 **때까지** ( ) 가끔씩 막연한 동경으로 **올려보곤** 했던 교회당 첨탑 위의 흰 십자가도.
438. 심지어 지식인들을 비롯한 사회 지도층 **사람들은** ( ) 독서행위가 이렇게 널리 퍼져 있는 사태에 적절한 조치를 취할 것을 **주장하기도** 하였다.
439. 이석채 **정보통신부장관도** ( ) 취임 **당시** ( ) 경쟁에서 **2등은** ( ) **안되고** 능력있는 1등 사업자를 선정하는 것이 원칙이라고 밝힌 **바** ( ) **있다.**
440. IOC의 다른 한 특징으로 여성위원(녀성위원)을 **하나도** ( ) **두지** 않는다는 보수성이 비난의 표적이 되기도 했다.
441. 그래서 **소년은** ( ) 잠시 궁리하다가 선장실로 **찾아갔습니다.**
442. **춤도** ( ) **서양춤은** ( ) 손발의 직각미(직각미)를 **추구하는데** 한국춤은 ( ) 흐느적거리는 곡선미를 **추구한다.**
443. 그 **분도** ( ) 지금 연구실에 **나와** 있을 겁니다.
444. 그러나 **척사위정론은** ( ) 민족의 위기에 일정한 공헌을 하였지만, 근대화의 높은 파도에 밀려 떠 **내려가고** 말았다.
445. 겨울철 겨냥 **신제품도** ( ) **잇달아**
446. 은행권 발행에 따라 경제 **활동도** ( ) **활성화되었다.**
447. 아직도 한두 **포기** ( ) 더 **있을지는** 몰라도 아까운 식물이 없어졌다.
448. 어느 **날** ( ) 공동묘지로부터 돌아오는 길에 현우 **어머니는** ( ) 밤마을 종수 어머니를 **만났다.**
449. 장기적으로 고용문제를 염려하지 않을 **수** ( ) **없다.**
450. 이렇듯 대통령의 **구상은** ( ) 여러가지 발전적 환경정책을 **담고** 있어 어떤 **것은** ( ) 우리 실정보다 훨씬 **앞서가고** 있다.
451. 용의 눈이 마치 사람의 **눈** ( ) **같고,** 호랑이의 네 발이란 것이 흡사 도마 다리같이 꼿꼿해서 우습기가 짝이 없었지만, 그래도 울긋불긋 물감칠을 그럴듯하게 해놓으니까 보기에 근사했다.
452. 최근에 구입한 것 같은 새책이었으나, 어떤 **부분은** ( ) 이미 까맣게 손때가 **묻어** 있었다.
453. 반면에 다른 **비평가들은** ( ) 사회주의 리얼리즘의 수용에도 불구하고 사회학주의적 미학관의 온전한 극복에는 여전히 실패하는 모습을 **보여준다.**

454. 이론적으로 체감온도 영하 30 C 내지 영하 60 C에서는 공기중에 노출된 **피부는** ( ) 30초이내에 언다니 **곧** 얼기는 얼리라.
455. 이러한 의미에서 대미관계의 **설정도** ( ) 대북 관계의 맥락에서만이 아니라 포괄적인 맥락에서 **이루어져야** 할 것이다.
456. **그들은** ( ) 한결같이 인간의 고통과 결핍, 공포와 원망(원망)이 빚어낸 우상들로밖에는 **비쳐지지** 않았으며, 그들을 향한 찬가 또는 **기구는** ( ) 그런 것들에 시달리는 인간의 절규로만 **들렸다.**
457. 서비스란 말 자체가 상대방이 잘 받아칠 **수** ( ) **있게끔** 봉사한다는 뜻이고 보면 서비스 에이스는 모순이다.
458. 특히 무고한 양민을 향한 도청 앞 발포책임자에 대한 처벌이 미흡하고, 불기소처분대상자가 지나치게 많은데다 **선정마저** ( ) 검찰이 자의적으로 **했다는** 것이다.
459. **조선시대까지는** ( ) 전통이랄 것이 **있었지만** 우리의 **신문화운동은** ( ) 그 전통을 부정하는 데서 **출발하였고** 따라서 현대의 **우리는** ( ) 단절된 전통, 곧 전통이 없는 곳에 **처해** 있다는 견해가 그것이다.
460. 섹시하고 쇼킹한 감각으로 선두를 다투는 두 디자이너가 동시에 내놓은 이들 **향수는** ( ) 속의 향기보다는 손에 잡고 쓰는 향수병의 쇼킹경쟁을 한 눈에 **드러낸다.**
461. 오이는 보습효과, 미나리는 피부탄력, 파슬리는 혈관확장, 당근은 신진대사 촉진, **토마토는** ( ) 피부를 매끄럽게 하는 작용을 **한다는** 것이 메이커측의 설명이다.
462. **남경사는** ( ) 피살자의 사진을 **꺼냈다.**
463. 이런 광경이나 **기쁨은** ( ) 문명세계에서는 **느끼지못한다.**
464. 타이핑하는 **손가락놀림까지** ( ) **계산된다** 하니 역사가 생긴 **이래** ( ) 가장 혹독한 노예상태가 **아닐 수** ( ) **없다.**
465. 또 5월에 미국을 방문했을 **때도** ( ) **그는** ( ) 같은 운을 **뗐다.**
466. 존경의 염이라고는 **눈** ( ) **씻고도** 볼 수 ( ) **없어 인사** ( ) **받는** 쪽이 오히려 불쾌할 정도다.
467. 어머니 품에 안겨보지 못하고 자란 아이가 **지능도** ( ) **뒤지고** 횡포하고 악의 구렁텅이로 빠질 요인이 많아지듯이 영상생활이 일상화되면 인간퇴보와 사회악이 급증하리라고 예언한 것은 미래학자 토플러다.
468. 전기를 통해 문학작품에 접근하는 풍요하고 **의미** ( ) **있는** 것이 된다.
469. 하지만 내가 어떻게 수소문해서 찾아갔을 **때는** ( ) 이미 **거기** ( ) **없었오**
470. 따라서 엥겔계수에 있어서 일본과의 **수치는** ( ) 점차 완만한 추세로 **감소되고** 있음을 알 **수** ( ) **있다.**
471. 돈이 없어 **유학** ( ) **못가는** 젊은 예술가들을 기업이 맡아 육성하겠다는 제도적 장치다.
472. 이와 같은 한국 다도의 정신이 어디서 왔느냐 하는 **것은** ( ) 한국내에서도 문제가 **되고** 있습니다.
473. 사건 주변에 **여자만** ( ) **떠오르면** 이상하리만치 집착하는 수사관의 일방적인 경향에

다 남경사가 제 기분에 취해 약간 과장하는 바람에 더욱 비정상적이 되어버린 그녀의 상(상)이 그에게는 수사불충분으로 보인 것임에 틀림없었다.

