The Analysis of Japanese Relative Clauses

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF INFORMATION ENGINEERING
OF THE TOKYO INSTITUTE OF TECHNOLOGY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF ENGINEERING

Supervised by Prof. Hozumi Tanaka

Timothy Baldwin
February 1998

Contents

1 Introduction **1**
- 1.1 Objectives and outline . . . 1
  - 1.1.1 Statement of purpose of this research . . . 1
  - 1.1.2 Methodological outline . . . 2
  - 1.1.3 Applications of this research . . . 2
- 1.2 A basic model of Japanese syntax . . . 2
- 1.3 Definitions . . . 3
  - 1.3.1 Coordination, cosubordination and subordination . . . 3
  - 1.3.2 Displaceability . . . 4
- 1.4 Thesis overview . . . 4

2 Background **5**
- 2.1 The structure of Japanese relative clauses . . . 5
- 2.2 Past research . . . 6
  - 2.2.1 Descriptive accounts . . . 6
  - 2.2.2 Relativisation and thematisation . . . 9
- 2.3 Relative clause type definitions . . . 9
  - 2.3.1 Case-role gapping relative clauses . . . 9
  - 2.3.2 Head restrictive relative clauses . . . 12
  - 2.3.3 Full clause-based idioms . . . 13
- 2.4 Distribution of the relative clause types . . . 13
- 2.5 The full relative clause type hierarchy . . . 13

3 Valency, Argument Types and Case **15**
- 3.1 Complement/adjunct distinction . . . 16
- 3.2 Argument status . . . 18
- 3.3 Case set . . . 21
  - 3.3.1 Core grammatical cases . . . 21
  - 3.3.2 Other complement cases . . . 24
  - 3.3.3 Local cases . . . 26
  - 3.3.4 Time cases . . . 27
  - 3.3.5 Oblique cases . . . 27

4 Verb class-based resolution **29**
- 4.1 Verb class hierarchy . . . 29
  - 4.1.1 Argument status and interpretation preference . . . 29
  - 4.1.2 Constructing the verb class hierarchy . . . 31
- 4.2 Physical movement . . . 32
  - 4.2.1 Distal movement . . . 32

4.2.2 Proximal movement . . . 32
4.2.3 Travelling . . . 34
4.3 Relational . . . 34
4.3.1 Inter-personal relational . . . 35
4.3.2 Generic relational . . . 36
4.3.3 Part-whole relational . . . 36
4.4 Valence variational . . . 38
4.4.1 Conflated ergative . . . 38
4.4.2 Partitive . . . 39
4.5 Other verb classes . . . 40
4.5.1 Copula . . . 40
4.5.2 Conjoining . . . 40
4.5.3 Quantative . . . 40
4.5.4 Existential . . . 41
4.5.5 Locational action . . . 42
4.5.6 Tool-aided action . . . 43
4.6 Quotative . . . 44
4.6.1 Subordinate clause gapping . . . 44

5 Miscellaneous processing **46**
5.1 Non-gapping expressions . . . 46
5.1.1 The extraction of non-gapping expressions . . . 46
5.2 Time-related adjuncts . . . 47
5.2.1 Temporal masking . . . 47
5.2.2 Time relative constructions . . . 48
5.2.3 Temporal expressions . . . 49
5.2.4 Temporal vs. Durational interpretations . . . 49
5.3 Cardinal adjuncts . . . 50
5.4 The default rule set . . . 50

6 Lexical ambiguity **52**
6.1 Verb lexical ambiguity . . . 52
6.2 Resolving verb lexical ambiguity . . . 53
6.2.1 Calculation of verb scores . . . 53
6.2.2 Complexity of inflectional content . . . 54
6.2.3 Evaluation of verb scoring . . . 54
6.3 Noun head lexical ambiguity . . . 55

7 Extensions to the basic algorithm **57**
7.1 Relative clause cosubordination . . . 57
7.1.1 Processing of clause cosubordination . . . 58
7.1.2 Gap correspondences . . . 58
7.1.3 The treatment of subordinate clauses . . . 59
7.2 Coordinated relative clauses . . . 59
7.3 Noun head coordination . . . 60

8 Evaluation **61**
8.1 Evaluation criteria . . . 61
8.1.1 Baseline evaluation . . . 61
8.2 Overall evaluation . . . 62

8.2.1 Evaluation of fixed expressions . . . 63
8.2.2 Evaluation of subordinate gapping . . . 64
8.2.3 Evaluation of clause cosubordination . . . 65
8.2.4 Evaluation of inter-personal relational verbs . . . 65
8.3 Verb class-specific evaluation . . . 67

9 Conclusions **74**
9.0.1 Future research . . . 75

A The Relative Clause Resolution System **76**
A.1 The system valency dictionary . . . 76
A.1.1 Valency frame types . . . 76
A.1.2 Valency frame type preferences . . . 79
A.1.3 Dictionary entry content . . . 80
A.2 NTT valency dictionary . . . 81
A.3 System valency dictionary extraction . . . 81
A.3.1 Manual correction/modification of the system valency dictionary . . . 82
A.3.2 Verb inflectional analyser . . . 83
A.3.3 Valency frame compatibility analyser . . . 83
A.3.4 Semantic dictionary / lexical analysers . . . 83

List of Tables

4.1 Distribution of verb class frequencies 32
4.2 Verb class frequency 33

6.1 Results for the verb scoring methods 55

8.1 Overall analysis 62
8.2 Analysis of fixed expressions 64
8.3 Results of subordinate gapping analysis 64
8.4 Results of cosubordinated clause analysis 65
8.5 Analysis of inter-personal relational verbs (without reciprocity) 66
8.6 Analysis of inter-personal relational verbs (with reciprocity) 66
8.7 Analysis of conflated ergative verbs 67
8.8 Analysis of Conjoining verbs 67
8.9 Analysis of the copula 68
8.10 Analysis of distal movement verbs 68
8.11 Analysis of empathy verbs 68
8.12 Analysis of excluding verbs 69
8.13 Analysis of exisential verbs 69
8.14 Analysis of experiential verbs 69
8.15 Analysis of generic relational verbs 70
8.16 Analysis of idioms 70
8.17 Analysis of including verbs 70
8.18 Analysis of locational action verbs 71
8.19 Analysis of non-quantative verbs 71
8.20 Analysis of partitive verbs 72
8.21 Analysis of proximal movement verbs 72
8.22 Analysis of quotative verbs 72
8.23 Analysis of tool-based action verbs 73
8.24 Analysis of travelling verbs 73

A.1 Case valence dictionary statistics 83

List of Figures

1.1 A valency frame model of Japanese syntax 3

2.1 The structure of Japanese relative clauses 6
2.2 The relative distributions of the proposed Japanese relative clause types 14
2.3 The relative clause type hierarchy and associated sub-classifications 14

4.1 The full verb class hierarchy (Original verb classes indicated in bold, partitioning nodes capitalised) 30

A.1 The full system hierarchy 77

Acknowledgements

First and foremost, I would like to express my heartfelt gratitude to Prof. Hozumi Tanaka in supervising me over the past two years, giving me the opportunity to come to Japan, and making every step of the journey to reach this point smooth and enjoyable. At the same time, Assoc. Prof. Takenobu Tokunaga has generously given of his time to help me take my first shakey steps towards becoming a genuine researcher. Thanks go also to the long-suffering residents of Room 807 and all other members of the Tanaka and Tokunaga laboratories, in being asked for opinions on reams of example sentences, getting me and my computer back on speaking terms on numerous occassions, and for doing it all with a genuine smile all the way.

I doubt that this research would have progressed anywhere near the stage it has, without the assistance of the NTT Machine Translation group in providing access to their unsurpassed range of resources. In particular, I cannot thank Francis Bond enough for patiently reading papers in varying stages, sharing his work, and generally being a good friend throughout the past two years. I look forward to getting back to my definiteness 'roots' with Francis now that I have this little fellow out of the way!

And most importantly, I would like to thank family and friends for being patient and understanding while I learnt how to do things the hard way. I would like to say that I will devote more of myself, my time, and my energy to being a better husband, brother, son, and friend from now on, but realistically know that I am only going to turn around and plunge myself back into the cesspool of relative clauses and put you all through more of the same!

Abbreviations

ABL	Ablative
ACC	Accusative
ALL	Allative
CAUSE	Causative
COM	Comitative
DAT	Dative
DIM	Familial Diminutive
GEN	Genitive
LOC	Locative
MUTUAL	Mutual
NEG	Negative
-NML	Nominaliser
NOM	Nominative
PASS	Passive
PAST	Past tense
PRES	Non-past tense
PROG	Progressive
QP	Question Particle
QUOT	Quotative
REN	Ren'yo
TOP	Topic
WHILE	While (nagara)

Chapter 1

Introduction

1.1 Objectives and outline

1.1.1 Statement of purpose of this research

The purpose of this research is to classify Japanese relative clauses according to the nature of the semantic relationship between the relative clause body and the noun head. We consider the relative clause construal process as being both syntactically and semantically governed, and weighted pragmatically. By this is meant that for any given relative clause complex, there will be a well-defined set of semantic links which can be drawn between the clause body and head, based on both case-role interpretation and head restriction from the relative clause; these semantic links will each be constrained by such factors as sortal preferences on the head, inflectional restrictions on the main verb, and case slot instantiation. In order to extract and label these semantic links, we propose a theory of case-role types and verb class correspondences, and assign a mini-rule set to each verb class for use by the relative clause resolution system.

The job of the system is first to syntactically and semantically determine the scope of the relation set, and then to apply pragmatic preferences to determine the plausibility of each such relation. The final system output should consist of an ordered list of the final interpretation candidates, with the highest ranking interpretation taken as the final unmarked interpretation, barring further adjustment of the relative weights by external modules.

To take an example relative clause complex of *manzokusita gakusei* "a satisfied student", the system should first be able to determine the unacceptability of case-role gapping interpretations such as Direct Object, Indirect Object and Co-actor, syntactically from the valency frame of *manzoku(-suru)*. At the same time, the semantics of the verb and noun head should lead to the disallowment of specialised head restrictive senses such as Inclusive and Exclusive, and adjunct interpretations of the locative and temporal types. This would leave the two candidates of Subject case slot gapping and relative clause-based head restriction, from which the system would be expected to correctly select the Subject gapping interpretation.

An additional sub-purpose of this research is to construct a broad-coverage verb class hierarchy which can potentially be used to predict valency frame alternations/transformations (in the manner of Levin (1993)), and case-role correspondences. In this, the role of verb classes is twofold: (i) to model relations between case-roles within the valency frame, and (ii) to document the semantic type of the situation/action described by the verb. It is hoped that this classification will have wider ranging applications to the analysis of discourse processes, and can be used both in extracting zero pronominal instances from Japanese text and describing the interaction of verb arguments within the discourse.

Lastly, argument types will be introduced to rank case-roles according to contribution to the predicate sense, and apply heuristics with which to predict argument functionality.

1.1.2 Methodological outline

The methodology for analysing relative clauses devised in this research has been formulated so as to minimise use of extra-linguistic (i.e. pragmatic) resources in extracting the semantic relationship between the head noun and its associated relative clause. Importantly, we maximise reliance on syntactic and morphological analysis to this end, and use only limited semantic classification techniques throughout the extraction process. While this may appear contradictory to the semantic motivation underlying the research, it is intended to demonstrate the intrinsic relationship between (syntactic) case marking and case-role identification in Japanese, and the syntactic predictability of inflection-based valence transformation. We thus do not necessarily wish to oppose ourselves to the semantic/pragmatic-founded work of authors such as Sato (1989), Matsumoto (1990, 1996, 1997), Neumann (1994), Kanzaki (1997), and Kanzaki and Isahara (1997), and willingly admit that pragmatic and discourse-related factors can override syntactic and semantic preferences. However, given that the current work is focused at relative clause processing in isolation, without access to discourse and situational context, this design decision to predict *unmarked* relative clause construal would appear unavoidable.

At the same time, all effort has been made to develop a modular system which can hypothetically interface with a discourse processor to determine the most plausible interpretation from the ranked candidates outputted by the proposed system, and potentially reweight them according to such factors as contextual salience and recency, as well as sortal preferences for the head noun in performing given case-roles.

1.1.3 Applications of this research

Japanese involves the extensive usage of relative clause constructions. Given that approximately 50% of sentences contain one or more relative clauses, and that there are approximately two relative clauses for every three sentences in an average text,[1] the accurate analysis of relative clauses would seem to be a vital element of any overall system. This in itself would seem to justify any attempt to provide an accurate, robust algorithm for analysing Japanese relative clauses.

Theoretically, the method presented herein should be congruent with any NLP system, given its self-containing nature and relatively low-cost interface. Particular applications which would benefit most from accurate relative clause analysis would be Japanese language understanding tasks, including information retrieval and document extraction, and machine translation methods from Japanese to a Western language such as English[2]. Admittedly, however, in order to offer tangible results to the machine translation community, further work is required to transfer the analysis types proposed in this research over to translation equivalents in the target language.

1.2 A basic model of Japanese syntax

Case and Valency provide valuable tools in describing Japanese syntax, and are called upon frequently throughout this paper. The predicate is taken to be the nucleus of the clause, and relies on Valency to define the range and type of "case slots" subsumed by that predicate, according to the predicate modality. Each **case slot** is associated with a distinct "case-role" and "case marking" type. The **case-role** (aka. theta role, Case, semantic role, etc.) is an account of the semantic contribution of that case slot to the semantics of the predicate, although it is defined based on grammatical position

[1] The ratios given are based on an analysis of the EDR corpus (EDR 1995). Of the total of 201340 sentences, 99673 contained one or more relative clauses, and the total number of relative clauses was 133655.

[2] The applicability of the system and general methods utilised therein to language generation tasks is a matter for further research, although I suggest that the verb class hierarchy can potentially contribute to any Japanese language task relying on the notions of Case and Valency.

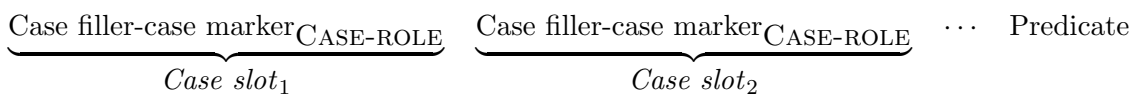

Figure 1.1: A valency frame model of Japanese syntax

for the core case-role set. To take an English example, *The party was held at David's house* contains a Subject and a Locative, in the form of *the party* and *at David's house*, respectively. Other examples of case-roles are Direct Object, Co-actor, Perlative, and Instrument.

Japanese marks each argument (**case filler**) with morphological case, in the form of a phrase-final case marker. Thus, in the Japanese equivalent for the above English example, *pāti-ha Dēbiddo no ie-de okonawareta* (party-TOPDavid GEN house-LOCwas held), the Subject *pāti* is marked with the topic marker and the Locative *Dēbiddo no ie* is marked with the locative case marker. These case-role/case marker tuples make up the content of each case slot.

That *Dēbiddo no ie* both performs the Locative case-role and is marked in the locative case is not entirely incidental, and peripheral case-roles commonly coincide with their canonical case marking type (another example of this is the Comitative case-role and comitative case marking), but that is not to say that this is either a necessary or sufficient condition. The case-role schema and case marking patterns should thus be considered as orthogonal issues, and the reader should bear in mind that in recovering the case-role of a given case slot, word order and the semantic content of the case filler also play integral roles.[3]

It is also true that a number of the proposed case-roles are associated with a unique case slot (in particular the Durational and Instrument case-roles), and that the core roles have a default case marker, but in no sense does the reverse apply, due to the conflation of case marking. To take an extreme case, the dative case marker (*ni*) can mark almost any case-role and provides minimal indication of the particular case-role of that case slot. For this reason of case marking variation, also, we consider case-roles and case marking as separate issues.

The **valency frame** for a given verb and verb sense is made up of individual case slots, with the scope of case marking for each slot often being plural. We will refer to the case marking paradigm for all case slots, considered in isolation of case-role correspondence, as the **case frame**. Case frames can thus be derived trivially from valency frames by discarding case-role information.

1.3 Definitions

1.3.1 Coordination, cosubordination and subordination

Clausal relations provide a valuable mechanism when analysing complex relative clauses, and are discussed variously throughout this thesis. In describing clausal relations, we apply a parameterised trichotomy comprising the "subordination", "cosubordination" and "coordination" types (Foley and Van Valin 1984; Van Valin 1984).

Two parameters are used to differentiate these three types: dependence and embedding. **Dependence** is form-based, and relates to whether the clause in question is syntactically dependent on surrounding clauses (either for operators or distributionally), or alternatively can stand alone as a complete sentence. **Embedding**, on the other hand, describes whether the clause is encapsulated within/functions as part of another clause, or is 'complete and distinct' (Van Valin 1984:542); herein, embedded clauses will be used to refer to quoted clauses marked with the quotative case marker.

[3]See (Blake 1994:13-8) for a discussion of this process in the context of a variety of language types.

Combining these two independent parameters, we in fact produce a four-way taxonomy, of which only the three classes given above are relevant for our purposes.

a.	**Coordination**	[−dependent, −embedded]
b.	Parenthesis	[−dependent, +embedded]
c.	**Cosubordination**	[+dependent, −embedded]
d.	**Subordination**	[+dependent, +embedded]

To clarify this distinction, *[David ran] [and Peter rode his bike]* is an instance of clause coordination, *[David ran] [because Peter was riding his bike]* is an instance of clause cosubordination (distributional dependence), and *[David ran to allow Peter [to ride his bike]]* is an instance of clause subordination.

1.3.2 Displaceability

Displaceability is a statement of the potential for a given case slot to be case-role gapped from a relative clause context, a notion original to this research. We wish to claim that it is possible to make an *a priori* judgement on the displaceability of a given case slot, independent of case filler type, and that if provided with a case filler which can be expressed in a displaceable case slot for a matrix clause instance, that case filler can also be gapped to become the noun head of a relative clause.