474. 옛것을 준수하고 모방하는 것만이 전통을 찾는 것인 **줄** ( ) **알다가는** 그 소중한 전통을 잃고 말 것이다.
475. 원시불교에서는 나한은 번뇌를 단진한 성자라고 하였으나, 후세 대승불교에 와서는 **이것은** ( ) 자리독선을 도모하여 회신멸지의 열반을 이상으로 **한다고** 하여 멸시하였다.
476. 아울러 **미국은** ( ) 이번에야말로 제네바합의대로 남북대화의 선행 없이는 어떠한 **합의이행도** ( ) **중단시켜야** 한다.
477. 옆에서 손으로 목화씨를 뽑고 있던 **어머니는** ( ) 멀거니 현우를 **돌아다봤다.**
478. **노동수단이란** ( ) 노동자와 노동대상 사이에 개입하여 이런 대상에 대한 그의 활동의 정도체로서 기여하는 하나의 물적 존재 또는 물적 존재들의 복합체를 **말한다.**
479. 그 놈이 떠나고 하룻만에 **아들놈도** ( ) **떠났으니까.**
480. **자동차는** ( ) 원래 사람이 차지했던 공간에 뒤늦게 **침입해** 왔기 때문에 좀 사양하는 기세를 보이고 있지만 **본심은** ( ) **그렇지** 못하다.
481. **남경사는** ( ) 궤짝을 열고 노트 뭉치를 뒤지기 **시작했다.**
482. 먼 훗날 **사가(사가)들은** ( ) 이 80년대를 가공할 자연의 섭리 파괴 연대로 **대서특필** ( ) **할** 것이 분명하다.
483. 수백만 년 지구의 역사를 통하여 자연이 마련해 두었던 **자원은** ( ) 그 양이 **한정되어** 있으나 그 자연의 젖줄에 매달리는 사람들이 많아졌다.
484. 빼고 나면 꼭 이겨야 한다는 외부 압력이며, 사특한 내부 욕심이며, 짐으로써 밀어닥칠 차가운 눈총이며 기량에 무리를 줄 심적 요인이 **눈** ( ) **녹듯** 녹아버린다.
485. 그렇다고 무턱대고 지도자를 자주 갈아 대어야 좋다는 **것은** ( ) **아니다.**
486. **양지(양지)는** ( ) 모두가 산성지라 인쇄물의 수명이 고작 1백 년을 못 **넘는다.**
487. 모든 것 하나하나가 **현우** ( ) **같은** 따위로서는 감히 참례해서는 안 되는 **세상** ( ) **같기만** 해서 쓸쓸하기 그지없었다.
488. 선거 **선진화** ( ) **이룩하자/**
489. **붕괴설까지** ( ) **나오는** 북한의 불안정한 상황, 북한에 대한 인식차이에서 나타나고 있는 한·미간의 이견, 독도문제로 깊어진 한·일간의 외교적 앙금 등 우리의 안보와 직결된 문제가 한두가지가 아니다.
490. 자신의 삶을 **초입부터** ( ) **헝클어논** 세계와 인생에 대한 의문이며 끝내는 분노와 격정으로 변해 귀중한 젊은날의 일부를 위악의 수렁 속에서 비틀거리게 한 종족의 오래된 신에 대한 실망을 달래어 줄 새로운 진리와 신을 찾아서였다.
491. **줄은** ( ) 10미터 **가량이나** ( ) 거저 먹히듯이 급속도로 **끌려갔다.**
492. 태극 무늬처럼 간소화된 **것도** ( ) **있고** 영구성을 나타내는 거북 모양의 구체적 **상징도** ( ) **있다.**
493. 2차 대전 **후** ( ) 폐허가 **된** 로마에서 올림픽을 개최하지 않을 **수** ( ) 없게 **된** 이탈리아

정부에서 그 기금을 염출할 길이 없어 이탈리아 사람들이 즐기는 축구 경기에 복권을 건 **익은** ( ) **있다.**

494. **가압식은** ( ) 내용물이 쉽게 응고되고 한 **번** ( ) **사용하면** 내용물을 다시 넣어 **사용해야** 한다.
495. 우리 한국의 놋다리밟기가 공민왕을 도강시킨 데서 비롯됐다 하듯이 부탄의 **놋다리밟기도** ( ) 그 많은 계곡을 건너는 도강 습속이 경기화한 것이 **아닌가** 싶었다.
496. 여기에 **선학원까지** ( ) **합쳐** 총 7개 단체가 총무원 세력과 대립하였다.
497. 부친의 의문을 풀어 주는 **대신** ( ) 또 새로운 물음으로 자신이 펼쳐 가려는 논의에 **끌어들이는** 것이었다.
498. 3일 **후** ( ) **귀하는** ( ) 명신물산을 **방문했다.**
499. 또 7대기본방향 가운데 그린 GNP, 즉 녹색국민총생산의 **개념도입은** ( ) 환경행정의 획기적 발전으로 **평가된다.**
500. 물론 **우리는** ( ) 자본주의적 생산 그 자체가 노동과 자본의 부등가 교환에 의한 적대성에 의거함을 잘 **알고** 있다.

Appendix E

Confusion Matrices

(a) X: FullContext$_1$, Y: DCD$_0$ (Agree 73.43%, Kappa 0.55)

	NOM	ACC	LOC	DAT	INST	COM	Sum
NOM	388	6	24	0	2	3	423
ACC	85	82	11	0	1	0	179
LOC	30	5	93	0	0	3	131
DAT	5	0	0	0	0	0	5
INST	20	6	6	0	10	0	42
COM	2	1	1	0	0	10	14
Sum	530	100	135	0	13	16	794

(b) X: FullContext$_2$, Y: DCD$_0$ (Agree 72.17%, Kappa 0.53)

	NOM	ACC	LOC	DAT	INST	COM	Sum
NOM	380	6	26	2	5	4	423
ACC	86	78	12	0	3	0	179
LOC	29	4	94	0	1	3	131
DAT	3	1	0	1	0	0	5
INST	19	6	7	0	10	0	42
COM	2	1	1	0	0	10	14
Sum	519	96	140	3	19	17	794

(c) X: FullContext$_3$, Y: DCD$_0$ (Agree 71.66%, Kappa 0.53)

	NOM	ACC	LOC	DAT	INST	COM	Sum
NOM	376	8	27	4	5	3	423
ACC	85	75	15	0	4	0	179
LOC	26	3	98	1	0	3	131
DAT	4	0	0	1	0	0	5
INST	17	8	8	0	9	0	42
COM	2	1	1	0	0	10	14
Sum	510	95	149	6	18	16	794

Table E.1: *Pairwise confusion matrices between full context annotations and DCD$_0$*

(a) X: LimContext$_1$, Y: DCD$_0$ (Agree 71.16%, Kappa 0.53)

	NOM	ACC	LOC	DAT	INST	COM	Sum
NOM	367	3	17	68	23	4	482
COM	6	10	0	0	3	0	19
INST	4	0	11	8	2	0	25
ACC	14	1	8	84	10	1	118
LOC	30	0	6	19	93	0	148
DAT	2	0	0	0	0	0	2
Sum	423	14	42	179	131	5	794

(b) X: LimContext$_2$, Y: DCD$_0$ (Agree 66.62%, Kappa 0.47)

	NOM	ACC	LOC	DAT	INST	COM	Sum
NOM	337	15	32	13	19	7	423
ACC	68	85	15	2	9	0	179
LOC	27	12	82	1	6	3	131
DAT	1	1	0	3	0	0	5
INST	15	10	5	0	12	0	42
COM	3	1	0	0	0	10	14
Sum	451	124	134	19	46	20	794

(c) X: LimContext$_3$, Y: DCD$_0$ (Agree 69.23%, Kappa 0.50)

	NOM	ACC	LOC	DAT	INST	COM	Sum
NOM	361	12	28	8	10	4	423
ACC	68	81	20	1	6	3	179
LOC	24	8	93	1	0	5	131
DAT	2	2	0	1	0	0	5
INST	18	8	7	0	9	0	42
COM	3	1	0	0	0	10	14
Sum	476	112	148	11	25	22	794

Table E.2: *Pairwise confusion matrices between limited context annotations and DCD$_0$*

(a) X: FullContext$_1$, Y: DCD$_1$ (Agree 74.94%, Kappa 0.57)

	NOM	ACC	LOC	DAT	INST	COM	Sum
NOM	397	8	19	0	2	3	429
ACC	57	81	9	0	1	0	148
LOC	36	7	98	0	1	3	145
DAT	14	0	0	0	0	0	14
INST	24	3	8	0	9	0	44
COM	2	1	1	0	0	10	14
Sum	530	100	135	0	13	16	794

(b) X: FullContext$_2$, Y: DCD$_1$ (Agree 74.56%, Kappa 0.57)

	NOM	ACC	LOC	DAT	INST	COM	Sum
NOM	393	8	20	1	3	4	429
ACC	58	76	11	0	3	0	148
LOC	32	6	101	0	3	3	145
DAT	11	1	0	2	0	0	14
INST	23	4	7	0	10	0	44
COM	2	1	1	0	0	10	14
Sum	519	96	140	3	19	17	794

(c) X: FullContext$_3$, Y: DCD$_1$ (Agree 74.69%, Kappa 0.57)

	NOM	ACC	LOC	DAT	INST	COM	Sum
NOM	390	10	20	2	4	3	429
ACC	57	75	14	0	2	0	148
LOC	30	4	105	1	2	3	145
DAT	11	0	0	3	0	0	14
INST	20	5	9	0	10	0	44
COM	2	1	1	0	0	10	14
Sum	510	95	149	6	18	16	794

Table E.3: *Pairwise confusion matrices between full context annotations and DCD$_1$*

(a) X: LimContext$_1$, Y: DCD$_1$ (Agree 74.56%, Kappa 0.58)

	NOM	ACC	LOC	DAT	INST	COM	Sum
NOM	367	3	17	2	5	5	399
ACC	6	10	0	0	4	0	20
LOC	4	0	11	0	2	3	20
DAT	14	1	8	0	0	0	23
INST	30	0	6	0	14	0	50
COM	2	0	0	0	0	11	13
Sum	423	14	42	2	25	19	525

(b) X: LimContext$_2$, Y: DCD$_1$ (Agree 71.91%, Kappa 0.55)

	NOM	ACC	LOC	DAT	INST	COM	Sum
NOM	357	13	26	9	18	6	429
ACC	37	92	12	1	6	0	148
LOC	30	12	90	1	9	3	145
DAT	6	0	0	8	0	0	14
INST	19	6	6	0	13	0	44
COM	2	1	0	0	0	11	14
Sum	451	124	134	19	46	20	794

(c) X: LimContext$_3$, Y: DCD$_1$ (Agree 73.93%, Kappa 0.57)

	NOM	ACC	LOC	DAT	INST	COM	Sum
NOM	377	10	24	6	9	3	429
ACC	41	86	14	0	6	1	148
LOC	25	9	101	2	1	7	145
DAT	10	1	0	3	0	0	14
INST	21	5	9	0	9	0	44
COM	2	1	0	0	0	11	14
Sum	476	112	148	11	25	22	794

Table E.4: *Pairwise confusion matrices between limited context annotations and DCD$_1$*

(a) X: FullContext$_1$, Y: DCD$_2$ (Agree 77.46%, Kappa 0.61)

	NOM	ACC	LOC	DAT	INST	COM	Sum
NOM	411	6	22	0	1	3	443
ACC	48	84	7	0	1	0	140
LOC	36	6	100	0	1	3	146
DAT	14	0	0	0	0	0	14
INST	21	3	6	0	10	0	40
COM	0	1	0	0	0	10	11
Sum	530	100	135	0	13	16	794

(b) X: FullContext$_2$, Y: DCD$_2$ (Agree 77.08%, Kappa 0.60)

	NOM	ACC	LOC	DAT	INST	COM	Sum
NOM	408	6	23	0	2	4	443
ACC	47	79	11	0	3	0	140
LOC	34	6	101	0	2	3	146
DAT	11	1	2	2	0	0	16
INST	19	3	5	1	12	0	40
COM	0	1	0	0	0	10	11
Sum	519	96	142	3	19	17	796