In the case of the verb *nar(-u)* "to become", for example, and the "Subject-NOMDirect.Object-DAT" valency frame, the Subject case slot is displaceable, but the Direct Object is not:

(1) a. *syōbōsi-ni* *natta* *kare*
fireman-DAT became he
(lit.) "he, who became a fireman"

b. * *kare-ga* *natta* *syōbōsi*
he-NOM became fireman
"the fireman he became"

1.4 Thesis overview

In Chapter 2, we define what is meant by "relative clause", and contextualise this research by giving a brief description of past treatments. We then describe the relative clause types around which this relative clause analysis pivots. Chapter 3 is an introduction to argument status, and their contribution to the proposed case-role schema. The full range of case-roles is illustrated with examples and descriptions of their synactic and semantic behaviour. In Chapter 4, we bring together both argument status and the case-role schema in detailing the verb class hierarchy proposed in this research, based upon which the modular rule sets applied in the resolution process are derived. Chapter 5 is an account of peripheral case-roles, and their interface to the verb class rule sets. Chapter 6 details the methods used to resolve lexical ambiguity both on the noun head and main verb, and again refers back to their relation to the rule sets. Chapter 7 describes inter-clausal and intra-noun head effects and their relevance to relative clause resolution, while Chapter 8 contains details of how the resolution system functions, what its parts are, and where they originate from. In Chapter 9, a thorough evaluation of the resolution system on different data sets is presented, including discussion of unexpected weaknesses in the system, before concluding the thesis in Chapter 10.

Chapter 2

Background

2.1 The structure of Japanese relative clauses

The general syntactic structure of Japanese relative clause complexes is given in figure 2.1 below. The noun phrase (NP) head is modified prenominally by a verb phrase (VP) clause body in comprising the overall noun phrase. We will refer to the NP head of the relative clause as the 'noun head' or 'head' throughout this paper, and the overall VP NP head-modified complex as a 'relative clause complex'.

The current research is aimed at verb-based relative clauses, but relative clauses incorporating adjectives and adjectival nouns are theoretically included under the generic category of VP, with the proviso that a copular connective (*na*) is required between the clause body and head in the case of adjectival nouns. Despite the structural parallelism that exists between verb-based and adjective-based relative clauses, however, it is suggested that there are certain semantic peculiarities displayed by adjectives which are not seen for verb-based relative clauses, drawing from the essentially stative nature of adjectives. Evidence of this can be seen in the work of Kanzaki (1997) and Kanzaki and Isahara (1997), but the absolute correlation between the two relative clause types is left as a moot question.

Semantically, the noun head and clause body stem verb cannot include clause-level grammatical constructions or grammatical markers which are unable to exist as independent NPs. This leads to the preclusion of the following types of VP NP constructions from our definition of relative clauses:[1]

a. Formal nouns/postadnominals (Martin 1975:664-740) such as *noti* "after", *tame* "for (the purpose of)" *bāi* "case/circumstance". These act as discourse/clause-level markers and unambiguously collocate with a clause body, deictic marker or noun specifier.

b. 'Relational' verb stems[2], as taken in the Hallidayean sense (see Halliday (1994:119-38)). This includes constructs of the type *to-iu* "called", *ni-kansuru* "concerning" and *ni-taisuru* "against/regarding".

It is important to note here that our definition of 'relative clause' for Japanese includes both NP complexes that involve case-role gapping (an adaptation of the traditional definition of relative clauses), and those for which the relative clause body simply restricts/exemplifies the noun head.[3] That is, our use of the term "relative clause" corresponds to Matsumoto's "noun-modifying construction" (1996), and differs significantly from the restricted Transformational Grammar sense of the word, which corresponds to the precept of 'case-role gapping' relative clausehood in our framework.

Admittedly, this appears to go beyond the bounds of the standard sense of relative clausehood, but the terminology is intended to reflect the syntactic parallelism that exists between these two relative

[1]There is limited scope to apply the methods described here to the processing of relative clause-type constructions produced with these operators, although this is left as a matter for future research.

[2]Referred to as "phrasal postpositions" by Martin (1975).

[3]Kameyama (1995), likewise, uses the term 'relative clause' in this wider sense.

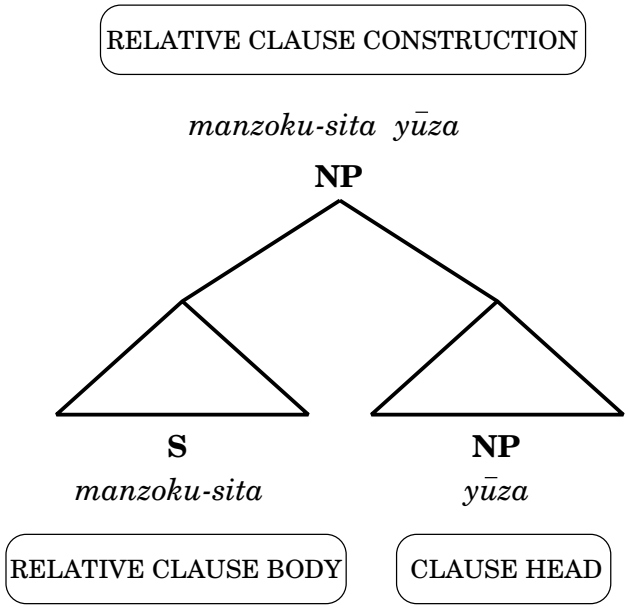

Figure 2.1: The structure of Japanese relative clauses

clause types. Additionally, while continual reference will be made to the notion of 'gapping', this is intended in a semantic case-role based sense, and we wish to distance ourselves from the research on syntactic gapping and movement that exists in the Chomskian literature (Nakau 1971; Okutsu 1974; Inoue 1976; Huddleston 1976; Shibatani 1978; Radford 1981). Clearly, similarities exist between the Chomskian treatment of gapping and its correlation to deep structure transformation, but, for our purposes, we view relative clause construal as a 'focusing' or 'aboutness' (Kuno 1976; Saito 1985; Kuno 1987) process rather than movement.

2.2 Past research

2.2.1 Descriptive accounts

Traditional descriptions of Japanese relative clauses have divided them into two main disjunctive categories, based on whether the noun head can be reinserted into the relative clause body to compose a matrix sentence (Martin 1975; Teramura 1975–78). Perhaps the most famous such account is that proposed by Teramura (1970, 1975–78, 1978, 1981, 1980), in which he describes the semantic relationship between the relative clause body and noun head either as the *uchi no kankei* "inner relationship" type or the *soto no kankei* "outer relationship" type, corresponding to the "sentence-insertable" and head noun content-supplementing sense designations, respectively. Following this typology, *[manzoku-sita] gakusei* "a satisfied student" is an inner relationship relative clause, and *[manzoku-sita] wake* "a reason for feeling satisfied" is an outer relationship relative clause.

On close observation, Teramura's schema attempts to classify inner relationship relative clauses as being syntactically-defined and outer relationship relative clauses as being semantically-defined, a point which is convincingly refuted by Matsumoto (1997) in citing cases where pragmatics influence interpretation of inner relationship relative clauses, such as:

(1) a. [*Donarudo Toranpu-ga* *katta*] *mise*
Donald Trump-NOM bought shop
"the store (which) Donald Trump bought"
b. [*Tomo-tyan -ga* *katta*] *mise*
Tomo-DIM-NOM bought shop
"the store (at which) little Tomo bought ( ) "
(from (Matsumoto 1997:43))

More recently, Saito (1985) reviewed the syntactic licencing of relative clauses, and suggested that syntax should account for the formation but not interpretation of inner relationship relative clauses. By way of relaxing the syntactic-dependence of inner-relationship-type relative clauses, it becomes possible to consider both inner and outer relationship type clauses in the same semantic light, and their equal susceptibility to pragmatic effects.

There is a further difficulty in defending the integrity of a Teramura-style proposal, however. Whereas the demarkation between inner and outer relationship relative clauses is portrayed as a distinct one, there are cases of relative clause complexes which seem to fit into both categories. Teramura recognises this shortcoming of his categorisation, and makes explicit mention of "truncated" constructions that straddle the divide between the two relationship types. An example of a truncated construction is:

(2) [*Tarō-ga* *pāti-ni* *deta*] *riyū*
Taro-NOM party-DAT attended reason
"Taro's reason for attending the party"

Within the unique indicated interpretation, the relative clause can be viewed both as exemplifying *riyū*, and as accommodating a gap for the noun head in the Purpose case slot:

(3) [[*Tarō-ga* *(nan no) riyū-de* *pāti-ni* *deta*] *no ka ha*]
Taro-NOM (what GEN) reason-LOC party-DAT attended -NMLQPTOP
watasi-ni *wakaranai.*
I-DAT not understand
"I do not understand [Taro's reason for attending the party]."

While this structural/analytical ambiguity seems pertinent for unmarked content noun heads, however, scope differences arise for internally modified noun heads, suggesting that the matter of truncation is not what it appears, and that the analytical ambiguity observed above is coincidental rather than terminal to Teramura's classification.

(4) [*Tarō-ga* *pāti-ni* *deta*] *hontō no riyū*
Taro-NOM party-DAT attended real GEN reason
"Taro's real reason for attending the party"

(5) ? *Tarō-ga* *hontō no riyū-de* *pāti-ni* *deta.*
Taro-NOM real GEN reason-LOC party-DAT attended
"Taro attended the party for a real reason." (≠ (4))

Similar observations can be made for time relative expressions (see Section 5.2.2), the other main source of truncation and contributor of claimed counterexamples to the inner/outer relationship dichotomy.

In Matsumoto's pragmatically-founded account of relative clauses, complexes such as (3) are interpreted as evidence for the inherent "framing" process involved in relative clause construal, through the predicate ("predicate framing") or the noun head ("nominal framing"), or both, with truncated relative clauses involving mutual predicate/nominal framing. Inner relationship relative clauses form a component of the predicate frame modifying type, and outer relationship relative clauses fall into the nominal and mutual predicate/nominal framing modifying types.

Matsumoto's frame semantic approach seems to provide valuable discrimintative power to the subdivision of Teramura's outer relationship relative clauses, both in terms of the source of the framing and additionally through the sub-type of the framing process. It lacks in credibility, however, in the semantic tenability of determining the relevant frame, identifying the available slots, and analysing compatibility between target arguments and the candidate slot set. Indeed, this first step of frame determination seems to be problematic in a cognitive context, even, for non-situational frames. For the well-defined experiential example of "eating", such as that described by Matsumoto for the predicate *tabe(-ru)* (Matsumoto 1997:61-63), the roles and frame are easily recoverable, but for a more abstract predicate such as "falling value", the location of an appropriate frame and role set would appear more confused and the scope of role matching less apparent. Additionally, Matsumoto's proposed methodology of evaluating case-role compatibility is not able to account for cases of bounded gapping.

Perhaps more serious, however, is that the discriminative nature of Matsumoto's expanded set of relative clause types produces its own demarkation problems. That is, as compared to the claimed difficulty in accounting for "truncated" relative clauses within Teramura's schema, Matsumoto's framework has inherent difficulties in delineating nominal framing ("NH-type") and mutual predicate/nominal framing ("CNH-type") relative clauses, and likewise nominal/predicate framing and predicate framing ("CH-type") relative clauses. To take an example from (Matsumoto 1997:159), *nioi* in (6) is suggested as both participating in the frame evoked by the modifying clause (predicate framing) and evoking its own "relational" frame to describe the cause or source of the smell (nominal framing). As such, (6) is described as being of the CNH-type.

(6) [*sakana-o* *yaku*] *nioi*
fish-ACC grill smell
"the smell of grilling fish"

This begs the question as to the status of (7) and (8).

(7) [*sakana-o* *yaku*] *kemuri*
fish-ACC grill smoke
"smoke from grilling fish"

(8) ? [*sakana-o* *yaku*] *oto*
fish-ACC grill sound
"the sound of grilling fish"

One addition body of research worthy of comment is the semantically-based work of Sato (1989) in proposing an expanded set of relative clause types. For all intensive purposes, Sato retains Teramura's inner relationship type as his *kaku yōso-gata* "case element-type", and partitions the outer relationship type into three subclasses, according to the nature of head modification. However, he then goes on to propose a fifth *kansetsu gentei-gata* "indirect restrictive-type", in which the noun head 'restricts a given case element in the relative clause by way of the genitive connective' (Sato 1989:9 – my translation). He cites examples of this effect such as (9), which corresponds to the genitively-connected matrix clause given in (10).

(9) [*tatemono-ga* *kowasareta*] *tosi*
building-NOM were destroyed city
"the city (where) buildings were destroyed"

(10) *(sono) tosi no tatemono-ga* *kowasareta*
(that) city GEN building-NOM were destroyed
"(the/that) city's buildings were destroyed"

Further discussion of this "indirect restrictive-type" will be made in defining bound gapping relative clauses (see below).

2.2.2 Relativisation and thematisation

One issue which has permeated the best part of (gapping) relative clause research, is the relationship between relativisation and thematisation. The first mention of this matter can be found in (Kuno 1973a:254-5), where Kuno proposes that relativisation correlates to deletion of the "theme" (topic) of the relative clause. A partial motivation for this analysis is the realisation that (true) topics cannot occur within relative clauses[4], and additionally that relativisation is basically a focusing phenomenon, similar to thematisation; there are also close syntactic parallels between the two processes, such as the deletability of case marking, and argument type compatibilities (both can involve any of adverbial clauses, complex noun phrases and sentential Subjects).

Despite the immediate problems inherent in Kuno's proposal (Muraki 1970; Kuno 1973a; McCawley 1976; Matsumoto 1997), it does offer an attractive guiding principle for differentiating between inner and outer relationship relative clauses, as most inner relationship relative clauses are sense-compatible with their topicalised noun head matrix equivalent, and outer relationship relative clause noun heads are almost exclusively not available to matrix topicalisation. We thus apply the topicalisation test implicitly as a litmus test to gauge the nature of relative clauses, but make no claims as to the equivalence or semantic coincidence of topicalisation and relativisation.

2.3 Relative clause type definitions

This research divides relative clauses into two major types, along the same lines as the traditional Teramura treatment, but adds a sub-type to the inner relationship type, by way of 'bound' relative clauses. Despite the obvious similarities to Teramura, we position ourselves very much away from the syntactic motivations of Teramura, and towards the semantic foundations of Matsumoto, Sato and Neumann.

2.3.1 Case-role gapping relative clauses

As previously described, "gapping" is used in a semantic sense to refer to the displacement of a case slot to the noun head position, in which respect case-role gapping relative clauses are distinctly removed from their origins in Teramura's inner relationship type. Within the "case-role gapping" class, we go on to define "bound" relative clauses, drawing on Sato's indirect restrictive-type.

Formally, case-role gapping is defined as the process whereby the noun head can be considered to fulfill a unique well-defined case-role subsumed within the relative clause body, for a given interpretation. That is not to say that all relative clauses have a unique interpretation, but rather that, given a particular interpretation, the case-role gap will be linked to a unique case-role within the relative clause. Considering example (11) below, the *gakusei* noun head can be seen to perform the role of the Subject of the main verb, *manzoku-sita*.

(11) [*manzoku-sita*] *gakusei*
was satisfied gakusei
"a satisfied student" (*lit.* a student who is satisfied)

(12) [*syōkai-sita*] *hito*
introduced person
a. "the person (who) introduced ( ) "
b. "the person (who) ( ) introduced"
c. "the person (to whom) ( ) introduced ( ) "

In contrast, the identity of the case-role gap for (12) is ambiguous between the Subject, Direct Object and Indirect Object case slots, but simultaneous gapping from these three case slots cannot occur.

[4]It is possible to have case slots marked with the topic marker, in non-topic contrastive usages.

That is, it is not possible to have an interpretation of the type *the person (who) introduced (self) ( )* or *the person (who) introduced (self) (to) (self)* without overt reflexive pronoun instances within the relative clause body. This leads to mutual exclusivity of case-role gapping between interpretations (12a), (12b) and (12c).

Even in the case of the personal reflexive pronoun *zibun* occupying a case slot within the relative clause, coindexing occurs clause-internally through case-role gaps/zero pronominal case-roles, rather than directly with the noun head.

(13) [ϕ_i *zibun*$_i$*-o* *syōkai-sita*] *hito*
SBJ self-ACC introduced person
a. "the person$_i$ (who) introduced him/herself$_i$"
b. "the person to whom ( ) $_i$ introduced him/herself$_i$"

For (13), this correlates to *zibun* being coindexed with the Subject case-role, which in turn corresponds to either the case-role gap (cf. (13a)) or some other independent, zero-pronominal discourse participant (cf. (13b)).

One feature of case-role gapping relative clauses is that whereas the case-role gap is defined uniquely for a given interpretation, the identity of the case slot from which gapping has occurred is not marked either as a trace within the relative clause, or as a relative pronoun-type marker. Moreover, there appear to be few restrictions on case-role positions from which gapping can occur, and when restrictions are found, they tend to be localised to that case-role in the given valency frame. Indeed, the main source of "restriction" is semantic, and derives from local sortal preferences defined through case frames.[5]

Returning to (12) above, any case-role ambiguity observable in the relative clause complex is removed in the matrix clause counterparts corresponding to the respective interpretations:

(14) a. *(sono) hito-ga* *syōkai-sita*
that person-NOM introduced
"that person introduced ( ) "
b. *(sono) hito-o* *syōkai-sita*
that person-ACC introduced
" ( ) introduced that person"
c. *(sono) hito-ni* *syōkai-sita*
that person-DAT introduced
" ( ) introduced ( ) (to) that person"

Bound relative clauses

Bound relative clauses are essentially Sato's indirect restrictive relative clause type, except that we consider the process to be semantically governed, and extend Sato's treatment to consider non-Subject case-role binding. The 'binding' facet of this relative clause type comes from its behaviour in anchoring a single attribute/'part' within the relative clause body. Note that the uniqueness of the gapped case-role seen for case-role gapping relative clauses, applies here for the bound case slot.

The basic binding function can be realised in a matrix clause context either as the topic or by genitive linkage to the bound case-role argument.

(15) a. [*10-gatu-kara* ***singakki-ga*** *hazimatta*] *pekin-daigaku*
October-ABL new term-NOM began Beijing University
"Beijing University, which began its new term in October"

[5]The actual process to determine which case slots case-role gapping can occur from, is largely left as a matter for future research.

b. *10-gatu-kara* ***pekin-daigaku no singakki-ga*** *hazimatta*
October-ABL Beijing University GEN new term-NOM began
"the Beijing University new term began in October"

c. ***pekin-daigaku-wa*** *10-gatu-kara* *singakki-ga* *hazimatta*
Beijing University-TOP October-ABL new term-NOM began
"the Beijing University new term began in October"

In (15a), *pekin-daigaku* binds the instantiated Subject case slot, a fact which is recoverable through the head-argument genitive linkage in the Subject position for (15b), and topic realisation in (15c). For comparison, (16) is an example of a Bounded Direct Object relative clause, although ambiguity exists with a canonical Subject gapping sense.