(c) X: FullContext$_3$, Y: DCD$_2$ (Agree 76.95%, Kappa 0.61)

	NOM	ACC	LOC	DAT	INST	COM	Sum
NOM	402	8	24	2	4	3	443
ACC	49	78	11	0	2	0	140
LOC	30	4	107	1	1	3	146
DAT	11	0	0	3	0	0	14
INST	18	4	7	0	11	0	40
COM	0	1	0	0	0	10	11
Sum	510	95	149	6	18	16	794

Table E.5: *Pairwise confusion matrices between full context annotations and DCD$_2$*

(a) X: LimContext$_1$, Y: DCD$_2$ (Agree 76.07%, Kappa 0.60)

	NOM	ACC	LOC	DAT	INST	COM	Sum
NOM	367	3	17	2	7	6	402
ACC	6	10	0	0	5	0	21
LOC	4	0	11	0	1	3	19
DAT	14	1	8	0	0	0	23
INST	30	0	6	0	12	0	48
COM	2	0	0	0	0	10	12
Sum	423	14	42	2	25	19	525

(b) X: LimContext$_2$, Y: DCD$_2$ (Agree 72.67%, Kappa 0.56)

	NOM	ACC	LOC	DAT	INST	COM	Sum
NOM	366	15	26	12	18	6	443
ACC	33	91	9	0	6	1	140
LOC	30	12	91	2	8	3	146
DAT	7	1	0	5	0	0	13
INST	15	4	7	0	14	0	40
COM	0	1	0	0	0	10	11
Sum	451	124	133	19	46	20	793

(c) X: LimContext$_3$, Y: DCD$_2$ (Agree 74.31%, Kappa 0.57)

	NOM	ACC	LOC	DAT	INST	COM	Sum
NOM	383	12	24	9	10	5	443
ACC	37	83	13	0	7	0	140
LOC	26	9	104	0	0	7	146
DAT	10	2	0	2	0	0	14
INST	20	5	7	0	8	0	40
COM	0	1	0	0	0	10	11
Sum	476	112	148	11	25	22	794

Table E.6: *Pairwise confusion matrices between limited context annotations and DCD$_2$*

(a) X: FullContext$_1$, Y: SCD (Agree 76.57%, Kappa 0.60)

	NOM	ACC	LOC	DAT	INST	COM	Sum
NOM	401	8	18	0	2	3	432
ACC	43	83	6	0	1	0	133
LOC	47	5	105	0	1	3	161
DAT	14	0	0	0	0	0	14
INST	24	3	6	0	9	0	42
COM	1	1	0	0	0	10	12
Sum	530	100	135	0	13	16	794

(b) X: FullContext$_2$, Y: SCD (Agree 76.20%, Kappa 0.59)

	NOM	ACC	LOC	DAT	INST	COM	Sum
NOM	399	8	18	0	3	4	432
ACC	42	78	9	0	4	0	133
LOC	44	4	107	0	3	3	161
DAT	11	1	2	2	0	0	16
INST	22	4	6	1	9	0	42
COM	1	1	0	0	0	10	12
Sum	519	96	142	3	19	17	796

(c) X: FullContext$_3$, Y: SCD (Agree 76.57%, Kappa 0.60)

	NOM	ACC	LOC	DAT	INST	COM	Sum
NOM	395	10	18	2	4	3	432
ACC	42	78	10	0	3	0	133
LOC	40	2	113	1	2	3	161
DAT	11	0	0	3	0	0	14
INST	21	4	8	0	9	0	42
COM	1	1	0	0	0	10	12
Sum	510	95	149	6	18	16	794

Table E.7: *Pairwise confusion matrices between full context annotations and SCD*

(a) X: LimContext$_1$, Y: SCD (Agree 75.82%, Kappa 0.60)

	NOM	ACC	LOC	DAT	INST	COM	Sum
NOM	367	3	17	2	7	5	401
ACC	6	10	0	0	1	0	17
LOC	4	0	11	0	2	3	20
DAT	14	1	8	0	0	0	23
INST	30	0	6	0	15	0	51
COM	2	0	0	0	0	11	13
Sum	423	14	42	2	25	19	525

(b) X: LimContext$_2$, Y: SCD (Agree 73.05%, Kappa 0.57)

	NOM	ACC	LOC	DAT	INST	COM	Sum
NOM	363	15	23	10	15	6	432
ACC	25	92	10	1	5	0	133
LOC	38	12	94	2	12	3	161
DAT	7	0	0	6	0	0	13
INST	18	4	6	0	14	0	42
COM	0	1	0	0	0	11	12
Sum	451	124	133	19	46	20	793

(c) X: LimContext$_3$, Y: SCD (Agree 75.94%, Kappa 0.59)

	NOM	ACC	LOC	DAT	INST	COM	Sum
NOM	380	12	20	7	8	5	432
ACC	30	88	11	0	4	0	133
LOC	35	7	110	1	2	6	161
DAT	11	0	0	3	0	0	14
INST	20	4	7	0	11	0	42
COM	0	1	0	0	0	11	12
Sum	476	112	148	11	25	22	794

Table E.8: *Pairwise confusion matrices between limited context annotations and SCD*

Bibliography

Abney, Steven. 1997. Stochastic Attribute-Value Grammars. *Computational Linguistics* 23(4).

Banko, Michele, and Eric Brill. 2001a. Mitigating the Paucity-of-Data Problem: Exploring the Effect of Training Corpus Size on Classifier Performance for Natural Language Processing. In *Proceedings of the Conference on Human Language Technology*.

Banko, Michele, and Eric Brill. 2001b. Scaling to Very Very Large Corpora for Natural Language Disambiguation. In *Proceedings of the 39th Annual Meeting and 10th Conference of the European Chapter (ACL-EACL '2001)*, 26–33.

Bies, Ann, Mark Ferguson, Karen Katz, and Robert MacIntyre. 1995. *Bracketing Guidelines for Treebank II Style, Penn Treebank Project*. Philadelphia: University of Pennsylvania.

Bird, Steven, and Edward Loper. 2004. NLTK: The Natural Language Toolkit. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics Demonstration Session.*

Blaheta, Don, and Eugene Charniak. 2000. Assigning Function Tags to Parsed Text. In *Proceedings of the 1st Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL '00)*, 234–240.

Blake, Barry J. 1994. *Case.* Cambridge Textbooks in Linguistics. Cambridge University Press.