(16) [*hon-o* *katta*] *sakka*
book-ACC bought author
a. "the author who bought the book"
b. "the author, (whose) book ( ) bought"

For complement case slots, binding is distinguishable from case-role gapping through the surface instantiation of the case-role in question. Naturally, the bound case slot must be instantiated for bound relative clauses (whether complement or otherwise), and the gapped case slot must be vacant for case-role gapping. Hence simple knowledge of the bound case-role defines the relative clause construal type for complement case slots, due to their uniqueness of surface form. However, the repeatability of adjunct case slots (see Section 3.1) means that simple adjunct instantiation does not necessarily lead to case-role gapping incompatibility. For the Locative case slot, multiple instantiation produces 'inner' and 'outer' Locative positions, and gapping can occur from either in the presence of the opposing mate.[6] Likewise, for the Temporal case slot, case-role gapping can occur for a temporally ground relative clause given that 'temporal masking' constraints are met (see Section 5.2.1), as seen in (17) below.

(17) [*17-niti-ni* *kaigi-ga* *hazimatta*] *zikan*
17th-DAT meeting-NOM started time
"the time the meeting started on the 17th"

This leads to total incompatibility of binding with the Temporal case slot, and compatibility only with 'dangling local relational nouns' for the Perlative case slot. **Local relational nouns** describe a location relative to some fixed point, such as *mae* "front", *ue* "top" or *migi* "right". **Dangling local relational nouns** are produced when a bare local relational noun occupies a case slot. An example of a dangling local relational noun with the Perlative case slot is given in (18), which is contrasted with the ungrammatical dangling Allative usage in (19).

(18) [*hikōki-ga* ***ue-o*** *tonda*] *sima*
plane-NOM top-ACC flew island
"the island which the plane flew over"

(19) * [*hikōki-ga* ***migi-ni*** *tonda*] *sima*
plane-NOM right-DAT flew island
"the island which the plane flew to the right of" (intended)

Dangling local relational nouns are also produced, however, for passives, from the Perlative or Direct Object case slots, suggesting that simple detection of a dangling local relational noun is not sufficient in itself to produce a bound relative clause.

[6] For locational action verbs, the inability to reproduce inner locative case-role gapping to complement an outer locative instance within the relative clause, is a symptom not of the inconsistency of inner locative gapping itself, but of the impossibility to maintain the outer locative in an outer position without inner locative collocation.

(20) [*sihō-o* *yama-ni* *kakomareta*] *koya*
all four sides-ACC mountain-DAT is surrounded hut
"a hut surrounded on all four sides by mountains"

(21) [*pēzi-ga* *otiteiru*] *hon*
page-NOM is missing book
"a book with missing pages"

(22) a. *hon no pēzi-ga* *otiteiru.*
book GEN page-NOM is missing
"The pages of the book are missing."
b. *hon-kara* *pēzi-ga* *otiteiru.*
book-ABL page-NOM is missing
"The book has missing pages."

2.3.2 Head restrictive relative clauses

Head restrictive relative clauses display identical surface syntactic structure to case-role gapping clauses. In the case of head restrictive relative clauses, however, the head represents a consequence, condition, requisite, simultaneous event, etc. of the modifying clause (Matsumoto 1997:103-130), or is simply restricted by the semantic content of the relative clause. Examples of head restrictive relative clauses are:

(23) [*katu*] *isi*
wins will
"the will to win"

(24) [*au*] *kikkake*
meets chance
"a chance to meet ( ) "

(25) [*sakana-o* *yaku*] *kemuri*
fish-ACC grills smoke
"smoke from grilling fish"

(26) [*gakkō-ni* *itta*] *kaeri*
school-DAT went return
"on the way back from school"

(27) [*pāti-ga* *hirakareru*] *zenzitu*
party-NOM is held the day before
"the day before the party was/will be held"

In (23) and (24), the respective heads of *isi* and *kikkake* are simply restricted by their modifying relative clause, whereas *kemuri* in (25) is an inferrable consequence of the event described by the associated relative clause. (26) and (27), on the other hand, are headed by nouns which are relational in the given settings. For *kaeri*, some locative source from which it is possible to return, must be supplied to the noun head to ground the directional relation, and similarly, *zenzitu* requires a locative reference point before *the day before* can be calculated. We return to discussion of (27) and its relation to Temporal gapping relative clauses in Section 5.2.2.

One test for head restrictive relative clauses is their incompatibility with a topicalised matrix clause context.

2.3.3 Full clause-based idioms

Full clause-based idioms are NP idioms constructed as a relative clause complex. Notice that the full complex construes the idiomatic sense, in which way full clause-based idioms are distinguished from fixed expression sense in the relative clause.

Due to the idiomatic unit nature of full clause-based idioms, it seems meaningless to attempt to analyse constructions of this type by way of the proposed case-slot gapping/head restrictive relative clause dichotomy. They are thus excluded from the classification process, and simply marked as idioms on detection.

Examples of full clause-based idioms are:

(28) [*mite* *minu*] *huri*
to look to not look pretence
"close one eyes on ( ) "

(29) [[*Tarō-to* *awaseru*] *kao*] *ga* *nai*
Taro-COM put together face NOM have not
"I am ashamed [to show my face in front of Taro]."

For (29), it could be argued that *kao* has been case-role gapped from the Direct Object case slot, but this seems to gain no benefit in terms of idiom analysis/comprehension.

2.4 Distribution of the relative clause types

In any discussion of relative clause types, it is worth considering the relative distribution of each type, and hence their relative importance in any analysis attempt (Figure 2.2). This calculation of distribution was carried out on the annotated relative clause corpus sourced during system evaluation, which originates from the EDR corpus (EDR 1995). Given the significant size of the relative clause corpus (4615 disambiguated relative clause instances), we suggest that the indicated figures are indicative of the relative proportions the various types could be expected to occupy for similar-styled texts.[7]

From the figure, it is clear that case-role gapping relative clauses far outweigh head restrictive clauses (85% vs. 14%), and hence we feel justified in having focused to this point on the analysis of case-role gapping relative clauses. The 1% figure for bound relative clauses represents a significant proportion, although again, emphasis clearly lies on full case-role gapping.

2.5 The full relative clause type hierarchy

Figure 2.3 provides an account of the full relative clause type hierarchy, and the sub-classifications associated with each relative clause type. The heavy weighting of sub-classifications for case-role gapping is noticeable, and a fuller analysis of semantic linking types for head restrictive relative clauses would undoubtedly help offset this unbalance.

[7] The major component of the sentences contained in the EDR corpus are from newspaper articles, with lesser numbers of sentences from scientific texts and other miscellaneous sources.

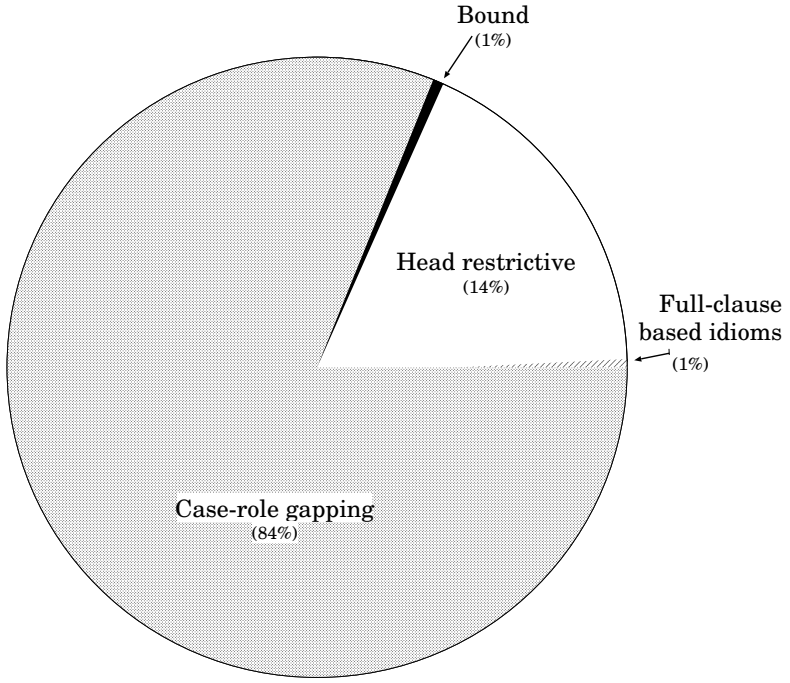

Figure 2.2: The relative distributions of the proposed Japanese relative clause types

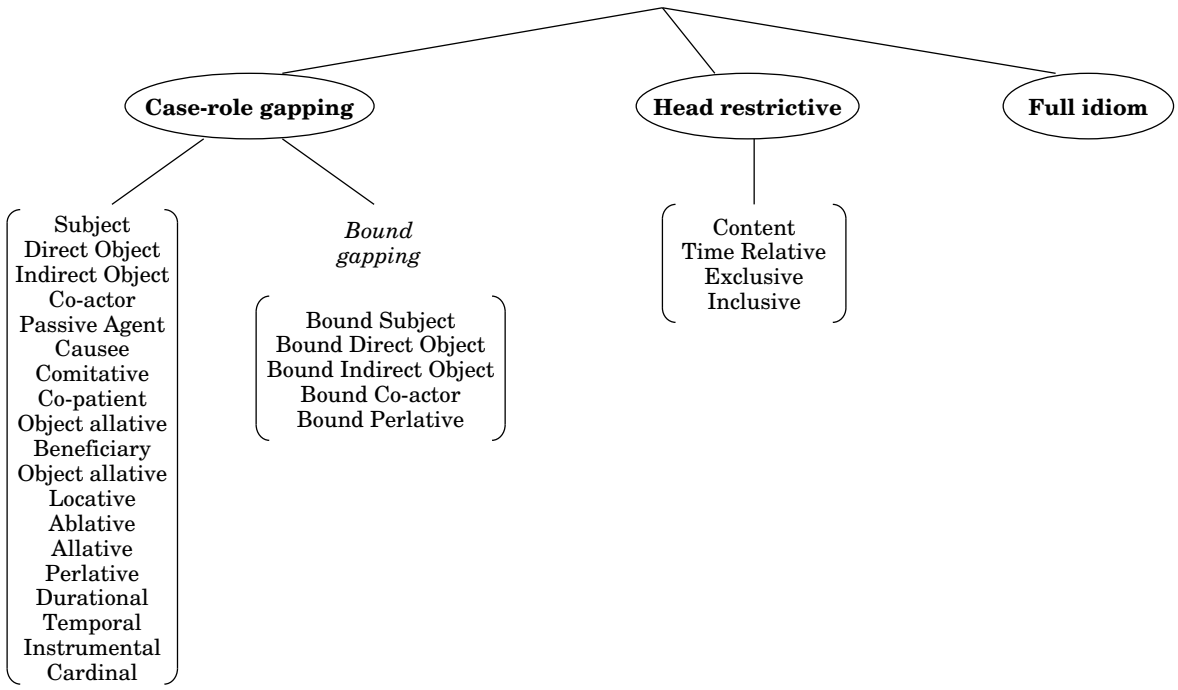

Figure 2.3: The relative clause type hierarchy and associated sub-classifications

Chapter 3

Valency, Argument Types and Case

In describing case-role gapping, we clearly require a set of case-roles powerful and wide-ranging enough to label all case slots in all valency frames. Additionally, in order to uniquely label each case slot contained in a given valency frame, the granularity of the case-role designation must be fine enough to account for semantic differences between case slots.

As an orthogonal issue, we introduce the concept of 'argument status', an expansion of the conventional complement/adjunct distinction. Argument status is used to predict argument obligatoriness, invaluable in distinguishing between zero pronominal and unrealised case slots, to introduce preferences between valency frames according to argument content, to gauge and generalise semantic consistency of usage, and to weight case-role gapping interpretations.

This final concern relates to which case-roles are preferred in a semantically neutral relative clause context of the type:

(1) [*taberu*] $\mathcal{X}$
eat

That is, assuming the semantics of the noun head $\mathcal{X}$ are inaccessible, what case-role interpretation (ignoring head restrictive relative clause interpretations for the time being) would be most likely? Inevitably, (obligatory) complements are preferred over optional complements, with the particular ranking of preference of complement case-roles often relating back to topicality/accessibility hierarchies (Keenan and Comrie 1977; Inoue 1976) such that, in the case of (1), the Subject case-role would be preferred over the Direct Object, followed by the Locative case-role, and possibly the Instrumental case-role. At the same time, however, the scope of case-roles available in a given clausal context is clearly constrained by the predicate valency frame, and it would not be possible to have an Indirect Object gapping interpretation, for example.

This illustrates the lowest level preference we can draw upon, defined by case-role immediacy and availability.

Working against any such default ranking are affinities for particular case-roles to take certain argument types in unmarked usages, and the potential for more specialised preferences for each case-role when in the context of a given valency frame and predicate sense. This first issue can be seen with the converse form of (1), in which the predicate is unspecified for a given lexical head:

(2) [$\mathcal{PRED}$] *basyo*
place

Here, the most likely case-role mapping would be onto the Locative case slot, *assuming locative adjunct compatibility for the predicate.* Failing any particular semantic correspondence to a predicate-independent case-role type, however, default preferences of the type seen above would apply. This exemplifies the opposite end of the scale of argument status, and adjuncy.

A third factor in this process is idiomatic usage, and overrides absolutely any local preferences

arising from case-role immediacy and adjuncy. For example, for the predicate *utu* "to strike" with an unmarked head $\mathcal{X}$, the most accessible case-role would be the Subject, and equally for the noun head *denpō* "telegram" with an unmarked predicate $\mathcal{PRED}$, adjuncy would not have any effect, but the moment these two are combined as *[utu] denpō* "the telegram ( ) sent", the Direct Object case-role overrules any preferences to take absolute preference. Equally, an unmarked instance of the predicate *a(-u)* "to coincide" would subsume a Co-actor case-role, but when collocating with the argument *keisan-ga* "calculation-NOM", the valency frame is reduced to a single topic position.

This chapter discusses an argument status hierarchy for use in predicting the syntactic interplay between arguments and predicates, and a case-role schema for documenting the semantic aspect of argument-predicate linkage.

3.1 Complement/adjunct distinction

The delineation between complements and adjuncts remains a contentious issue in valency research, with the most commonly occurring disagreement relating to the treatment of non-obligatory complements. Unlike adjuncts, which are by definition optional[1], complements are not necessarily obligatory. That is, all obligatory constituents are complements, and all adjuncts are omissible, but an optional argument can be either an adjunct or a complement.

In an attempt to defuse the controversy surrounding the classification process, we introduce a number of the more successful diagnostics that have been proposed to test for complement-hood/obligatoriness.[2]

Perhaps the best-known diagnostic is the **elimination test** (Helbig and Schenkel 1973)[3], in which the element in question is eliminated from the original sentence, and the derived sentence tested for grammaticality. In the case that a grammatical sentence is produced, the eliminated element is judged to have been non-obligatory, while if an ungrammatical sentence is produced, the element is seen to be syntactically obligatory.

(3) a. **The dog** ate **his dinner** eagerly.
b. * Ate **his dinner** eagerly.
c. **The dog** ate eagerly.

Based on the elimination test, *the dog* is obligatory in (3) (and therefore a complement), but *his dinner* is optional. At the same time, no indication is given as to the complement/adjunct status of *his dinner*, pointing to the need for orthogonal diagnostics to fully characterise a given sentence element.

In the case of Japanese, the applicability of the elimination test is highly restricted, in that most "obligatory" case slots are expressible with zero pronouns. Simple elimination of elements hence produces interpretational ambiguity as to the genuine non-existence of that element, and a zero anaphoric status, although disambiguation independent of sentential context is generally possible. While recognising this danger, however, we suggest that elimination at least provides an indicator in the obligatoriness evaluation process, and is valuable in comparing the eliminated contexts of different arguments, as arguments higher in the argument status hierarchy (see below) produce more distinct markedness.

A diagnostic which draws on the concept of "argument linking" is **repeatability**, suggested by Smith (1996:66). Here, Smith proposes that multiple instances of a given adjunct type generally produce grammaticality, whereas even if grammaticality is produced for multiple complement constituents, the produced interpretation will allocate distinct complement roles to each individual element. This

[1] In this, we wish to distance ourselves from Uszkoreit's (1987) claim of a full four-way adjunct/complement, optional/obligatory contrast in German. In terms of the valency binding hierarchy introduced below, Uszkoreit's "obligatory adjuncts" are treated as middles.

[2] Prominent tests not mentioned here include 'back-formation' (Helbig and Schenkel 1973), 'substitution' (Brinker 1972), 'passivisability' (Emons 1974), 'addability' (Heinz 1978) and the 'do-so' test (Somers 1984; Somers 1987).

[3] Refer to Somers (1987:13-4) for a description of problems associated with the elimination test and the closely related test for 'extractability' (Brinker 1972).

process can be exemplified with the following sentences.

(4) **Tarō-ga* ***Hanako-ni*** ***Jirō-ni*** *atta*
Taro-NOM Hanako-DAT Jiro-DAT met
"Taro met up with Hanako and Jiro" (intended)

(5) *Tarō-ga* ***Hanako-ni-mo*** ***Jirō-ni-mo*** *atta*
Taro-NOM Hanako-DAT-also Jiro-DAT-also met
"Taro met up with both Hanako and Jiro"

(6) *Tarō-ga* ***Hanako-to-Jirō-ni*** *atta*
Taro-NOM Hanako and Jiro-DAT met
"Taro met up with Hanako and Jiro"

(7) *tugi-wa* ***18-niti-ni*** ***2-zi-ni*** *kite-kudasai*
next-TOP 18th-DAT 2 o'clock-DAT please come
"Please come next on the 18th at 2 o'clock"

Naturally, this does not extend to cases of coordination for a given case position, but rather refers to separate surface realisations of that case slot. Hence, the acceptability of the in-case slot coordination in (6) and iterative *-mo -mo* construction in (5) do not threaten the integrity of the test.

The process of a distinct complement role being forced on a repeated element can be seen with the verb *pass* below, whereby *the salt* in (8c) is forced into an Indirect Object role, producing a structure paralleling that in (8b).

(8) a. Mary passed the salt
b. Mary passed Peter the salt
c. ?Mary passed the salt the pepper

One limitation of this diagnostic is its inability to account for the multiple-subject construction (Kuno 1973b:34, 68-78) in Japanese, where multiple nominative-marked constituents are generated in the Subject position.