Brants, Thorsten, Wojciech Skut, and Brigitte Krenn. 1997. Tagging Grammatical Functions. In *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP '97).*

Buchholz, Sabine. 2002. Memory-Based Grammatical Relation Finding. Ph.D. thesis, Tilburg University.

Carletta, Jean. 1996. Assessing Agreement on Classification tasks: the Kappa Statistic. *Computational Linguistics* 22(2): 249–254.

Carletta, Jean, Amy Isard, Jacqueline C. Kowtko Stephen Isard, Gwyneth Doherty-Sneddon, and Anne H. Anderson. 1997. The Reliability of a Dialogue Structure Coding Scheme. *Computational Linguistics* 23(1): 13–31.

Carreras, Xavier, and Lluís Màrquez. 2001. Boosting Trees for Clause Splitting. In *Proceedings*

of the 5th Conference on Natural Language Learning (CoNLL '01), 73–75.

Cha, Jeongwon, Geunbae Lee, and Jong-Hyeok Lee. 2002. Korean Combinatory Categorial Grammar and Statistical Parsing. *Computers and the Humanities* 36(4): 431–453.

Chang, Suk-Jin. 1993. *Information-Based Korean Grammar*. Seoul: Language and Information Association. In Korean.

Cho, Namho. 2002. Hyeondae Gugeo Sayong Bindo Josa [Word Frequency Survey of Modern Korean]. Research report, The National Academy of the Korean Langauge.

Cho, Pyoungok, and Chyulhyung Ok. 1997. Construction of semantic hierachies of Korean Nouns. In *Proceedings of the 9th Conference on Hangul and Korean Information Processing*, 313–319. In Korean.

Choi, Hyeonbae. 1937/1983. *Urimalbon [Our Grammar]*. Jeongeumsa.

Choi, Jaehi. 1999. Gugeoui Gyeog Pyoji Bisilhyeon Hyeonsanggwa Uimi Haeseog [Case Marker Unrealisation Phenomenon and Semantic Interpretation in Korean]. *Hangeul* 245: 49–78.

Chomsky, Noam. 1981. *Lectures on Goverment and Binding Theory*. Foris.

Chomsky, Noam. 1986. *Barriers*. No. 13 in Linguistic Inquiry Monograph. MIT Press.

Chomsky, Noam. 1993. A Minimalist Program for Linguistic Theory. In K. Hale and S. J. Keyser, eds., *The View from Binding 20: Essay in Linguistics in Honor of Sylvain Bromberger*, 1–52. Cambridge, MA: MIT Press.

Chomsky, Noam. 1995. *The Minimaplist Program*. Cambridge, MA: MIT Press.

Chung, Hee Jung. 1988. '-E'leul Jungshimeulo Bon Tossiui Uimi: '-e'wa '-go, -leul'ui Uimi Bigyo [The Meaning of Particles with a Focus on '-e': Semantic Comparison on '-e' and '-go, -leul']. *Gugeohag* 17.

Chung, Hee Jung. 1998. A Study on Korean Nouns: on Syntactic Functions Based on Meaning. Ph.D. thesis, Dept. of Korean Language and Literature, Yonsei University.

Chung, Hoojung. 1999. Resolving Syntactic Ambiguity Using Lexical Information. Master's thesis, Department of Computer Science, Korea University. In Korean.

Chung, Hoojung, and Hae-Chang Rim. 2004. Unlexicalized Dependency Parser for Variable Word Order Languages based on Local Contextual Pattern. In *Proceedings of the Fifth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2004)*, vol. 2945 of *Lecture Note in Computer Science*, 112–123. Springer-Verlag.

Cohen, Jacob. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20: 37–46.

Collins, Michael J. 1999. A Head-Driven Approach to Statistial Natural Language Parsing. Ph.D. thesis, Department of Computer and Information Science, University of Pennsylvania.

Collins, Michael J., and James Brooks. 1995. Prepositional Phrase Attachment Through a Backed-off Model. In *Proceedings of the Third Workshop on Very Large Corpora*, 27–38.

Crystal, David. 2002. *A Dictionary of Linguistics and Phonetics (Language Library)*. Blackwell Publishers, 5th edn.

Dagan, Ido, and Shuly Wintner. 2004. Statistical and Learning Methods in Natural Language Processing. Available from `http://cs.haifa.ac.il/~shuly/teaching/04/statnlp/`.

de Lima, Erika F. 1997. Assigning Grammtical Relations with a Back-off Model. In *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP '97)*. Also available as `http://xxx.soton.ac.uk/ps/cmp-lg/9706001`.

Déjean, Hervé. 2001. Using ALLiS for Clausing. In Walter Daelemans and Rémi Zajac, eds., *Proceedings of the Fifth Workshop on Computational Language Learning (CoNLL-2001)*, 64–66. Toulouse, France.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum Likelihood From Incomplete Data via EM Algorithm. *Journal of Royal Statistical Society Series* 39: 1–38.

Eom, Mi-Hyun, Dae-Gyu Shin, Byung-Jun Lim, and Dongyul Ra. 1996. Resolving Structural Ambiguity of Korean Based on Multiple Parse Filtering. In *Proceedings of the 8th Conference on Hangul and Korean Language Processing*, 443–451. In Korean.

Eugenio, Barbara Di, and Michael Glass. 2004. The Kappa statistic: a second look. *Computational Linguistics* 30(1).

Ferro, Lisa, Marc Vilain, and Alexander Yeh. 1999. Learning Transformational Rules to Find Grammatical Relations. In *Proceedings of the third Conference on Computational Language Learning (CoNLL-99)*, 43–52.

Fillmore, Charles J. 1968. The Case for Case. In Emon Bach and R. T. Harms, eds., *Universals in Linguistic Theory*, 1–90. New York: Holt, Rinehart and Winston.

Fillmore, Charles J. 1969. Toward a Modern Theory of Case. In A. D. Reibel and S. A. Schane, eds., *Modern Studies in English: Readings in Transformational Grammar*, 361–375. Englewood Cliffs, N.J.: Prentice-Hall.

Hachey, Benjamin C. 2002. Recognising Clauses Using Symbolic and Machine Learning Approaches. Master's thesis, Department of Theorethical and Applied Linguistics, University of Edinburgh.

Hammerton, James. 2001. Clause identification with Long Short-Term Memory. In Walter Daelemans and Rémi Zajac, eds., *Proceedings of the Fifth Workshop on Computational Language Learning (CoNLL-2001)*, 61–63. Toulouse, France.

Heinz, W., and J. Matiasek. 1994. Argument Structure and Case Assignment in German. In J. Nerbonne, K. Netter, and Carl Pollard, eds., *German Grammar in HPSG*, 199–236. Stanford: CSLI Publications.

Heo, Ung. 1983. *Gugeohag [Korean Linguistics]*. Saemmunhwasa.

Higginbotham, James. 1999. Thematic Roles. In Robert A. Wilson and Frank C. Keil, eds., *The MIT Encyclopedia of the Cognitive Sciences*, 837–838. The MIT Press. Also available at `http://cognet.mit.edu/MITECS/Entry/higginbotham`.

Hong, Jaesung. 1986. Gyocha Jangso Boeo Gumun Yeongu [A Study on Locative Alternation Construction]. *Hangeul* 191.

Hong, Saman. 1987. *A Study on Korean Special Particles*. Hagmunsa, revised edn.

Im, Hong-Pin. 1972. Gugeoui Jujehwa Yeongu [A Study on Topicalisation in Korean]. *Gugeoyeongu* 28.

Im, Hong-Pin. 1979. 'Eul/leul' Josaui Uimiwa Tongsa [The Syntax and Semantics of the Particle '-eul/-leul']. *Hangughag Nonchong* 2. In Korean.

Im, Hong-Pin. 1987. *Gugeo-ui Jaegwisa Yeongu [A Study on Korean Reflexives]*. Seoul: Singumunhwasa. In Korean.

Im, Hong-Pin. 1993. A study on thesaurus of Korean words. Research report, National Academy of the Korean Language. In Korean.

Jelinek, Fredrick, and Robert L. Mercer. 1980. Interpolated Estimation of Markov Source Parameters from Sparse Data. In *Proceedings of Workshop on Pattern Recognition in Practice*, 381–397.

Kang, Beom-mo, Hyohyun Jang, and Jaemin Yun. 1998. Hangughak Munheonui Jeonsanhwalul Wihan TEI Buhohwa Banganui Eungyonggwa Hwagjang [Application and Extension of TEI Encoding Scheme for the Computerisation of Korean Studies Documents. Research report, Korea Research Foundation.

Kang, Beom-mo, and Hung-gyu Kim. 2001. 21st Century Sejong Project - Compiling Korean Corpora. In *Proceedings of the 19th International Conference on Computer Processing of Oriental Languages (ICCPOL 2001)*, 180–183. Seoul.

Kang, Beom-mo, and Hung-gyu Kim. 2004. Sejong Korean Corpora in the Making. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, 1747–1750.

Kang, Myung-Yoon. 1988. Topics in Korean Syntax. Ph.D. thesis, MIT.

Kang, Myung-Yoon. 1996. Ijung Jugyeog Gumune Daehan Choesojuuijeog Jeobgeun [A Mimimalist Approach to the Double Nomainative Construction]. *Hangugeohag* 4. In Korean.

Kang, Young-Se. 1986. Korean Syntax and Universal Grammar. Ph.D. thesis, Harvard University.

Kaplan, Ronald M., and Joan W. Bresnan. 1982. Lexical Functional Grammar: A Formal System For Grammatical Representation. In Joan W. Bresnan, ed., *The Mental Representation of Grammatical Relations*. Cambridge, Mass.: MIT Press.

Katz, Slava M. 1987. Estimation of Probabilities from Sparse Data for the Language Model

Compoment of a Speech Recogniser. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 35(3): 400–401.

Kim, Gwi Hwa. 1994. *A Study on Case of the Korean Language*. Seoul: Hankuk Munhwasa.

Kim, Hung-gyu, Jongsun Hong, and Haechang Rim. 2001. 21segi Sejonggyehoeg Geugeo Gichojalyo Guchug [21st Century Sejong Project Korean Basic Resource Constuction]. Reserarch report, Korean Ministry of Culture and Tourism.

Kim, Hung-gyu, and Haechang Rim. 2003. 21segi Sejonggyehoeg Geugeo Gichojalyo Guchug [21st Century Sejong Project Korean Basic Resource Constuction]. Reserarch report, Korean Ministry of Culture and Tourism.

Kim, Hung-gyu, Hombin Rim, and Haechang Rim. 2000. 21segi Sejonggyehoeg Geugeo Gichojalyo Guchug [21st Century Sejong Project Korean Basic Resource Constuction]. Reserarch report, Korean Ministry of Culture and Tourism.

Kim, Kwangjin, Younghoon Song, and Junghyun Lee. 1993. Implementation of the System Dividing Simple Sentences from Embedded Sentence in Korean. In *Proceedings of the 5th Conference on Hangul and Korean Information Processing*, 25–34. In Korean.

Kim, Kyounghee. 1996a. A Simple Sentence Separator using the Verb Pattern Classification Information of Nueral Network. Master's thesis, Dept of Computer Science, Inha University. In Korean.

Kim, Seon Ho. 1996b. A Prediction of Lexical Relation Based on Statistical Information. Master's thesis, Department of Computer Science, Yonsei University. In Korean.

Kim, Yongha. 1999a. *Hangugeo Gyeoggwa Eosunui Choesojuui Munbeob [The Minimalist Syntax of Korean Case and Word Order]*. Hangugmunhwasa.

Kim, Young-Hee. 1973. Hangugeoui Gyeogmunbeob Yeongu [A Study on Case Grammar for Korean]. Master's thesis, Yonsei University.

Kim, Young-Hee. 1998. Mupyogyeogui Jogeon [The Conditions of Unmarked Case]. In *Hangugeo Tongsaloneul Wihan Nonui [Arguments for Korean Syntax]*, chap. X, 263–292. Seoul: Hangukmunhwasa.

Kim, Young-Hee. 1999b. Bojogeowa Gyeogpyosi [Complements and Case Marking]. *Hangeul* 244: 75–109.

Kim, Youngjoo. 1990. The Syntax and Semantics of Korean Case. Ph.D. thesis, Harvard University.

Kim Sang., Erik F. Tjong. 2001. Memory-Based Clause Identification. In Walter Daelemans and Rémi Zajac, eds., *Proceedings of the Fifth Workshop on Computational Language Learning (CoNLL-2001)*, 67–69. Toulouse, France.