(9) *yama-ga* *ki-ga* kirei desu
mountain-NOM tree-NOM pretty are
'The mountains have beautiful trees' adapted from Kuno (1973b:69)

Smith predicts this fact in identifying the nominative case marker as the default case marker in Japanese (p. 98). Because of this observation, however, repeatability is only applicable to non-subject case slots.

One last test worthy of mention, which is specific to Japanese, is **quantifier floating** (Kuroda 1980; Miyagawa 1988; Miyagawa 1989b). Quantifier floating occurs when a numeric classifier associated with a noun can be transposed to the right of the associated case slot. For this process to successfully occur, the noun phrase occupying the source case slot must necessarily be an obligatory element (Jacobsen 1992:41), and hence a complement. Quantifier floating correctly identifies the nominative case slot in the sentence pair of (10) as a complement.

(10) *gakusei-ga* *hutari* *kita* $\Longleftrightarrow$ *hutari no gakusei-ga* *kita*
student-NOM two (people) came two (people) GEN student-NOM came
"Two students came"

3.2 Argument status

This research hinges largely around the "argument status" of each case slot. In analysing argument status, we follow Bond and Shirai (1997) in combining the six-degree scale of valency binding proposed by Somers (1984, 1987), with the notion of 'shadow complements' as suggested by Pustejovsky (1995). The resultant argument status hierarchy is as follows.

1 Integral complement
2 Shadow complement
3 Obligatory complement
4 Optional complement
5 Middle
6 Adjunct
7 Extra-peripheral

Category 1 – Integral complements

Integral complements are generally constrained as to word order, are obligatory, have a lexically fixed head, and are not anaphorically realiseable. Additionally, Japanese integral complements display only minor inflection-driven case marker variation, with the accusative case alternating with the nominative for potential modality. However, there is often limited scope for modification within the complement, multiplicity of case marking in the unmarked form is observable, and adverbial insertion can occur between the complement and verbal stem in restricted contexts. An example of an integral complement is *the bucket* in *kick the bucket* (taken from Bond and Shirai (1997)), or in the case of Japanese *hidoime* in *hidoime-ni au* "to have a hard time of it". It is worthwhile noting that whereas *the bucket* cannot be modified internally, *hidoi-me* can be modified by the adverb *monosugoku* "extremely" to produce the accentuated idiom *monosugoku-hidoime-ni au* "to have a really hard time of it", an effect which is equally applicable to the English gloss.

Integral complements generate idiomatic verb sense, and multiple integral complement instances can occur within a single valency frame.

Category 2 – Shadow complements

Pustejovsky (1995:63-7) coined the term 'shadow complement' to describe 'parameters which are semantically incorporated into the lexical item' (p. 63), although the same phenomenon has been identified independently by Talmy (1996:241) in his treatment of 'blocked complements'. By this process, Pustejovsky refers to verbs such as *butter*, in which a default description of the material is entailed by the semantics of the verb, to the degree that a redundant surface realisation of that default material becomes ungrammatical (cf. (11b)). On the other hand, restricted instances of the default argument can be realised lexically (cf. (11c)), and replacement of the default argument by a synonym generally produces a grammatical usage (cf. (11d)).

a. Mary buttered her toast.
b. *Mary buttered her toast with butter.
c. Mary buttered her toast with freshly churned butter.
d. Mary buttered her toast with margarine.

adapted from Pustejovsky (1995:65)

An example of a shadow complement from Japanese is *zeikin-o* "tax-ACC" with the predicate *nōzei-suru* "to pay tax", which behaves analogously to *butter* above.

A side-effect of classifying shadow complements lower in the hierarchy than integral complements is the misrepresentation of them being somehow less tightly bound to the predicate than integral

complements. In actual fact, as was seen for *butter*, they can be perceived as being so tightly bound to the predicate as to be non-realiseable in an unmarked form. However, unlike integral complements, synonym replacement is generally allowable (see (11d) above), and restrictions on word order are relatively relaxed in cases when a surface complement representation is possible. At the same time, we can observe some scope for in-complement modification, as was observed for integral complements. Shadow complements are thus less rigidly restricted in surface form, supporting the given positioning below integral complements.

An additional feature of shadow complements is the cline of acceptance of unmarked instances of the default complement in a matrix clause context. To take the above case from Japanese, *zei-o nōzei-suru* was unanimously unacceptable to the native speakers consulted, whereas the matrix collocation of the shadow complement *byōin-ni* "hospital-DAT" with *nyūin-suru* "to go into hospital/be hospitalised" received a relatively neutral response. Given that our usage of shadow complements is aimed at text analysis, we avoid making grammaticality judgements for such unmarked matrix occurrences of the default argument. Despite this relaxation of the constrained nature of shadow complements, however, we maintain a treatment independent of that for integral complements. That is, integral complements are necessary to derive the associated idiomatic sense of the predicate, in comparison with surface realisations of shadow complements which simply reinforce/extend the inherent verb sense generated by the predicate.

This leaves open the question of the status of *song* in the construction *sing a song*. The default argument of *song* is unarguably encoded in the predicate, a fact which is evoked in the intransitive usage of *sing*, and only synonyms, hyponyms, and modified instances of *song* are allowed as Direct Object. The exclusion of arguments in constructions of this type from the shadow complement classification, stems from the acceptability of proper **hyponym** replacement, such as *sing a shanty* or *sing a rollicking tune you heard on the radio*. That is, semantic restriction on the Direct Object slot denotates a hierarchical semantic set of **both** synonyms and hyponyms, with the default of *song* at the root, unlike *butter* or *zei* "tax" which are highly restricted in themselves and are replaceable only with a limited range of synonyms, and modified instances of that default sense.

The inherent optional nature of shadow complements should be clear from the relative ungrammaticality of an unmarked surface occurrence of the default element. Application of the repeatability diagnosis, then, leads to the expected complement status of shadow complements, producing an optional complement (category 4) categorisation for shadow complements. On the surface, this would appear to cast doubt on the placement of shadow complements above obligatory complements within the valency binding hierarchy. We justify the given analysis from the observation that whereas elimination of surface shadow complements is possible, doing so reverts its semantic content to the default; for optional complements (see below), the same process of elimination simply leads to underspecification. Moreover, any potential for synonym replacement is highly constrained, to a much higher degree than for obligatory complements.

For the above reasons, the category 2 placement of shadow complements between integral complements and obligatory complements would appear to be well-founded.

Categories 3 & 4 – Obligatory/optional complements

Complements are defined as being 'strictly subcategorised by the predicate for semantic class, syntactic function ... and often syntactic form' (Somers 1987:28). Within this complement definition, however, we make a distinction between 'obligatory' and 'optional' complements, in line with the original valency-binding framework proposed by Somers (1987:27). In this, we diverge from the treatment given in (Bond and Shirai 1997).

It is perhaps easiest to describe the difference between obligatory and optional complements by way of an example, in which the two complement types are indicated in bold and underlined, respectively.

(12) a. **Peter** handed over **the documents**.
b. *Handed over **the documents**.
c. ***Peter** handed over.
d. **Peter** handed over **the documents** to the man in dark glasses.

Here, the elimination test clearly identifies both *Peter* and *the documents* as obligatory complements, on account of the ungrammaticality of (12b) and (12c), respectively. In direct comparison, *the man in dark glasses* is an optional constituent. This third, optional constituent would, however, appear to play an equally salient role in the activity portrayed, in that it states the Recipient of the goods, in direct opposition to the Source of *Peter*. Subsequent assessment of the Recipient slot with the repeatability diagnostic described in section 3.1 clearly supports this intuition as to the complement status of *the man in dark glasses*.

Through the combination of these above results, *to the main in dark glasses* can be seen to be an **optional** complement.

Unlike the more restrictive complement types given above, however, 'basic' complements (i.e. complements not subsumed within the specialised categories 0, 1 and 2 given above) are compatible with anaphoric processes. In the case of Japanese, this equates to basic complements being realisable with zero pronouns, an effect which clouds the obligatory/optional complement dichotomy. Note that for English, the grammar requires that anaphoric derivatives of obligatory complements have a surface lexical form, a fact which can be used to determine obligatoriness. In fact, the obligatory/optional complement distinction in English basically corresponds to whether that complement is *bound* by the grammar to have a surface form, noting the deceitful behaviour of verbs which display distinct transitive and intransitive usages such as *sing* (see above):

(13) a. Bill sang **the national anthem** with great gusto.
b. Bill sang with great gusto.

The only tool we offer to diagnose obligatoriness for a given complement position in Japanese is quantifier floating (see above). The complement status of optional complements can then be verified by way of repeatability.

Category 5 – Middles

Middles are proposed by Somers as an idiosyncratic 'in-between' classification made up of elements which share the characteristics of both complements and adjuncts. Naturally they are non-obligatory, but the same close association can be observed with the governing verb. Examples of middles taken from English are Instrumental, such as *with a hammer* in *hit the nail with a hammer*, and Beneficiaries, such as *the squire* in '*The gamekeeper shot the squire a rabbit.*' (taken from Somers (1987:25)). These two case types also generally produce middle elements in Japanese.

A significant class of middles particular to Japanese is that of onomatopaeic adverbials, and notably phonomimes and phenomimes (Shibatani 1990:153-7). The strong correspondence between onomatopaeic expressions and particular verbs supports this view, as is seen for *kusukusu* "titter", *nikoniko* "grin" and *kutukutu* "chuckle", which collocate only with the verb *wara(-u)* "to smile/laugh".[4]

Category 6 – Adjuncts

Adjuncts, again, are necessarily optional, but unlike middles tend to display semantic consistency across usage with distinct predicate classes. Naturally, pragmatic restrictions will exist as to local semantic compatibility with a given predicate, but, in general, their use is unpredictable. Co-occurrence of adjuncts of the same semantic type commonly occurs, as suggested by the repeatability test described above, and word order restrictions are relatively relaxed.

[4]All these expressions also occur with the light verb *suru* "to do" in an inherent 'laughing' sense.

An example of an adjunct is the Japanese Locative case slot, such as *kono-ie-de* "this house-LOC" in:

(14) *Tarō-ha* ***kono-ie-de*** *umaresodatta*
Taro-TOP this house-LOC was born and brought up
"Taro was born and brought up in this house."

(15) *asu-no-hiru-ni* ***kono-ie-de*** *matiawaseyō*
tomorrow-GEN-midday-DAT this house-LOC let's meet
"Let's meet at this house at midday tomorrow."

(16) ***kono-ie-de-wa*** *uti-no-inu-ga* *itumo* *tukue-ya-isu-no-sita-de*
this house-LOC-TOP us GEN dog-NOM always table and chair GEN under-LOC
netagaru
wants to lie
"(When) in this house, our dog is always wanting to lie under tables and chairs"

Sentence (16) displays the repeatable nature of adjuncts.

Category 7 – Extra-peripherals

Extra peripherals are optional sentence modifying constituents, and constitute the outermost argument category; as one would expect, they are almost impervious to both word order and semantic restrictions. In both English and Japanese, adverbs form the main component of extra-peripherals. Particular examples are *suddenly* and *often*, and *wazato* "intentionally" and *sorosoro* "soon".

3.3 Case set

To talk of case-role gapping, we must have a well-defined toolset of **case-roles** with which to tag case slots within the relative clause, and identify the source case slot in cases of gapping. The particular case-role paradigm employed in this research draws primarily on **grammatical relations** rather than strict case relations (cf. Fillmore's (1968) Case Grammar, Starosta's (1988) Lexicase, Somers's (1987) case grid). By this is meant that the 'core' case-roles are syntactically relational, drawing on the notions of 'Subject', 'Direct Object' and 'Indirect Object'.

At the same time, however, the argument status hierarchy proposed in Section 3.1 is evoked in deriving the intermediate 'Co-actor' and 'Co-patient' roles from the traditional Direct and Indirect Object case-roles. In addition, the 'grammatical' basis applies only to the core case-roles, and the remaining eleven peripheral case-roles are defined based on semantic criteria.

In order to validify the proposed case-role set, we endeavour to assign diagnostics to each case-role, as described in each case-role description below, although the reader is cautioned that these are specific to Japanese and no cross-lingual universality for either the case-roles or diagnostics is claimed.

The proposed 18-way case-role set is grouped according to argument status and semantic similarity, into the **core**, **other complement**, **local**, **temporal** and **oblique** sets, as follows.[5]

3.3.1 Core grammatical cases

'Core grammatical cases' are complements (obligatory, optional or otherwise), and are defined surface syntactically, in the manner of Relational Grammar (Perlmutter 1980; Blake 1990).

[5]For reference purposes, Machine Translation System Laboratory (MTSL) (1995), Nomura and Muraki (1996) and Bond and Shirai (1997) describe case-role sets for use in Japanese-English machine translation systems.

Subject

The Subject is traditionally defined as the general 'doer' of the action, such as *the dog* in *The dog gnawed the rope*. While this description is relatively uncontroversial for active clauses, it leads to two distinct treatments of passive subjects. The first is to take a Fillmore Case-style approach and identify that entity which corresponds to the active subject, in the "underlying" or "logical" subject sense. The alternative method is to concentrate solely on surface syntactic marking in identifying the **grammatical** subject. Thus, in *The rope was gnawed by the dog*, *the dog* comprises the logical subject (coinciding with the grammatical subject of the active voice equivalent), and *the rope* the grammatical subject.

In this research, we adopt this second, grammatical notion of subjecthood.

Subjects are necessarily obligatory or integral complements, and are characterised by nominative case marking. Japanese Subjects cannot be unambiguously detected through either word order, inflection, case marking, or the concept of surface syntactic obligatoriness. Rather, we must fall back on a number of linguistic tests to ascertain the Subject argument in a given sentence context.

The first such test is *zibun*-binding, and the observation that instances of the reflexive pronoun *zibun* can generally only bind to a clausal Subject position.[6]

(17) *Tarō-ga* *Hanako-ni* *Zirō-o* *zibun no ie-de* *syōkaisita.*
Taro-NOM Hanako-DAT Ziro-ACC self GEN house-LOC introduced
(lit.) "Taro$_i$ introduced Jiro to Hanako in self's$_i$ house." (Shibatani 1990:283)

Hence, in (17), *Taro* is the Subject.

One further test is subject honorification (Shibatani 1990:283), and involves the use of honorific *o* V-*ni naru* marking on the main verb to indicate deference to the Subject:

(18) a. *syatyō-ga* *waratta.*
president-NOM laughed
"The company president laughed"

b. *syatyō-ga* *o-warai-ni natta.*
(Subject honorific form of a.)

Naturally, the Subject entity must be animate and pragmatically worthy of honorification for this test to be applicable.

Despite this seemingly overbearing constraint, subject honorification provides an unambiguous means of determining the Subject case position through analysis of suitable situational participants, assuming that it is possible to uniquely identify one of those candidate participants as being worthy of honorification. That is, by identifying a case slot as containing a Subject filler in a given lexical context, for a given case frame, we can generalise that case slot as being the Subject in other lexical contexts, assuming consistency of case marking.

An application of this process is the identification of the Subject position of *wakar(-u)* "to know/understand" in a dative-nominative "ergative" marking context.[7] First, it is necessary to generate a sentence context involving animate participants in all candidate case slots, one of which must be unambiguously superior in social standing to the others. Such a sentence is given in (19a), where *syatyō* is the entity worthy of honorification. Next, we consider the "appropriateness" of subject honorification (cf. (19b)) and object honorification (cf. (19c)), and correlate these findings with our *a priori* honorification judgement. In (19), the appropriateness of (19b) suggests the datively marked *syatyō* as occupying the Subject position, leaving the nominatively marked *syain* as the Direct Object.

[6] See (Iida 1996) for documentation of significant exceptional cases in which *zibun* binds to non-subjects.

[7] Here again, note the non-coincidence between prototypical case-roles from the case marking types and the actual case-roles, with the Subject occupying a datively marked case slot and the Direct Object occupying a nominatively marked slot.

(19) a. *syatyō-ni* *syain no koto-ga* *yoku* *wakaru.*
president-DAT employee GEN -NML-NOM well understands
"The president understands well his employees."

b. *syatyō-ni* *syain no koto-ga* *yoku* *o-wakari-ni naru.*
(Subject honorific form of a.)

c. * *syatyō-ni* *syain no koto-ga* *yoku* *zonziru.*
(Object honorific form of a.)[8]

Direct Object

Direct Objects generally indicate the entity/entities affected by the action described by the main verb. As such, they are expressible only as obligatory complements, or in terms of our argument status hierarchy, as obligatory or integral complements.

One language-inspecific test for Direct Objects is that, in the absence of a Causee argument, they are commonly passivisable to the Subject position. This was the process observed above for *the rope* in *The rope was gnawed by the dog.*

Direct Objects are prototypically marked with accusative case.

Indirect Object

Indirect Objects represent the 'recipient' or 'beneficiary' in an action. Similarly to Direct Objects, Japanese Indirect Objects are passivisable (cf. (20), in which the Indirect Object is transformed into the Subject position), but only for *(di)transitive* verb senses (cf. (21)).

(20) a. *Tarō-ga* ***Hanako-ni*** *tegami-o* *okutta.*
Taro-NOM Hanako-DAT letter-ACC sent
"Taro sent Hanako a letter."

b. ***Hanako-ga*** *Tarō-ni* *tegami-o* *okurareta.*
Hanako-NOM Taro-DAT letter-ACC was sent
"Hanako was sent a letter by Taro."

(21) a. *tegami-ga* ***Hanako-ni*** *watatta*
letter-NOM Hanako-DAT reached
"The letter reached Hanako"

b. ***Hanako-ga*** (*tegami-niyotte*) *watarareta*
Hanako-NOM letter-by was reached
"Hanako was reached by the letter." (intended)

The default case marking for Japanese Indirect Objects is dative, and all Indirect Objects are optional complements.

Co-actors

In terms of traditional grammatical analysis, the Co-actor case-role straddles the boundary between the Direct and Indirect Object positions. It resembles the Direct Object case-role in argument status, in that all Co-actors are obligatory complements (but never integral complements). From the perspective of case marking, however, the Co-actor case slot is dative or comitative case marked, and hence most similar to Direct Objects. As an additional concern, Co-actors are not passivisable (cf. (22b)), in which respect they parallel intransitive Indirect Object usages.

[8] *Zonji(-ru)* is a lexical object honorific equivalent of *wakar(-u).*

(22) a. *Tarō-ga* *Hanako-ni/to* *atta.*
Taro-NOM Hanako-DAT/COM met
"Taro met Hanako."

b. **Hanako-ga* (*Tarō-niyotte*) *awareta.*
Hanako-NOM Taro by was met
"Hanako was met by Taro." (intended)

Under causativisation, on the under hand, Co-actors are transformed into Co-patients, (and hence coordinated with the Causee case-slot – see below):

(23) *Jirō-ga* *Tarō-o* *Hanako-to* *awa-se-ta.*
Jiro-NOM Taro-ACC Hanako-COM meet-CAUSE-PAST
"Jiro introduced Taro to Hanako." (lit. "Jiro caused Taro to meet Hanako.")

One phenomenon which sets Co-actors apart from **all** Direct and Indirect Object usages is that Co-actor case fillers can be coordinated within the Subject case slot to retain the same basic sentential semantic (ignoring focus/theme variation). That is, they occur with *reciprocal verbs* and are mutually exchangeable.

(24) *Hanako to Tarō-ga* *atta.*
Hanako and Taro-NOM met
"Hanako and Taro met."

This conflation of agency in the Co-actor case-role points to strong semantic resemblance between coordinating and comitative case marking roles of the particle '*to*', although no claim is made as to the exact nature of this correspondence.

The crucial difference between the two case-roles comes in the optional complement status of the Comitative, and it hence not being intrinsically defined for any verb. For example, while it is perfectly natural to say *Tarō-to(-issyo-ni) itta* "(I) went along with Taro", one would certainly not want to make the claim that the Comitative is defined within the valency frame for *ik(-u)* "to go". This informal observation leads to the Comitative and Co-actor case-roles existing in disjunctive distribution.

The optional nature of the Comitative case-role correlates to it being displaceable only with a dangling expanded comitative case marker in the relative clause body.

(25) a. *Jirō-ga* *Tarō-to(issyon-ni)* *pāti-ni* *itta*
Jiro-NOM Taro-COM party-DAT went
"Jiro went along to the party with Taro."

b. [*Jirō-ga* *issyo-ni* *pāti-ni* *itta*] *Tarō*
Jiro-NOM COM party-DAT went Taro
(lit.) "Taro, who Jiro went along to the party with."

c. * [*Jirō-ga* *pāti-ni* *itta*] *Tarō*
Jiro-NOM party-DAT went Taro
"Taro, who Jiro went along to the party with." (intended)

3.3.2 Other complement cases

All 'other complement cases' are optional complements, with the sole exception of the Passive Agent, which is an obligatory complement for adversative passives.

Passive agent

The Passive Agent case-role is derived through passivisation, and marked datively or with *ni-yotte*.[9] In the case of a "direct" passive (Miyagawa 1989b; Miyagawa 1989a), the Passive Agent is an optional complement, whereas Passive Agents in "adversative" passive contexts are obligatory complements (Hoshi 1994).

Causee

The Causee case-role is generated through causativisation, from Subject case slot transformation. It constitutes an obligatory complement (cf. Passive Agents) and is marked either datively or accusatively.

(26) a. ***Tarō-ga*** *Hanako-ni* *hon-o* *yonda.*
Taro-NOM Hanako-DAT book-ACC read
"Taro read a book to Hanako."

b. *Jirō-ga* ***Tarō-ni*** *Hanako-ni* *hon-o* *yomaseta.*
Jiro-NOM Taro-DAT Hanako-DAT book-ACC made read
"Jiro made Taro read a book to Hanako."

Co-patient

Co-patients mimic Co-actors in case marking (dative or comitative) and in their being non-passivisable (cf. (27b)), but differ in that they coordinate with the Direct Object or Causee case-slots (rather than the Subject – cf. (27c)), are optional complements and are unaffected by causativisation. That is, Direct Objects being unaffected by causativisation leads to consistency of case-role coordination.

(27) a. *Tarō-ga* *enzin-o* ***haikikan-to*** *kumiawaseta*
Taro-NOM engine-ACC exhaust pipe-COM joined
"Taro joined the engine with the exhaust pipe."

b. * ***haikikan-ga*** (*Tarō-niyotte*) *enzin-o* *kumiawasareta*
exhaust pipe-NOM Taro-by engine-ACC joined
"The exhaust pipe was joined to the engine (by Taro)." (intended)

c. *Tarō-ga* ***haikikan to enzin-o*** *kumiawaseta*
Taro-NOM exhaust pipe and engine-ACC joined
"Taro joined the engine and exhaust pipe."

Target

The Target case-role describes the optional target of experiential verbs (see Section 4.5.4) and is marked datively.

(28) a. *kabu no wariate-ga* *atta*
stock GEN allotment-NOM there was
"There was an allotment of stock."

b. *kozin-ni(taisite)* *kabu no wariate-ga* *atta*
individual-DAT stock GEN allotment-NOM there was
(lit.) "There was an allotment of stock to individuals."

[9]The dative case marker and *ni-yotte* are essentially interreplaceable, excepting that *ni-yotte* cannot be used with adversative passives (Kuroda 1979). There is also a slight semantic difference between the two marking types, relating to the degree of "affectivity"of the Subject of the passive clause.

Object Allative

The Object Allative case-slot refers to the optional physical medium reference associated with resultative action verbs such as *kak(-u)* "to write (on)" and *kake(-ru)* "to hang (on)"; it is marked datively.

(29) *Tarō-ga* *zibun no namae-o* ***kami-ni*** *kaita*
Taro-NOM self GEN name-ACC paper-DAT wrote
"Taro wrote his name on the paper."

3.3.3 Local cases

Japanese local case slots are consistently adjuncts. As suggested in the definition of adjuncts, this produces a relatively consistent spread of semantic variation across all patterns of usage, although predictable diversification is seen in case marker collocation, according to the directionality of the local case slot. The relative consistency of semantic scope of local case usages means that a single "locative filter" can be utilised to analyse the locative compatibility of a given noun head/case filler.

A default incompatibility with all local case types is assumed, unless otherwise marked; local case slots are generally unaffected by verbal modality, with the notable exception of the "resultative" *-te ar(-u)* auxiliary verb.

Locative

The Locative case slot construes the generic positional case-role, and is marked with the dative or locative (*de*) case markers.

(30) *Tarō-ga* ***Tōkyō-ni*** *sundeiru.*
Taro-NOM Tokyo-DAT is living
"Taro lives in Tokyo."

(31) *kaigi-ga* ***Meruborun-de*** *okonowareta.*
conference-NOM Melbourne-LOC was held
"The conference was held in Melbourne."

Locatives can occur in either an 'inner' or 'outer' reading, with Inner Locative being the default. Unless otherwise stated, the term Locative will be used to refer to Inner Locatives throughout this paper. For a discussion of these Locative types and patterns of distribution, refer to Section 4.5.4.

Ablative

The Ablative case slot indicates the local *source* of a directional action, and is marked with the particle of the same name.

(32) *Hanako-ga* *harubaru* ***Sapporo-kara*** *yattekita.*
Hanako-NOM all the way Sapporo-ABL came
"Hanako came all the way from Sapporo."

Allative

The Allative case slot indicates the local *target* of a directional action, and is marked datively or by the allative *e/made* case particles.

(33) *Tarō-ga* *sengetu* ***Kyōto-made/ni/e*** *borosya-de* *unten-site kita.*
Taro-NOM last month Kyoto-ALL-DAT-ALL old car-LOC drove there and back
"Taro drove to Kyoto and back in his bomb (of a car) last month."

Perlative

The Perlative case slot indicates the location through/across which the directional action of the main verb occurs, and is marked accusatively.

(34) *Tarō-wa* *izen* *tomodati no hune-de* ***Nihonkai-o*** *watatta-koto-ga aru.*
Taro-TOP previously friend GEN boat-LOC Japan Sea-ACC has crossed
"Taro has previously crossed the Japan Sea in a friend's boat."

3.3.4 Time cases

Japanese time cases are used for time-related reference. They are expressed as adjuncts and marked datively or with null marking. By default, verbs are assumed to be compatible with time case reference.

Durational

The Durational case-role indicates the length of time an action or state lasted. Unmarked durational nouns such as *kikan* "interval" are generally datively marked (with or without a pre-dative *zyū* durational marker), whereas durational complexes involving cardinal reference are most commonly associated with null marking.

(35) ***sēru no kikan-ni*** *mise-o* *odozureta*
sale GEN interval-DAT shop-ACC visited
" ( ) visited the shop during the sale."

Temporal

The Temporal case-role indicates a distinct point in time, and is marked with the dative case marker for cardinal date/time references, and with null/iterative (*mo*) marking for generic temporal expressions such as *kyō* "today" and *kyonen* "last year".

(36) ***kyō*** *Tarō-ga* *kuru.*
today Taro-NOM comes
(lit.) "Today, Taro will come."

(37) ***mikka-ni*** *Tarō-ga* *kuru.*
3rd-DAT Taro-NOM comes
(lit.) "On the third, Taro will come."

3.3.5 Oblique cases

Instrumental

The Instrumental case-role is used to describe a tool or instrument used in performing an action. The extension of tool types is highly verb-specific, supporting a middle status and default incompatibility with Instrumental reference. Instrumentals are marked with the locative (*de*) case marker.

(38) *Tarō-ga* ***hasami-de*** *tegami-o* *aketa*
Taro-NOM scissors-LOC letter-ACC opened
"Taro opened the letter with a pair of scissors."

Cardinal

The quantity/degree of an action is expressed with the Cardinal case-slot, with the semantic scope spanning from unit-based mention such as *sokudo* "speed", to physical extent and price, such as with

kakaku “price”. Cardinals are marked with the locative case marker or null marking, are adjuncts, and are by default incompatible??? with all verbs.

(39) *hikōki-ga* ***monosugoi hayasa-de*** *sora-o* *tonda*
plane-NOM extreme speed-LOC sky-ACC flew
“The plane flew across the sky at extreme speed.”

Chapter 4

Verb class-based resolution

4.1 Verb class hierarchy

The most fundamental mechanism called upon in realising this research is the derivation of a verb class hierarchy to cluster verbs based on valency, argument preferences, inter-case-role relations and argument type preferences. By way of linking verb classes to valency variation and inter-case-role relations, it is possible to slot optional arguments in and out of the valency frame as required, and apply inter-case-role preferences. The way this 'linking' is achieved in this research is that mini-rule sets are stipulated *a priori* for each verb class, and applied in parallel for the full verb class characterisation of the verb in question to produce multiple analysis types. Such rule sets are made up of IF-THEN-ELSE conditionals, frequently consisting of a single rule. Each rule set is designed to produce at most one output, with semantic incompatibility potentially leading to failure of any rule being triggered. The various outputs are then combined based on analytical coincidence, and weighted variously to produce a unique relative clause interpretation (see Chapter 6).

4.1.1 Argument status and interpretation preference

Commonly, the use of valency frames in a task of this type is closely linked to the introduction of sortal preferences for each case slot modelled, so as to be able to differentiate between distinct senses for a given verb stem (*verb sense disambiguation*), and at the same time disambiguate the case-role of each argument (*case-role disambiguation*). In the case of our system, however, the decision was made to alleviate local sortal preferences, so as to avoid consideration of verbal polysemy as much as possible. In return, argument status is used as an indicator of interpretive salience/accessibility, such that lower argument status interpretations will generally override higher value interpretations. For example, given an integral complement interpretation and an adjunct type interpretation, the integral complement will be preferred outright.

The top level argument preference set is thus:

> Full clause-based idiom ≫ Integral complement ≫ Shadow complement ≫ Middle ≫ Other argument types

That is, full-clause based idioms are strictly preferred over integral complements, which in turn take precedence over shadow complements, and so on. The placement of middles above obligatory and optional complements may seem controversial in light of the higher argument status of these types. However, the well-defined lexical nature of middles makes them more stable through relativisation than general complements.

The remainder of argument types (obligatory/optional complements, and adjuncts) interplay on a finer level, with selection between obligatory and optional complements being made from the mappings

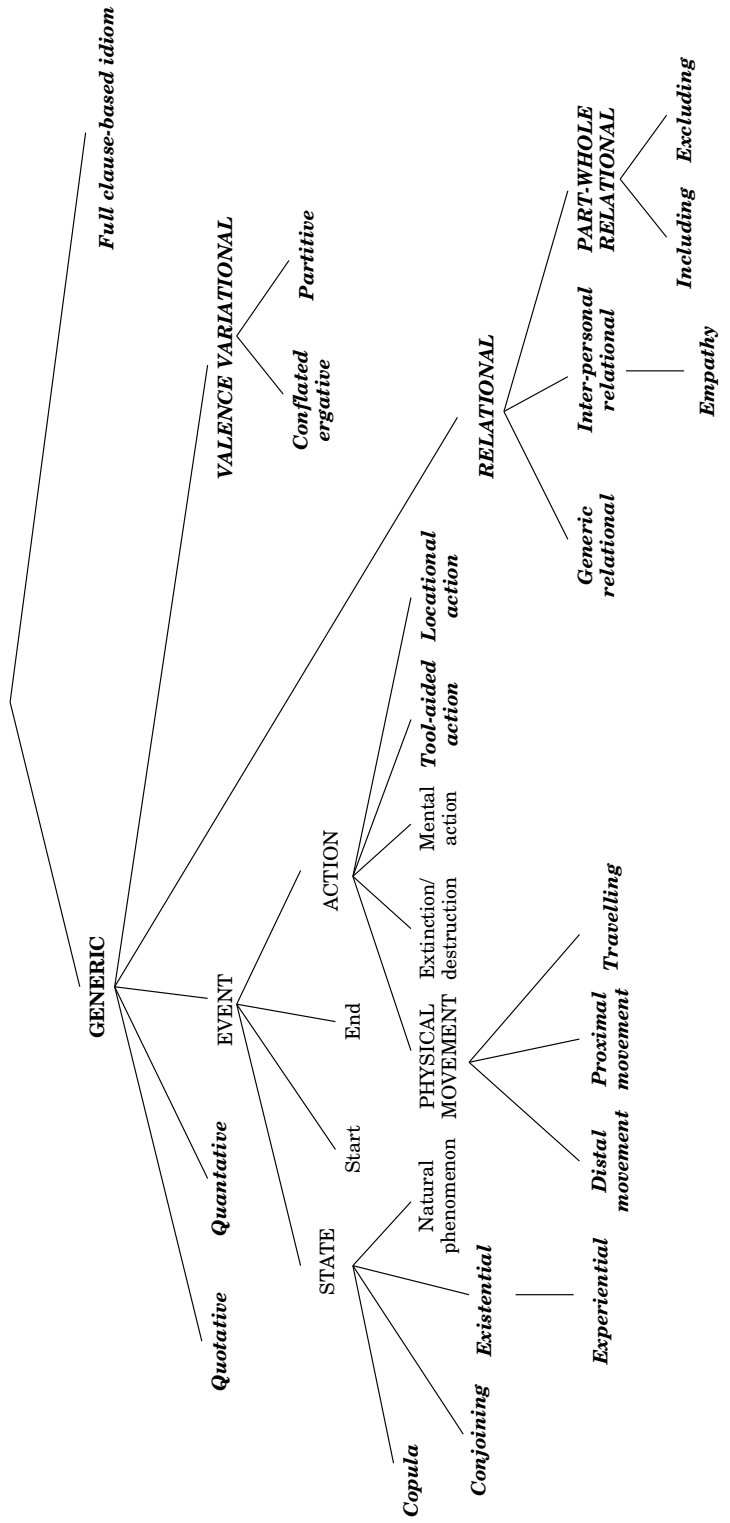

Figure 4.1: The full verb class hierarchy (Original verb classes indicated in bold, partitioning nodes capitalised)

between them inherent in the valency frame, and adjuncts weighted according to the prototypicality of the noun head with that adjunct type.

We claim that the proposed treatment of adjunct case slots is one of the potential strengths of the system, in terms of consistency of application and simplicity of processing. Given that adjuncts behave relatively consistently across all usages, it is possible to simply define adjunct type compatibilities for each adjunct case-role, and apply a uniform semantic treatment to the calculation of adjunct correspondence of arguments. In terms of verb class representation, this equates to associating verb classes to each adjunct type, determining a default (in)compatibility judgement (see Section 3.3), and marking those cases for which the particular verb sense does not coincide with this judgement.

4.1.2 Constructing the verb class hierarchy

In constructing the verb class hierarchy, the NTT "verbal semantic attribute" ('verb class' hereafter) hierarchy (Nakaiwa *et al.* 1994; Nakaiwa and Ikehara 1997) provided a solid starting point, which we were able to expand upon and modify for our purposes. Within the NTT transfer dictionary, verb classes are included as a general representation of the semantic type of each verb sense, in a cognitive or real world-applicability sense. They are called upon in the ALT J/E system during the discourse processing of zero pronouns (Nakaiwa and Ikehara 1994; Nakaiwa and Ikehara 1995; Nakaiwa *et al.* 1995; Nakaiwa and Ikehara 1996), and in English article and classifier generation. However, due to the principal intention of the verb class hierarchy to document clause-level semantics, rather than intra-clausal case-role relations, the verb class hierarchy does not in itself meet the lower level representational needs of our system, in replacing the need for case slot-based type constraints. Ultimately, therefore, existing verb class descriptions were combined in the single default valency frame entry generated for each verb stem (see Section A.1.1), and complemented with additional verb class types devised independently for our relative clause analysis purposes.

Verb class multiplicity

Multiplicity of verb class characterisation exists to a limited extent in the original NTT verb entries (Nakaiwa and Ikehara 1997:219), with the original average number of verb classes associated with each entry at 1.12. For the dictionary used in our system, the average number of verb classes per entry rose significantly to 1.45, with the distribution of verb class frequencies across the dictionary as given in Table 4.1.

This multiplicity of rule class categorisation complicates the relative clause analysis process somewhat, as we can often expect multiple verb class-based analyses for the relative clause, leading to scope for multiple analyses. Weighting and ranking schemes interfaced to the various rule sets are described in Chapter 6, but suffice to say that all analyses are combined additively, and the degree of semantic compatibility is weighted appropriately. The final output is then the analysis type which receives the greatest overall score.

Our verb class hierarchy

The final verb class hierarchy is as presented in Figure 4.1[1], in which labels indicated in bold are verb classes developed independently in this research, and NTT verb classes to which a particular rule set has been attached; capitalised labels are dummy nodes used to delineate/structure the verb class hierarchy but which have no verb class semantic in themselves. The hierarchy presented does not detail the full extent of the 107 verb classes contained within the full NTT transfer dictionary and maintained in the verb entry extraction process, but rather those classes/class clusters specifically

[1]Many thanks to Hiromi Nakaiwa of NTT for providing English translations of the full range of NTT verb classes.