Krippendorff, Klaus. 1980. *Content Analysis: an Introduction to its Methodology*. Beverly Hills, CA: Sage Publications.

Landis, J. Richard, and G. G. Koch. 1977. The Measurement of Observer Agreement for Cat-

egorial Data. *Biometrics* 33: 159–174.

Lapata, Maria. 1999. Acquiring Lexical Generalizations from Corpora: A Case Study for Diathesis Alternations. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL '99)*, 397–404.

Lapata, Maria. 2001. The Acquisition and Modeling of Lexical Knowledge. A Corpus-based Investigation of Systematic Polysemy. Ph.D. thesis, School of Informatics, University of Edinburgh.

Lee, Hui-Feng, Insu Kang, and Jong-Hyeok Lee. 1998. Resolution of Ambiguous Grammatical Functions of Korean Using Conceptual Patterns and Statistical Information. In *Proceedings of the 10th Hangul and Korean Information Processing*, 261–266.

Lee, Hyun A, Jong Hyeok Lee, and Geun Bae Lee. 1997a. Noun Phrase Indexing using Clausal Segmentation. *Journal of Korean Information Science Society* 24(3): 302–311. In Korean.

Lee, Iksop, and S. Robert Ramsey. 2000. *The Korean Language*. Anbany: State Univerity of New York Press.

Lee, Kong Joo, Gil Chang Kim, Jae-Hoon Kim, and Young S. Han. 1997b. Restricted Representation of Phrase Structure Grammar for Building a Tree Annotated Corpus of Korean. *Natural Language Engineering* 3: 215–230.

Lee, Kong Joo, Jae-Hoon Kim, and Gil Chang Kim. 1997c. An Efficient Parsing of Korean Sentences Using Restricted Phrase Sturucture Grammar. *Computer Processing of Oriental Languages* 1997(1): 49–62.

Lee, Kong Joo, Jae-Hoon Kim, and Gil Chang Kim. 1997d. Probabilistic Language Model for Analyzing Korean Sentences. In *Proceedings of the 17th International Conference on Computer Processing of Oriental Languages (ICCPOL '97)*, 392–395.

Lee, Kwangho. 1988. *Gugeo Gyeogjosa 'eul/leul'ui Yeongu [A Study on the Case Particle '-eul/-leul' in Korean]*. Seoul: Tabchulpansa. In Korean.

Lee, Songwook, Tae-Youb Jang, and Jungyun Seo. 2001. The Grammatical Function Analysis between Adnoun Clause and Noun Phrase in Korean. In *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium (NLPRS '01)*, 709–713.

Lee, Songwook, Tae-Youb Jang, and Jungyun Seo. 2002. The Grammatical Function Analysis between Korean Adnoun Clause and Noun Phrase by Using Support Vector Machines. In *Proceedings of COLING2002*.

Lee, Sun-Hee. 2004. A Lexical Analysis of Select Unbounded Dependency Constructions in Korean. Ph.D. thesis, Dept. of Linguistics, Ohio State University.

Levin, Beth. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago: University of Chicago Press.

Li, Hui-Feng, Jong-Hyeok Lee, and Geunbae Lee. 1998. Identifying Syntactic Role of Antecedent in Korean Relative Clause using Corpus and Thesaurus Information. In *Pro-*

ceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL '98), 756–762.

Loos, Eugene E., Dwight H. Day, Paul C. Jordan, and J. Doublas Wingate. 1997. *Glossary of Linguistic Terms*. SIL International. Available from `http://www.sil.org/linguistics/GlossaryOfLinguisticTerms/Index.htm`.

Manning, Christopher D., and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Mitchell, Tom. 1997. *Machine Learning*. McGraw-Hill.

Molina, Antonio, and Ferran Pla. 2001. Clause Detection using HMM. In Walter Daelemans and Rémi Zajac, eds., *Proceedings of the Fifth Workshop on Computational Language Learning (CoNLL-2001)*, 70–72. Toulouse, France.

Nam, Kishim. 1972. *Gugeoui Wanhyeong Bomumbeob Yeongu [A Study on Complementation in Korean]*. Seoul: Tabchulpansa.

Nam, Kishim. 1993. *Gugeo Josa-ui Yongbeob: '-e'wa '-lo'leul Jungsimeulo [The Usage of Korean Particles: focusing on '-e' and '-lo']*. Seoul: Bagijeong. In Korean.

Nam, Kishim, and Youngeun Koh. 1993. *Pyojun Gugeomunbeoblon [Standard Korean Grammar]*. Seoul: Tab Chulpansa, revised edn. In Korean.

Nam, Yunjin. 1997. A Quantitative Study on Modern Korean Particles. Ph.D. thesis, Department of Korean Language and Literature, Seoul National University. In Korean.

Ohno, S., and M. Hamanishi. 1981. *New Synonym Dictionary*. Tokyo: Kadokawa Shoten. In Japanese.

Park, Hyun-Jae. 2000. Design and Implementation of Two-Level Clausal Segmentation System using Sense Information. Master's thesis, Dept of Information and Telecommunication Engineering, University of Inchon. In Korean.

Park, Sunham. 1970. 'Gyeogmunbeobe' Ibgaghan Gugeoui 'Gyeobjueo'-e Daehan Gochal [A Study on Double Nominatives in Korean based on the Case Grammar]. *Eohagyeongu* VI(2).

Patrick, Jon D., and Ishaan Goyal. 2001. Boosted Decision Graphs for NLP Learning Tasks. In Walter Daelemans and Rémi Zajac, eds., *Proceedings of the Fifth Workshop on Computational Language Learning (CoNLL-2001)*, 58–60. Toulouse, France.

Pollard, Carl. 1994. Towards a Unified Account of Passive in German. In J. Nerbonne, K. Netter, and Carl Pollard, eds., *German Grammar in HPSG*, 273–296. Stanford: CSLI Publications.

Pollard, Carl, and Ivan A. Sag. 1988. *Information-Based Theory of Syntax and Semantics. Volume 1: Fundamentals*. Stanford: CSLI, Stanford Univ.

Pollard, Carl, and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. Chicago and

Stanford: Univ. of Chicago Press and CSLI, Stanford Univ.

Ratnaparkhi, Adwait. 1998. Statistical Models for Unsupervised Prepositional Phrase Attachment. In *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL '98)*, 1079–1085.

Resnik, Phillip. 1993. Selection and Information: A Class-based Approach to Lexical Relations. Ph.D. thesis, University of Pennsylvania.