No. verb classes for entry (Total: 11533)	**Frequency** (Total entries: 7951)	**Proportion**
1	5463	68.71%
2	1798	22.61%
3	457	5.75%
4	144	1.81%
5	52	0.65%
6	16	0.20%
7	11	0.14%
8	4	0.05%
9	2	0.03%
10	3	0.04%
14	1	0.01%

Table 4.1: Distribution of verb class frequencies

utilised in this research. For details of the "full clause-based idiom" class, the reader is referred to Section 2.3.3.

Individual descriptions of the newly developed verb classes, and associated rule sets where applicable, are provided below.

4.2 Physical movement

As the name suggests, physical movement verbs indicate a directional movement, as distinct from stasis. This can take the form of **distal** movement *towards* a particular location, **proximal** movement *away from* a particular location, or **travelling** motion *through* a locus.

4.2.1 Distal movement

Distal movement verbs are associated with compatibility with the Allative case-role, realised interchangeably through the dative and allative (*e*/*made*) case markers. Examples of distal movement verbs are *ik(-u)* "to go" and *muka(-u)* "to head towards".

```
IF (locative head AND uninstantiated Allative case slot) RETURN ALLATIVE;
```

4.2.2 Proximal movement

Proximal movement verbs are associated with compatibility with both the Ablative and Allative case-roles. Case marking for the Ablative case-role involves either the accusative or ablative (*kara*) case marker, and Allative case-role marking is identical to that for distal movement verbs, that is dative or allative. The default local case-role, however, is the Ablative, and Allative case-roles are generally only tenable with lexically marked **allative nouns** such as *saki* "destination". That is, the focus of proximal movement verbs is on the starting point of the action, and any unmarked locative is automatically associated with the Ablative case-role, but this preference can be overridden given overt allative-type marking.

Examples of proximal movement verbs are *ririku(-suru)* "to take off" and *tōzakar(-u)* "to recede/go away (from)".

Verb class	Frequency (Total entries: 7951)	Percentage of entries containing given verb class
Conflated ergative	79	0.99
Conjoining	25	0.31
Copula	3	0.04
Distal movement	124	1.56
Empathy	6	0.08
End	76	0.96
Excluding	1	0.01
Existential	126	1.58
Experiential	3	0.04
Extinction/destruction	77	0.97
Generic relational	289	3.63
Idiom	18	0.23
Including	3	0.04
Inter-personal relational	689	8.67
Locational action	1309	16.46
Mental action	2173	27.33
Natural phenomenon	101	1.27
Quantative	59	0.74
Partitive	49	0.62
Proximal movement	13	0.16
Quotative	160	2.01
Start	72	0.91
Tool-aided action	22	0.28
Travelling	66	0.83

Table 4.2: Verb class frequency

```
IF (allative noun head) RETURN ALLATIVE;

ELSE IF (locative head AND uninstantiated Ablative case slot) RETURN ABLATIVE;
```

4.2.3 Travelling

The focus for travelling verbs is on the Perlative case-role, and the route taken in the travelling motion; this is marked with the accusative. However, similarly to proximal travelling verbs, there is potential to refer to the destination through the use of ablative nouns, for which the case marking is allative or dative.

Tōr(-u) "to travel/pass through" and *tob(-u)* "to fly (across)" are both instances of travelling verbs.

```
IF (allative noun head) RETURN ALLATIVE;

ELSE IF (locative head AND uninstantiated Perlative case slot) RETURN
PERLATIVE;
```

4.3 Relational

Relational verbs are characterised by relating a source and target entity. In the case of inter-personal relational verbs, both entities are generally human, whereas generic relational verbs are associated with a broader range of both animate and abstract arguments. For all relational verbs, the focus is on the source entity, which in the context of relative clauses means that the default gapped case-role in cases of ambiguity between the source and target case slots, corresponds to the source. For target case-role gapping to occur, one or more of the following conditions must be met: (a) the source entity must be lexically realised, (b) the head must be an allative noun, or (c) there must be marked empathy on the source entity (see below). The most commonly occurring personal allative noun is *aite* "opponent", although the non-personal allative *saki* "direction/goal" can be equally acceptable for generic relational verbs in certain contexts.

Target entities can occur in any of the Co-actor, Co-patient and Indirect Object case-roles, with the particular case-role defined by the predicate and exclusivity of these case-roles occurring in a given valency frame.

Co-actor targets are obligatory in nature (a fact which derives directly from the definition of the Co-actor case-role), which makes them slightly more tenable to unmarked case-role gapping than the other two target argument types. They occur for "reciprocal" verbs such as *a(-u)* "to meet" (inter-personal relational) and *itti(-suru)* "to correspond" (generic relational). The 'reciprocity' of Co-actor target elements can be observed in (1), where ambiguity exists between Subject and Indirect Object case-role gapping.

(1) [*au*] *hito*
meets person
a. "people who meet ( ) "
b. "people ( ) meets"

Clearly, the two glosses correspond to the same situation, and if there is to be any constraint on the two case-roles, it is that the most topical/empathised entity occupies the Subject case slot for inter-personal relational verbs. We return to this matter in discussion of the system evaluation in Section 8.2.4.

Similarly to Co-actors, Co-patients are obligatory and occur for reciprocal-sense verbs. Here, however, the relational correspondence occurs with the Direct Object case slot, with differing degrees of reciprocity. In the case of *kumiawase(-ru)* "combine", for example, full interreplaceability is possible, whereas complications occur for 'replacement'-sense verbs such as *tyenzi(-suru)* "to change over" and *kōkan(-suru)* "to replace". Even with *kōkan(-suru)*, however, a higher degree of reciprocity is seen than for the English *replace*, in that the Direct Object source element can indicate the replacing item given appropriate replacing-type markedness on the case filler.[2] Thus, while recognising that implicit directionality is evident for certain Co-patient marked source types, the Direct Object and Co-patient case slots can generally be interchanged.

Indirect Object-type targets generally refer to the Recipient or Beneficiary of the described action, and are optional (again, obtained from the definition of Indirect Objects). Unlike Co-actors and Co-patients, Indirect Object targets produce a definite sense of directionality of the action, are not reciprocal (seen in the non-equivalence of (2a) and (2b)), and cannot be coordinated with the target case-role while retaining the same sense, as occurred above for Co-actors and Co-patients (cf. (2c)). Examples of Indirect Object targets occur with the verbs *watas(-u)* "to hand over" and *aisatu(-suru)* "to greet".

(2) a. ***Tarō-ga*** ***Hanako-ni*** *tegami-o* *okutta.*
Taro-NOM Hanako-DAT letter-ACC sent
"Taro sent Hanako a letter."

b. ***Hanako-ga*** ***Tarō-ni*** *tegami-o* *okutta.*
Hanako-NOM Taro-DAT letter-ACC sent
"Hanako sent Taro a letter." (≠ a.)

c. ***Hanako to Tarō-ga*** *tegami-o* *okutta.*
Hanako and Taro-NOM letter-ACC sent
"Hanako and Taro sent a letter (to ( ))." (≠ a.)

4.3.1 Inter-personal relational

Inter-personal relational verbs relate two animate entities, in the source case slot α (the Subject or Direct Object) and target case slot β (any of the Co-actor, Indirect Object and Co-patient case slots).

The only personal allative considered is *aite*. For details of the determination of the animacy of the noun head, see Section 6.3.

```
IF (animate head)
    IF (personal allative noun head AND uninstantiated target case slot β)
    RETURN β;
    ELSE IF (uninstantiated source case slot α) RETURN α;
    ELSE IF (uninstantiated target case slot β) RETURN β;
```

Empathy

'Empathy' verbs (Kuno 1978) form a proper subset of inter-personal relational verbs, and are defined by their incompatibility with a first person pronoun in the target case slot for simple inflection usages.

(3) * *Tarō-ga* *watasi-to* *atta.*
Taro-NOM I-COM met
"Taro met with me."

[2] Mechanical and civil engineering-related instruction manuals frequently contain sentences such as *atarasii boruto-o kōkansuru* (new bolt-ACCreplaces) "replace ( ) (with) the new bolt", in syntactically unmarked usages.

This high degree of empathic focus on the source entity produces the effect that target case-role gapping can occur without a surface realised source entity, for unmarked head nouns. Unfortunately, this often leads to lack of focus-based preferences between the source and target case slots in the case that both are uninstantiated, unless the head noun intension is marked allatively or empathically. At the same time, however, temporal or locative grounding tends to weight the focus towards the target case slot, as does the past tense.

(4) a. [*au*] *hito*
meets person
"the person who met ( ) " vs. "the person ( ) met"

b. [*nitiyōbi-ni* *atta*] *hito*
Sunday-DAT met person
"the person ( ) met on Sunday"

c. [*Nihon-de* *au*] *hito*
Japan-LOC meets person
"people (one) meets in Japan"

The handling of this marginal preference for the target case slot is a somewhat brutal one, in that simple existence of past tense inflection or local grounding is seen to generate unambiguous target gapping. However, given that there are no factors working to reverse the preference back in the other direction, the given treatment seems sufficient.

The algorithm for empathy verbs interfaces with that for inter-personal relational verbs through sequentiality, in that the following rule is applied prior to the inter-personal relational verb algorithm, and if an output is returned, that analysis type is automatically returned from the inter-personal relational verb algorithm.

```
IF (animate head AND uninstantiated target case slot α AND relative clause is
in past tense or is locally grounded) RETURN α;
```

4.3.2 Generic relational

Generic relational verbs are identical to inter-personal relational verbs, except that there is no semantic restriction on the source and target case slots. The rule set is thus basically the same as that for inter-personal relational verbs (see above).

```
IF (allative noun head AND uninstantiated target case slot β) RETURN β;

ELSE IF (uninstantiated source case slot α) RETURN α;

ELSE IF (uninstantiated target case slot β) RETURN β;
```

4.3.3 Part-whole relational

Part-whole relational verbs are subdivided into the including and excluding verb classes, and exemplify/exclude some part of the whole characterised by the modified noun head.

Including

Verbs contained in the including verb class can be used in a non-restrictive exemplification form, realisable in the simple non-past or past tense. The exemplar set is construed in the accusative case[3], which must be present to trigger the including sense, and no further arguments can collocate with the main verb.

(5) [*tōzai-o* *fukumeta*] *Ōsyū*
east and west-ACC included Europe
(lit.) "Europe, including (both) the east and the west"

Including relative clauses are head restrictive, and the system output on detection of this relative clause type is the tag for this modifying type, i.e. INCLUSIVE.

The fact that including relative clauses are not case-slot gapping can be seen by considering a simplex derivation of (6a) below.

(6) a. [*Beisutāzu-o* *fukumu*] *zen yakyū tīmu*
Baystars-ACC includes all baseball teams
"all baseball teams, including the Baystars"

b. *zen yakyū tīmu-ga* *Beisutāzu-o* *fukumu*
all baseball teams-NOM Baystars-ACC includes
"All baseball teams include the Baystars."

Clearly, the scope of the existential quantifier *zen* "all" is not equivalent between (6a) and (6b), excepting the case where the Subject in (6b) is treated as being textually containing through quotation, thus restricting the scope of the quantifier to the Subject NP. However, this quotative interpretation is not available in (6a) and hence does not constitute direct equivalence.

Members of the including verb class include *fukume(-ru)* and *hazime-to(-suru)*.

```
IF (simple main verb inflection AND unique accusatively marked argument) RETURN
INCLUSIVE;
```

Excluding

Excluding verbs extensionally restrict the modified head noun by identifying elements which are to be excluded from the default denotation. The exclusion sense of these verbs is produced for simple tense usages with only the accusative case slot instantiated.[4]

(7) [*nitiyō-o* *nozoku*] *mainiti*
Sunday-ACC excludes everyday
"everyday, excluding Sundays"

While usages such as (7) can be related back to the unmarked simplex sense *nozoku*, scope differences occur between excluding relative clauses and the corresponding simplex clause derivant, as was seen for the including verb class above:

(8) *mainiti-kara* *nitiyō-o* *nozoku*
everyday-ABL Sunday-ACC excludes
"to exclude Sunday from every day"

[3]Verb arguments can also be marked with the iterative (*mo*) marker.

[4]Note that for excluding verbs, the unique verb argument cannot be marked with the iterative (*mo*) marker, unlike including verbs.

This supports a head restrictive relative clause treatment for excluding relative clauses.

The excluding verb class consists uniquely of the verb *nozoku*.

```
IF (simple main verb inflection AND unique accusatively marked argument) RETURN
EXCLUSIVE;
```

4.4 Valence variational

Valence variational verbs are associated with distinct valency values, with a well-defined case slot mapping between the different valency frame types. In traditional valency frame dictionary treatments, such correspondences are not represented explicitly, and separate entries are simply allocated for the different valency types. The inadequacy of this method is described below for the various valence variational sub-types.

4.4.1 Conflated ergative

Conflated ergative verbs have been independently identified by Jacobsen (1992:212) as a noteworthy exceptional case in his proposed transitivity account of Japanese. As identified by Jacobsen, conflated ergative verbs are unambiguously of the Sino-Japanese type and display both transitive and intransitive usages, in ergative correspondence.

(9) a. *Tarō-ga* ***sagyō-o*** *kaisi-sita*
Taro-NOM work-ACC started
"Taro started work."
b. ***sagyō-ga*** *kaisi-sita*
work-NOM started
"Work started."

Considering this effect from the perspective of relative clauses, inherent ambiguity of valency frame type arises in cases where no overt Subject is given.

(10) a. [*Tarō-ga* *kaisyō-sita*] *mondai*
Taro-NOM solved problem
"the problem Taro solved"
b. [*kaisyō-sita*] *mondai*
solved problem
"the problem ( ) solved" vs. "the solved problem"
c. [*kaisyō-sita*] *hito*
solved person
"the person who solved ( ) "

In (10a) above, that the transitive sense is evoked is recoverable from the animate Subject, forcing *mondai* into the Direct Object position. For (10b), on the other hand, it is not possible to determine the transitivity of *kaisyō(-suru)* through sortal preferences, as they apply equally to the Subject case slot for the intransitive sense and Direct Object case slot for the transitive sense (in which case a zero subject interpretation would be produced). In fact, the only means of resolving this conflict is discourse processing, and the determination of a suitably salient discourse entity, capable of filling the transitive Subject position. Failure to locate such an entity would point to intransitive valency, and successful identification of an animate Subject would suggest transitive valency.

Given the current discourse-independent nature of our relative clause treatment, this type of high level processing cannot be called upon. The solution to the problem is thus to assume the unmarked

intransitive interpretation in the case that an overt Subject is not supplied within the relative clause. This correlates to hedging on the transitivity issue, as no assumption is made one way or the other as to the full content of the valency frame, and the resultant Subject analysis is equally applicable to both intransitive and transitive Subject analysis. Indeed, the only instance in which this analysis would prove incorrect is where gapping has occurred from the Direct Object case slot of the transitive sense of the verb in question, for a zero Subject.

```
IF (uninstantiated Subject case slot) RETURN SUBJECT;

ELSE RETURN DIRECT OBJECT;
```

4.4.2 Partitive

Partitive verbs contain a "part"/attribute in the nominatively marked Subject case slot, and can optionally collocate with a topic-marked "whole", to which the clausal attribution applies.

(11) a. *iro-ga* *aseta.*
colour-NOM faded
"The colour faded."
b. *sētā-wa* *iro-ga* *aseta.*
sweater-TOP colour-NOM faded
"The colour faded out of the sweater."

The whole and part can alternatively be coordinated in the Subject position, suggesting the clause initial topic construction as a "major subject" (Tateishi 1994), and hence a displaced 'whole' in a relative clause context as generating a bound gapping clause.

(12) *sētā no iro-ga* *aseta.*
sweater GEN colour-NOM faded
(lit.) "The colour of the sweater faded."

In terms of our case-role schema, the optional "whole" topic is classified as a (second, anchoring) Subject. For relative clause analysis, the gapping of the anchoring "whole" case slot translates to a bound gapping relative clause instance, with Subject gapping. Gapping from the "part" case slot in the absence of the "whole", on the other hand, constitutes simple Subject gapping, noting that gapping of the part in the presence of the whole is not possible.

(13) [*iro-ga/no* *aseta*] *sētā*
colour-NOM/GEN faded sweater
"a sweater which has faded in colour"

(14) [*aseta*] *iro*
faded colour
"a faded colour"

(15) * [*sētā-ga* *aseta*] *iro*
sweater-NOM faded colour
"the color which faded from the sweater" (intended)

The ungrammaticality of relative clause complexes of type (15), combined with the constraint that gapping of the part can occur only in the absence of the whole, means that (14) is not associated with the pragmatically acceptable second interpretation of *(lit.) the colour which faded from ( )*. However, analysis of (14) by means of the valency frames attributed to the simple 'part' and complex

'whole-part' senses would produce this second interpretation, of strictly equivalent acceptability to *a faded colour* due to the identical type restrictions on the 'part' case slot in the two frames. For this reason, the partitive verb class would appear semantically justified.

The current algorithm is limited in its potential to capture correspondences of this type, by the lack of a broad-coverage world knowledge source, with which to derive part-whole relationships. Thus, the actual handling of "whole" Bound Subject gapping of the type given above, is simplified to assume that Bound Subject gapping occurs only in the context of full complement case instantiation. What this means in real terms is that given full complement case instantiation, the system should prefer a Bound Subject interpretation over a head restrictive relative clause interpretation, these two analysis types being the only two possible alternatives. Clearly, therefore, considerable scope exists to improve the current treatment of partitive verbs, and this is left as an item for future research.

```
IF (fully instantiated complement case frame) RETURN BOUND SUBJECT;
```

4.5 Other verb classes

4.5.1 Copula

The copula verb class is used to tag the various realisations of the Japanese copula. The copula is relatively orthodox in behaviour within the context of relative clauses, in that case-role gapping can only occur from the Subject case slot, and incompatibilities exist with both temporal and local case-roles.[5]

The copula verb can take the forms *de-ar(-u)*, *de-gozar(-u)*, *de-irrasya(-ru)* and *da*.

```
IF (Subject case slot uninstantiated) RETURN SUBJECT;
```

4.5.2 Conjoining

Conjoining verbs closely resemble the copula from the standpoint of case-role gapping, by way of gapping only from the Subject position and being incompatible with both temporal and local case-roles. As implied by the nomenclature, conjoining verbs semantically 'conjoin' or 'relate' concept pairs, but differ from relational verbs in that gapping cannot occur from the target (non-subject comparator) case slot.

Examples of conjoining verbs are *uwamawar(-u)* "to exceed" and *kanren(-suru)* "to relate to"

```
IF (Subject case slot uninstantiated) RETURN SUBJECT;
```

4.5.3 Quantative

Quantative verbs are exempt from the default adjunct compatibility for time and cardinal adjuncts, with quantative arguments implicitly expressible through the valency frame-defined 'maximally peripheral complement case slot'. In real terms, the maximally peripheral complement case slot is the final (rightmost) complement represented within the valency frame.

[5]Due to the modular nature of the given verb class system in handling adjuncts, it is neither possible nor desirable to code multiple adjunct (in)compatibilities within a single verb class, and the observed adjunct incompatibilities for copula verbs are not applied directly from the copula verb class within our system.

(16) [*kakatta*] *zikan*
took time
"the time taken"

For (16), the noun head of *zikan* "time" would, by default, point to a Temporal or Durational case-role gapping analysis. The quantative membership of *kakar(-u)* overrides this analysis type, however. Instead, the nominatively marked Subject case slot is identified as the 'quantative case slot', and an attempt is made to map *zikan* onto the Subject case slot. Finding that the Subject case slot is uninstantiated, the system correctly returns a Subject analysis for the relative clause complex.

Examples of quantative verbs include *kakar(-u)* "to cost/take (time)" (maximally peripheral/quantative argument case slot = Subject) and *tuiyas(-u)* "to consume" (maximally peripheral/quantative argument case slot = Direct Object) .

```
IF (quantative noun head AND uninstantiated maximally peripheral complement
case slot α) RETURN α;
```

4.5.4 Existential

Existential verbs are stative verbs which can include mention of the locus??? of the state. As a direct consequence of the adjunct status of the locative case-role, multiple mention of locus can occur, with the separate locative case slots marked distinctly as being 'inner' and 'outer' positions (inner/outer terminology taken from Halliday (1970) and Platt (1971)). The (basic) Inner Locative is marked datively and is the default, whereas the (peripheral) Outer Locative is marked with the locative case marker (*de*), and occurs only in conjunction with the Inner Locative. This marks a point of departure from prototypical adjunct repetition, by way of repeatability not extending to the case marking level. Allocation to the two case slots is determined according to the relative specificity or local granularity of the locative case fillers, with the finer grained case filler occupying the inner case slot.

(17) *Tarō-ga* *Pari-de(ha)* *ōkina ikkenya-ni* *sundeiru*
Taro-NOM Paris-LOC(TOP) large house-DAT is living
"Taro lives in a big house in Paris."

In (17), for example, *ōkina ikkenya* corresponds to the Inner Locative case slot, and *Pari* to the outer case slot. The role of granularity in demarking these case slots is evident in that *ōkina ikkenya* is geographically contained within the extension of *Pari*, and the two locatives can be coordinated by use of the genitive connective (*no*) producing *Pari no ōkina ikkenya* "big house in Paris". Plugging this genitive coordinated locative back into the original clause, we see that the two locatives are conflated within the *Inner* Locative case slot:

(18) *Tarō-ga* *Pari no ōkina ikkenya-ni* *sundeiru*
Taro-NOM Paris GEN large house-DAT is living
"Taro lives in a big house in Paris."

Additionally, if we consider (17) in the absence of the mention of *ōkina ikkenya*, we see that *Pari* is forced from the Outer Locative case slot into the Inner:

(19) a. * *Tarō-ga* *Pari-de* *sundeiru*
Taro-NOM Paris-LOC is living
"Taro lives in Paris." (intended)

b. *Tarō-ga* *Pari-ni* *sundeiru*
Taro-NOM Paris-DAT is living
"Taro lives in Paris."

From this, it is clear that the outer locative case slot occurs only in simplex conjunction with an Inner Locative, and conversely that any singular locative case-role mention for existential verbs must occur in the Inner Locative case slot (and hence be datively marked for existential verbs).

Returning to consideration of relative clauses, this produces an immediate result for locative gapping existential verbs. That is, the head of a relative clause containing an Outer Locative and no Inner Locative *must* have been gapped from the Inner Locative case slot, given that the Inner/Outer Locative dichotomy is preserved under case-role gapping. Hence, given that an Outer Locative can only exist in the presence of an Inner Locative mate, 'dangling' Outer Locatives indicate cases of Inner Locative gapping.

(20) [*Tarō-ga* *Pari-de* *sundeiru*] *ie*
Taro-NOM Paris-LOC is living house
"the house Taro lives in in Paris"

Additionally, in the absence of any locative in the relative clause body, Locative gapping must occur from the Inner Locative case slot, noting that the Outer Locative case slot is equally tenable to case-role gapping as the Inner Locative.

(21) [*Tarō-ga* *ōkina ikkenya-ni* *sundeiru*] *Pari*
Taro-NOM big house-DAT is living Paris
(lit.) "Paris, where Taro lives in a big house"

Examples of existential verbs are *sum(-u)* "to live/inhabit" and *kizon(-suru)* "to exist".

```
IF (locative head)

    IF (Inner Locative case slot uninstantiated) RETURN INNER LOCATIVE;
    ELSE IF (Outer Locative case uninstantiated) RETURN OUTER LOCATIVE;
```

Experiential

Experiential verbs[6] form a proper subset of existential verbs, and are additionally compatible with an optional Target case slot, realised in the dative case.

Examples of experiential verbs are *i(-ru)* "to be/have" and *ar(-u)* "to be/have".

```
IF (uninstantiated Subject case slot) RETURN SUBJECT;

ELSE RETURN TARGET;
```

4.5.5 Locational action

Locational actions closely resemble experiential verbs, in that they are actions which can be attributed a locus or locii of action, although here, case marking is in the locative case (*de*) for both Inner and Outer Locatives. This leads to one major difference with existential verb-based relative clauses, in that a single Locative within the relative clause body is automatically interpreted as the Inner Locative. Thus, Inner Locative gapping only arises in the complete absence of any locative argument, and the Inner Locative gapping relative clause equivalent of a double Locative matrix clause produces ungrammaticality[7]:

[6] We refer to experiential verbs as 'stative' in (Baldwin *et al.* 1997b).

[7] Note that grammaticality is produced for the bound LOCATIVErelative clause equivalent, due to the preservation of the Inner and Outer Locative roles:

(22) a. ***kono ie de-wa*** *uti no inu-ga* *itumo* *tukue ya isu no sita-de*
this house-LOC-TOP us GEN dog-NOM always table and chair GEN under
netagaru
wants to lie
"(When) in this house, our dog is always wanting to lie under tables and chairs"

b. * [***kono ie-de*** *uti no inu-ga* *itumo* *netagaru*]
this house-LOC us GEN dog-NOM always wants to lie
tukue ya isu no sita
table and chair GEN under-LOC
(lit.) "under tables and chairs, where our dog is always wanting to lie when in this house"

Here again, however, the system is not intended to be able to make grammaticality judgements, and given an input such as (22b), *kono-ie* is simply assumed to fill the Inner Locative role.

```
IF (locative head)

    IF (Locative case slot uninstantiated) RETURN INNER LOCATIVE;

    ELSE RETURN OUTER LOCATIVE;

IF (uninstantiated Subject case slot) RETURN SUBJECT;
```

4.5.6 Tool-aided action

Tool-aided verbs are implicitly associated with a well-defined set of Instrumental which facilitate the action (a 'tool-set'). Instrumental denotations are verb-specific, and Instrumental mentions occur in the locative case.

One test for tool-aided actions is the recoverability of a tool sense for a generic object-referring noun head such as *mono* "thing" or *yatu* "thing", when modified by a relative clause containing simple non-past inflection on the main verb and no overt Subject.

(23) [*hasamu*] *yatu*
grabs thing
"a grabbing thing/implement" (lit. "a thing (with which) to grab")

Examples of tool-aided actions are *ut(-u)* "to strike/hit" (the associated tool-set includes such entries as *batto* "bat" and *kanaduti* "hammer") and *kak(-u)* "to write" (tool-set includes *pen* "pen" and *fude* "brush").

```
IF (head ∈ verb-defined tool-set) RETURN INSTRUMENTAL;

ELSE IF (generic object-referring head AND simple non-past main verb tense AND
relative clause subjectless) RETURN INSTRUMENTAL;
```

(i) ***kono-ie-de*** *uti-no-inu-ga* *itumo* *sita-de* *netagaru*
this house-LOC us GEN dog-NOM always under-LOC wants to lie
tukue-ya-isu
table and chair
(lit.) "tables and chairs, under which our dog is always wanting to lie when in this house"

4.6 Quotative

Quotative verbs are compatible with clause quotative (subordinating) usages, with the subordinated clause marked with the quotative (*to*) case marker. When nominalised, quotative verbs can generally express indirect and direct quotation through message linking with the *to-no* complex case marker or *to-i(-u)* relational verb.[8] For quotative verbs, case-role gapping can occur from within the subordinate clause ("subordinate clause gapping" - see below).

Examples of quotative verbs are *i(-u)* "to say", *omo(-u)* "to think" and *tutae(-ru)* "to report".

4.6.1 Subordinate clause gapping

For quotative verbs, case-role gapping can extend across a 'bridging' clause to a subordinated clause. Bridging clauses are defined as accommodating the subordinate gapping process without containing a gap themselves.[9] Bridging clauses must be headed by a quotative main verb, and rely on the subordinate clause being marked either with the 'quotative' case marker (*to*), in which case the clause takes finite inflection, or the nominative case marker for nominalised subordinate clauses (*koto* or *no* nominaliser).

(24) [[*100-ton izyō* *aru*] *to* *mi-rare-ru*] *zaiko*$_i$
100.tonnes over to be-PRES QUOT consider-PASS-PRES stock
'stock considered to be over 100 tonnes (in quantity)'

(25) [[*ziken-ni* *kanyosi-ta*] *to* *nihon-ga*
incident-DAT contribute-PAST QUOT Japan-NOM
miteiru] *kuni*$_i$
considers country
'countries which Japan believes to have contributed to the incident'

(26) [[*kaigi-ni* *sanka-suru*] *koto-ga* *kakunin-sareteiru*] *hito*
meeting-DAT attends -NML-NOM is confirmed person
"people who are confirmed to attend the meeting"

In order for subordinate gapping to occur for *quoted* subordinate clauses, the main verb in the superordinate relative clause must be potential or passive, or alternatively the superordinate relative clause must contain a surface representation of the clause Subject; in the case of a *nominalised* subordinate clause, the main verb in the superordinate clause must be passive.

If these inflectional/syntactic requirements are met, gapping resolution takes place at the subordinate clause level, based on the valency frame and inflectional content of the subordinate main verb. Interestingly, the same scope of gap types exists at the subordinate level as at the matrix relative clause level. We can thus apply the proposed matrix clause rules unchanged, excepting that subordinate gapping can only occur across a single bridging clause and hence recursion must be limited to a depth of one. If a case-role gap is detected within the subordinate clause, the system returns not only the identity of the gapped case slot, but also the fact that the gap is subordinate rather than superordinate.

For quoted subordinate clauses, a passive main verb produces coindexing between the superordinate and subordinate clause Subject positions, and an active main verb leads to the subordinate Subject being coreferent with the superordinate Direct Object.

[8] As mentioned in Section 2.1, the frequently occurring relational verb *to-i(-u)* is excluded from consideration in this research, and hence despite its compatibility with the quotative case marker, it is not classified as a quotative verb.

[9] Naturally, if the subordinate clause is to contain a case-role gap, an alternate gap cannot exist within the superordinate clause, according to the 'gap uniqueness' maxim.

(27) *(sono) kuni (no koto)-o* *Nihon-ga* [*ziken-ni* *kanyosi-ta*]
(that) country (GEN -NML)-ACC Japan-NOM incident-DAT contribute-PAST
to *mite-iru*
QUOT considers
"Japan considers (that) country to have contributed to the incident."

These sub/superordinate case slot correspondences are evoked in marking case slot incompatibility at a given clause level, in the case of case slot instantiation at the other level. For example, if the superordinate Direct Object case slot were instantiated in (27), the subordinate Subject case slot would be excluded from the case-slot gapping candidate set for gapping resolution at the subordinate clause level, and vice versa.

While recognising that these inter-clausal case slot correspondences superficially contradict our stipulation that gapping occurs from a unique case slot in a given interpretation, we consider the co-indexed case slots to have been conflated into one, and analyse the gap as existing in the subordinate clause. Indeed, the only consideration of the corresponding superordinate case slots comes in checking for zero content during gapping resolution, and conversely, for stipulating local gapping incompatibility in the case of instantiation of either of the case slots in question.

Full conflation of the superordinate and subordinate case slots can be seen through the unacceptability of a co-instantiation of the corresponding case slots in a matrix clause setting.

(28) * ***(sono) kuni (no koto)-o*** *Nihon-ga* [***(sono) kuni-ga*** *ziken-ni*
(that) country (GEN -NML)-ACC Japan-NOM (that) country-NOM incident-DAT
kanyosi-ta] *to* *mite-iru*
contribute-PAST QUOT considers
"Japan considers (that) country to have contributed to the incident." (intended)

That the gap exists at the subordinate clause level, rather than the superordinate clause level, can be verified by application of the subject honorification test on the subordinate and superordinate clauses.

(29) a. [[*kaigi-ni* *deta*] *to* *sareteiru*] *sensei*
meeting-DAT attended QUOT is said teacher
"teachers who are said to have attended the meeting"
b. [[*kaigi-ni* *o-de-ni natta*] *to* *sareteiru*] *sensei*
(Subordinate subject honorific form of a.)
c. * [[*kaigi-ni* *deta*] *to* *o-sare-ni natteiru*] *sensei*
(Superordinate subject honorific form of a.)

The grammaticality of (29b) over (29c) clearly supports the proposed subordinate case-role gapping treatment.

```
IF (passive or potential main verb OR superordinate Subject position
instantiated)

    Mark any subordinate gap incompatibilities based on superordinate case
    content;

    IF (gapping resolution of the subordinate clause identifies a gap α)
    RETURN SUB-α;

    ELSE RETURN CONTENT;

ELSE mark any superordinate gap incompatibilities based on subordinate case
content;
```

Chapter 5

Miscellaneous processing

5.1 Non-gapping expressions

Non-gapping expressions are defined as noun heads which are generally associated with a head restrictive interpretation of the containing relative clause complex. One example of a non-gapping expression is *mokuteki* "purpose" in (1) below.

(1) [*hataraku*] *mokuteki*
works purpose
"the purpose for working"

Within the current system, detection of a full non-gapping expression head[1] is taken to automatically produce a head restrictive clause sense. The fallacy of this strategy and potential for non-gapping expressions to be involved in gapping interpretations, is evidenced in (2).

(2) [*kare-ga* *nobeta*] *mokuteki*
he-NOM gave purpose
"the purpose he gave"

Despite this realisation of the inherent limitations of the proposed analysis method, non-gapping expressions provide a low-cost and remarkably effective means of filtering off head restrictive relative clauses, with the overall benefit derived through their use far outweighing the inherent noise they produce in analysis. Eventually, the system accuracy is hoped to be brought up to a level equivalent to the performance gain availed by non-gapping expressions, at which point they will lose their worth and a more thorough, context-dependent means of differentiating between gapping and head restrictive clauses will become necessary.

Examples of non-gapping expressions are:

mokuteki "purpose", *ugoki* "movement/trend", *hōsin* "direction/trend", *kanzi* "feeling", *zizitu* "fact/truth"

```
IF (full non-gapping expression head) RETURN CONTENT;
```

5.1.1 The extraction of non-gapping expressions

Ideally, non-gapping expressions could be mechanically extracted from an annotated corpus, by imposing a threshold likelihood of participation in head restrictive senses, and testing all heads against

[1] Note that non-gapping expressions must constitute the full noun head to be able to guarantee non-gapping.

this threshold value. However, given the limited size of the (annotated) corpus currently used, and its closed set nature, this method could not be applied. Instead, nouns were experientially evaluated for propensity to case-role gapping, and those for which case-role gapping was possible only in relatively restricted domains, were included in the non-gapping expression dictionary.

A further automatic learning procedure could be applied to learn verb collocations which produce case-role gapping sense for non-gapping expression heads. Here again, however, a richer source of annotated relative clause instances would be required than is currently available.

5.2 Time-related adjuncts

The following is a description of time-related adjunct types. Time-related adjuncts are by default compatible with all verbs and time-related interpretations override any other interpretations, in the case of ambiguity.

5.2.1 Temporal masking

Temporal masking describes the process of representing the temporal extension of a linguistic unit. A 'temporal vector' of discrete time units is employed to this end, containing a slot for each of the year, month, day, and time units:

Year	Month	Day	Time

Instances of each of these temporal units in the target linguistic unit are marked by 'switching on' the corresponding slot in the vector. For example, *1998-nen no 2-gatu* "1998 GEN February" would result in the vector:

1	1	0	0

Our interest in temporal vectors lies in their application to the analysis of Temporal gapping relative clauses, where an effect of concurrent case-role instantiation quite distinct to that for local adjuncts, is produced. Recall that for locatives, we introduced the notion of inner and outer case-roles, which were related through granularity/specificity, and interrelated at a high semantic level. For temporal case-roles, the reverse is true, in that while there appears to be an inherent limitation of two on the number of temporals which can include in a given context (including the noun head in the case of relative clauses), the only constraint on multiple Temporals is that their temporal vector representations are not permitted to overlap. In computer hardware terms, when the two temporal vectors are logically AND'ed together, the resultant temporal vector must consist of all zeroes. This leads to the following grammaticality judgements for relative clauses:

(3) [*1-gatu-ni* *kaigi-ga* *okonawareta*] *hi*
January-DAT meeting-NOM was held day
"days in January on which meetings were held"

* [*22-niti-ni* *kaigi-ga* *okonawareta*] *hi*
22nd-DAT meeting-NOM was held day
(lit.) "days on which meetings were held on the 22nd"

We return to consider temporal vectors in the next section.

5.2.2 Time relative constructions

Time relative constructions are produced either by the noun head being a **time relative expression**, or through the combination of the head being a **time relative complex** and particular main verb inflectional requirements being met.