Russell, Stuart. J., and Peter Norvig. 1995. *Artificial Intelligence: A Modern Approach*. Englewood Cliffs, NJ: Prentice Hall.

Seo, Kwang-Jun, Ki-Chun Nam, and Key-Sun Choi. 1999. A Probabilistic Model for Dependency Parsing Using Ascending Dependency. *Computer Processing of Oriental Languages* 12(3): 309–323.

Seo, Sangkyu. 1999. Eon-eo Yeongu-ui Toguroseo-ui Keompyuteo[Computer as a Tool of Linguistic Research]. *Research on Language and Information* 1: 271–309. In Korean.

Seo, Young-Hoon. 1998. Hangugeo Gumun Tagged Corpus Guchuk Mich Kumun Bunseok Deiteo Sajeon Gaebal [Korean Tree Tagged Corpus Construction and Syntactic Analysis Data Dictionary Development]. Tech. rep., Korea Electronics and Telecommunication Research Institute.

Shin, Changsun. 1975. Gugeoui 'Jueo Munje' Yeongu [A Study on 'Subject Problem' in Korean]. *Munbeobyeongu* 2.

Siegel, Sidney, and N. John Castellan. 1988. *Nonparametric Statistics for the Behavioural Science*. New York: McGraw-Hill, 2nd edn.

Skut, Wojciech, Brigitte Krenn, Thorsten Brants, and Hans Uszkoreit. 1997. An Annotation Scheme for Free Word Order Languages. In *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP '97)*.

Sohn, Ho-Min. 1999. *The Korean Language*. Cambridge Language Surveys. Cambridge University Press.

Stevenson, Suzanne, and Paola Merlo. 2000. Automatic Lexical Acquisition Based on Statistical Distributions. In *Proceedings of the 18th International Conference on Computational Linguistics*, 815–821. Saarbrüken, Germany.

Sung, Gwangsu. 1974. Gugeo Gyeogmunbeob Silon [An Essay on the Korean Case Grammar]. *Inmunlonjib* 19.

Teufel, Simone. 2000. Argumentative Zoning: Information Extraction from Scientific Articles. Ph.D. thesis, School of Informatics, University of Edinburgh.

van Rijsbergen, C. J. 1979. *Information Retrieval*. London: Butterworths, 2nd edn.

Vapnik, Vladimir N. 1995. *The Nature of Statistical Learning Theory*. New York: Springer.

Viterbi, A. J. 1967. Error Bounds for Convolution Codes and an Asymptotically Optimum

Decoding Algorithm. *IEEE Transactions on Information Theory* IT-13: 1260–1269.

Woolford, Ellen. 1999. Grammatical Relations. In Robert A. Wilson and Frank C. Keil, eds., *The MIT Encyclopedia of the Cognitive Science*, 355–357. MIT Press.

Yang, In-Seok. 1972. Korean Syntax: Case Markers, Delimiters, complementation, and Relativization. Ph.D. thesis, Dept. of Linguistics, University of Hawaii, Honolulu.

Yang, Jaehyung, and Yung Taek Kim. 1993. Identifying Deep Grammatical Relations in Korean Relative Clauses Using Corpus Information. In *Proceedings of Natural Language Processing Pacific Rim Symposium '93 (NLPRS '93)*, 337–344.

Yang, Jaehyung, and Yung Taek Kim. 1994a. Korean Analysis using Multiple Knowledge Sources. *The Journal of the Korea Information Science Society* 21(7): 1324–1332. In Korean.

Yang, Jaehyung, and Yung Taek Kim. 1994b. A Statistical Approach to the Resolution of Ambiguous Grammatical Functions In Korean Noun Phrases. *The Journal of the Korea Information Science Society* 21(5): 808–815. In Korean.

Yang, Jaehyung, and Kwangseb Shim. 1999. Case Ambiguity Resolution Using Thesaurus and Subcategorisation Information. *The Journal of the Korea Information Science Society (B)* 26(9): 1132–1141. In Korean.

Yoo, Dongseok. 1995. *Gugeoui Maegae Byeonin Munbeob [Parameterised Grammar of Korean]*. Seoul: Singumunhwasa.

Yoo, Eun-Jung. 1993. Subcategorization and Case Marking in Korean. In Andreas Kathol and Carl J. Pollard, eds., *Papers in Syntax*, no. 42 in OSU Working Papers in Linguistics, 178–198. Department of Linguistics, Ohio State University.

Yoo, Hye Won. 2002. A Study on the Case Alternation Constructions in Korean. Ph.D. thesis, Dept. of Korean language and literature, Korea University. In Korean.

Yoon, Deok Ho, and Yung Taek Kim. 1989a. Analysis Techniques for Korean Sentences based on Lexical Functional Grammar. In *Proceedings of the 1st International Workshop on Parsing Technologies (IWPT '89*, 368–378.

Yoon, Deok Ho, and Yung Taek Kim. 1989b. A Study on the Analysis Methods of Korean Sentence using the Unknown GR Attributes on LFG. *The Journal of the Korean Information Science Society* 16(5): 434–444. In Korean.

Yoon, Jun-tae, Seon-ho Kim, and Man-seok Song. 1997. New Parsing Model Using Global Association Table. In *Proceedings of the 5th International Workshop on Parsing Technology*.

Yoon, Juntae. 1998. Syntactic Analysis for Korean Sentences Using Lexical Association Based on Co-occurrence Relation. Ph.D. thesis, Department of Computer Science, Yonsei University. In Korean.

Yu, Hyeongseon. 1995. Gugeoui Jugyeog Jungchul Gumune Daehan Tongsa·Uimilonjeog

Yeongu [A Syntactic/Semantic Study on Korean Double Nominative Construction]. Ph.D. thesis, Korea University. In Korean.

Yu, Hyun Kyung. 1997. A Classificatory Study on Korean Adjectives. Ph.D. thesis, Dept. of Korean Language and Literature, Yonsei University, Seoul. In Korean.

Yu, Hyun Kyung, and Sun-Hee Lee. 1996. Gyeog Josa Gyochewa Uimiyeog [Case Particle Alteration and Semantic Roles]. In *Gugeo Tongsalonui Munjewa Jeonmang [The Problems and the Prospect of Korean Syntax]*, no. 3 in Gugeo Munbeobui Tamgu [Research on Korean Grammar], 129–171. Seoul: Taehagsa.