Examples of time relative expressions are:

> *sono-hi* "that day", *yokuzitu* "the following day", *tōzitu* "that day"

These can collocate with any main verb inflectional type to generate a time relative interpretation.

Time relative complexes are generally produced by attaching a postfix to a phrase describing a time span. Instances of time relative complexes are:

> *1-kagetu-go* "one month later", *nan-nitika-mae* "a few days before", *2-okunen-mae* "200 million years before"

The two affixes which can collocate with time relative complexes are *-go* "after" and *-mae* "before". For *-go*, the stem verb must be in the simple past tense to produce a time relative construction, whereas *-mae* requires the simple present tense. If the head is a time relative complex but tense and aspectual requirements are not met, a Temporal case-role relative clause is produced.

The effect of the tense and aspect of the stem verb in variously producing a time relative construction and a non-relative temporal construction, is illustrated by:

(4) *kyōryū-ga* *sunde-ita* *yaku-2-okunen-mae*
dinosaurs-NOM were living about 200 million years ago
"about 200 million years ago, when dinosaurs lived"

(5) *kyōryū-ga* *sumu* *yaku-2-okunen-mae*
dinosaurs-NOM to live-PRES about 200 million years ago
"about 200 million years before dinosaurs lived"

Relative clause complex (4) constitutes an absolute temporal construction, and hence Temporal case-slot gapping, whereas the simple present tense in (5) leads to the production of a time relative construction.

The justification for the characterisation of time relative constructions as being non-gapping lies in the semantic incompatibility that exists between the time relative interpretation of the construction (*extra-clausal*), and the interpretation produced for Temporal case-role gapping (*intra-clausal*). This can be seen in the glosses of sentences (4) and (5) above.

Additionally, returning to use of temporal vectors, the observed masking effects and mutual exclusivity between temporal expressions co-existing in a single clause, are not observed for time relative constructions (cf. the ungrammaticality of (7)):

(6) [*1963-nen-ni* *Kenedī daitōryō-ga* *ansatu-sareta*] *yokutosi*
1963-DAT Kennedy president-NOM was assassinated the next year
(lit.) "the year after President Kennedy was assassinated in 1963"

(7) * [*1963-nen-ni* *Kenedī daitōryō-ga* *ansatu-sareta*] *tosi*
1963-DAT Kennedy president-NOM was assassinated year
(lit.) "the year President Kennedy was assassinated in 1963"

The temporal vector for both *1963-nen* and *yokutosi* in (6) becomes:

1	0	0	0

Hence, in applying one as a mask over the other, the inital flag remains set.

This provides a type test for time relative expressions, and allows us to draw a distinct line between time relative clauses and Temporal gapping clauses. This is worthy of particular note, as time relative clauses have been largely misrepresented as case-role gapping relative clauses, with gapping occurring from the temporal case-role (Matsumoto 1997:53).

5.2.3 Temporal expressions

Temporal expressions consist of time-related NP heads which are ground either absolutely or situationally. They consist of **absolute temporal expressions**, **generic temporal expressions** and **non-relative temporal constructions**.

Absolute temporal expressions are of the type:

16-niti "the 16th", *sakunen* "last year", *mainiti* "everyday"

That is, they constitute the set of temporal expressions which are well defined within the context of the surrounding text.

Generic temporal expressions are of the type:

kikan "period", *zikan* "time", *nendo* "year/fiscal year", *hi* "day"

These express generic temporal categories and are semantically restricted by the clause body. They can be likened to lambda expressions in that they are ground time-type case slot 'casts', without having the semantic extra-clausal and intra-clausal semantic incompatibility described below for time relative constructions.

Non-relative temporal constructions are temporal constructions which involve a time relative complex head, but which do not fulfill the stem verb inflectional requirements of a time relative construction (see above). Non-relative temporal constructions produce Temporal case-slot gapping relative clause sense.

Note that there is a certain degree of reliance on the surrounding context as to whether a temporal expression is absolute or generic, in that most absolute expressions can be forced to take a generic reading. This difference is most noteworthy when analysing restrictive and non-restrictive relative clauses, a matter which is beyond the scope of the current research.

5.2.4 Temporal vs. Durational interpretations

We have defined two distinct time-type case-roles in our case-role set, and obviously require some means to differentiate between them. For certain relative clause complexes, construal is uniquely defined, as occurs with non-relative temporal constructions being mapped onto the Temporal case-role. For absolute temporal expressions, the same applies, in that only the only time-type compatibility exists with the Temporal case-slot. However, for generic temporal expressions, the procedure becomes dependent on an array of factors, including the verb sense, verb inflection, instantiation of a temporal case slot within the relative clause body, and noun head type.

These are modelled heuristically as:

```
IF (non-relative temporal construction OR head suffixed durationally) RETURN
DURATIONAL;

ELSE IF (head is ''zikan'')

    IF ('start' or 'end' main verb) RETURN TEMPORAL;
    ELSE IF (potential inflection on main verb) RETURN DURATIONAL;
```

```
        ELSE IF (temporal case slot instantiated OR 'locational action'/'physical
        movement'/'existential' main verb) RETURN DURATIONAL;
        ELSE RETURN TEMPORAL;
    ELSE IF (generic temporal head)
        IF (temporal case slot instantiated) RETURN DURATIONAL;
        ELSE RETURN TEMPORAL;
```

These heuristics are evaluated in the overall evaluation of Section 8.2.

5.3 Cardinal adjuncts

Cardinal adjuncts are treated as being compatible with only the locational action, physical movement and mental action verb classes.

```
IF (cardinal noun head AND locational action/physical movement/mental action
main verb) RETURN CARDINAL;
```

5.4 The default rule set

In the event that none of the rules for the different verb classes is triggered to produce a non-Locative analysis type, a default rule set is applied.

The first stage of the rule set is to match the case slot contents of the relative clause against the valency frame, and ascertain remaining complement compatibilities. If the Subject case slot is found to be uninstantiated and the noun head to be animate, a Subject interpretation is returned. This treatment is founded on the "humanness hierarchy" proposed by Kuno and Kaburaki (1977)[2], and the realisation that humans and other animate arguments are more highly empathised than inanimate objects, and hence occupy the Subject case slot more easily.

Failing this, first person pronoun heads are analysed as producing a Bound Subject relative clause sense, based on the "Speech-act Participant Empathy Hierarchy" of Kuno and Kaburaki (1977).

If the Subject case slot is instantiated and the noun head is not in the first person, then we map the noun head onto the most accessible uninstantiated obligatory complement case slot. Accessibility is defined through argument status and the linear ordering of the case slots in the case frame.[3]

For this obligatory complement case slot mapping process to fail, all obligatory case slots in the valency frame must be instantiated, and the noun head cannot be compatible with any middle or adjunct case slot, as they would have been triggered prior to the defualt rule set. The only remaining interpretations for the relative clause, hence, are bound gapping and head restrictive. Abstract heads are assumed to be unavailable to bound gapping, and hence automatically tagged as head restrictive. For non-abstract heads, the system attempts to identify a complement case slot which is instantiated with a common noun (working down the accessibility hierarchy), and if successful, that case slot is returned as containing the bound case-role.

```
IF (uninstantiated Subject case slot)
```

[2]The same preference for animate Subjects has been implemented into Nishida *et al.*'s (1980) Case-based machine translation system.

[3]Case slots in the valency frame are ordered in descending order, from back to front, with optional complements excluded from analysis at this stage.

IF (animate head) RETURN SUBJECT;

ELSE IF (first person head) RETURN BOUND SUBJECT;

IF (uninstantiated obligatory complement case slot) THEN (identify most accessible obligatory case slot α) RETURN α;

ELSE IF (abstract head) RETURN CONTENT;

ELSE (determine highest accessible common noun-filled complement case slot β) RETURN BOUND β

Chapter 6

Lexical ambiguity

This chapter describes methods of resolving lexical ambiguity in the main verb and noun head, through statistical/representational preference and thesaurus use, respectively.

6.1 Verb lexical ambiguity

Plurality of successfully parsed entries results from a combination of both full and partial verb homophony and homography.

Full verb homophony is a direct result of the existence of multiple inter-replaceable writing systems within Japanese (hiragana, katakana and kanji), and occurs when two distinct verb entries coincide in both conjugational type and phonetic content of the verb stem/auxiliary verb complex. It is distinguishable from polysemy by virtue of the fact that disambiguation is achievable through use of the *kanji* form of the verb stem. An example of full verb homophony is "*a(-u)*", for which three heterogeneous kanji forms produce the distinct entries corresponding to the generic glosses of "to meet" (会う), "to coincide" (合う) and "to encounter" (逢う). Full homophony can alternatively be produced through combinations of auxiliary verb morphemes, such that "*miau*" is ambiguous between *mi-a(-u)* "to see--MUTUAL" and *mia(-u)* "to correspond".

Full verb homography is analogous to full verb homophony, except that the ambiguity exists in the kanji-based representation for coinciding conjugational types. In this case, disambiguation is possible through the kana phonetic version of the verb in question. An example of a full homograph occurs for the verbs *tome(-ru)* "to stop$_{TRANS}$" and *yame(-ru)*to quit/put an end to, for which a common kanji ("止") corresponds to the "*to-*" and "*ya-*" prefixes, respectively.

Partial verb homophony, meanwhile, occurs for verbs which differ in conjugational type, but agree in phonetic content of the verb stem. In this case, heteronomy of kana representation is produced for only certain inflectional types. In the case of our example of *a(-u)*, *ar(-u)* "to have" shares the verb stem of *a-*, and a heteronym is produced in the simple past tense, in the form of *atta*. Here again, however, kanji representation allows us to resolve the lexical ambiguity. Partial verb homography closely resembles partial verb homophony, except that the lexical ambiguity is produced in the kanji form, and resolvable through the use of kana. One example of partial homophony is produced for the simple past tense verbs *i-tta* "to go--PAST" and *okona-tta* "to carry out/hold--PAST", in that a single kanji ("行") is used to represent both "*i-*" and " *okona-*", respectively.

Note that in both of the classifications of partial heteronymic correspondence, the degree of coincidence is usually highly restricted, unlike full verb heteronymy. For the *i-tta*/*okona-tta* ambiguity, for example, partial heteronymy occurs only in the simple past tense or for progressive/perfective aspect.

6.2 Resolving verb lexical ambiguity

The most immediate method of resolving representational ambiguity is through statistical means. In this, two methods of statistical weighting were tested, the first based on naive probability and the second on representational preferences. For both methods, statistical scores were computed only in cases where multiple "generalised" (i.e. non-fixed expression) verb senses existed for a common verb stem. Fixed expression verb senses were automatically allocated a score of one, on the assumption that the fixed case element content is mutually exclusive for a given verb stem, and that the system should prefer idiomatic verb senses over generalised verb senses in the case that an idiomatic verb sense is plausible.

Due to the difficulty in predicting partial verb homophony/homography, all verbs sharing a common stem are treated as being fully heteronymic. Note however that for most inflectional types, coincidence of inflectional form does not result. In this case, the preferred verb sense is the one which has the highest relative score, ignoring the fact that the various scores in question may not total to one.

In terms of the interface between statistical weighting and the rule set, the rule set is applied as is for each parseable verb entry, and weights are summed for each resultant output. The unique system output is determined simply by calculating the highest summed weight, and randomly selecting between multiple analysis types of highest score.

6.2.1 Calculation of verb scores

The collation of frequencies is based on the EDR corpus (EDR 1995), and the verb sense annotations given for each verb occurrence. This is the same corpus as was used to extract all relative clause test sets described in this paper, and hence forms a closed test set. Whereas no direct reliance is made on verb sense by our system, the EDR corpus provides a means of determining lexical correspondences between different verb forms. To return to our example of *atta* above, all occurrences of *atta* are attributed a verb sense index, which correspond to different verb 'sense sets'. Contained in these sense sets are one or more representational alternatives of the verb stem *a-*, detectable through the verb stem representation contained in the original system dictionary entry. While there is not guarantee of disjunction between the alternative forms of *atta* and their respective sense sets, in almost all cases seen in the EDR corpus, full disambiguation was possible through the granularity of the verb sense index. In cases where sense ambiguity remained, the frequency of the original verb index was equally distributed between polysemous candidates.

One unfortunate characteristic of the EDR corpus is the uncommonly high numbers of index mismatches and 'nil' verb senses (unanalysed/unanalyseable verb senses). In the calculation of verb scores, index mismatches were simply disregarded from the data, while 'nil' indices were treated as described below for the separate scoring methods.

Frequencies are calculated *a priori* and normalised (significant to three figures) to produce the probability of occurrence of that form of the given stem verb.

Naive probability of occurrence

The naive probability of occurrence (NPO) of lexical form a of verb entry f (represented as a_f) is computed simply by totalling the number of usages of verb senses corresponding to a_f, and normalising over the total occurrences of a. Smoothing is achieved by evenly distributing 'nil' occurrences for a between entries a_i, where the total number of distinct entries a_i is represented as $|a|$ in equation (6.1). Thus, high levels of 'nil' occurrences will produce roughly standardised probabilities for all entries a_i, whereas lower levels of 'nil' occurrences will lead to nearer correspondence between relative frequency and normalised probability. This is intended to reflect the assumption that 'nil' senses suggest inherent ambiguity, and that higher levels of 'nil' values indicate lower confidence on the part

of the EDR developers in annotating usages of a.

$$NPO(a_f) = \frac{\frac{freq(a_{nil})}{|a|} + freq(a_f)}{\sum_i freq(a_i)} \tag{6.1}$$

Normalised representational preference

The representational preference (RP) of lexical form a of verb entry f (i.e. a_f) is defined as the confidence with which one can predict that a will be used to represent f, with the mean confidence predicted as 1. Smoothing is carried out through a double application of Jeffrey's estimate (Good 1965), that is by adding one to both the numerator and denominator. In this way, low-frequency verb entries and lexical forms can be smoothed to a value near the mean confidence of one (or to exactly one for zero-frequency entries), but at the same time high-frequency items are relatively unaffected. Additionally, instances of zero denominators are avoided, and the confidence is guaranteed to be strictly greater than zero.

Occurrences of the 'nil' index are not included in the RP calculation, such that entries found only with the 'nil' index return a representational preference of one.

$$RP(a_f) = \frac{1+freq(a_f)}{1+\sum_{i \neq a} freq(i_f)} \tag{6.2}$$

This is normalised over the representational preference for all source entries a_i, to produce the normalised representational preference $NRP(a_f)$.

$$NRP(a_f) = \frac{RP(a_f)}{\sum_i RP(a_i)} \tag{6.3}$$

6.2.2 Complexity of inflectional content

The only representational ambiguity not covered by these two scoring systems is instances where inflectional morphemes have produced an ambiguity which was not predictable from the stem verb (see the example of *miau* in section 6.1). This shortcoming is resolved by introducing the concept of 'complexity of inflectional content' (CIC), in which we penalise higher numbers of component inflectional morphemes. The penalty is computed *in situ* based on the number of inflectional morphemes contained in the verb, relative to the parse of simplest inflectional content (*min_infl*); the simplest parse receives a complexity of one. Thus, in the case of "*miau*", *mia-u* "to correspond-PRES" has a complexity of one, and *mi-a-u* "to see-MUTUAL-PRES" has a complexity of two. Weighting is achieved through the use of the constant parameter α. That is, the relative contribution of CIC can be enhanced by increasing α, hence exponentially increasing the value of the denominator and reducing the overall verb score (VS). At the same time, the parse of simplest inflectional content receives a complexity of one, and its VS is hence unaffected by variation in the value of α.

Complexity of inflectional content is compatible with both methods of statistical weighting given above, such that the VS for lexical form a of entry f (i.e. a_f) using statistical weighting measure SW is computed by:

$$VS(a_f) = \frac{SW(a_f)}{(CIC(a_f) - min\_infl + 1)^\alpha} \tag{6.4}$$

6.2.3 Evaluation of verb scoring

Preliminary evaluation was carried out to determine the relative effectiveness of the naive probability of occurrence (NPO) and normalised representational preference (NRP) methods, and contribution of CIC. The test sets used for this purpose were the full set of annotated relative clauses used in

	Overall accuracy (4411)	**Accuracy on case-role gapping clause instances** (3650)
Baseline	84.6%	90.8%
NPO ($\alpha = 1$)	86.0%	92.3%
NPO ($\alpha = 10$)	86.0%	92.3%
NRP ($\alpha = 0$)	85.9%	92.1%
NRP ($\alpha = 1$)	85.8%	92.2%
NRP ($\alpha = 10$)	85.9%	92.2%
Optimal	88.4%	94.5%

Table 6.1: Results for the verb scoring methods

developing the system, and the subset of gapping relative clauses. The sizes of the two test sets are indicated in brackets below each heading.

The baseline method for evaluation purposes simply selects the verb sense of highest probability when multiple parses are produced, which equates to utilising the naive probability method in computing the verb score, with α set to zero. The optimal achievable result for the system is determined by testing for membership of the correct analysis in the full set of analysis types produced for all successful parses. Given that verb scores simply rank these candidates, it is impossible for the other methods to better this non-deterministic method.

Table 6.1 lists the comparative results for the various methods[1], including evaluation of varying values of α for both the NPO and NRP methods. The 1.4% point difference between the overall accuracy for the baseline method and that for the NPO method with various values of α is a direct indication of the effects of weighting according to inflectional complexity, although the ineffectiveness of an increased α value is unexpected.

Likewise for the NRP method, whereas results are significantly higher than those for the baseline method, altering α produced only minor improvement. Indeed, performance with α set to zero (i.e. without consideration of CIC) marginally outperformed NRP with α set to one, although the statistical significance of this difference is questionable. This would tend to suggest that there is some interference in the choice of representational form of the verb stem given complex inflection, a fact which was borne out on summary inspection of the data. That is, the kanji form of the verb stem is generally utilised if auxiliary verbs are also given in a kanji representation, and full hiragana representation is generally reserved for simple inflection uses, such that a hiragana occurrence of "miau" would tend to point to the simple inflectional 'mia-u' stem (see section 6.1).

Perhaps more noticeable, however, is that the NPO method slightly outperforms NRP, which leads to the conclusion that representational preference in isolation is outweighed by the brute force of likelihood of sense.

Based on these results, we adopt the NPO method for the remainder of this paper, with α set to one.

6.3 Noun head lexical ambiguity

Noun head lexical ambiguity arises because of the polysemous representation utilised in the NTT thesaurus (Ikehara *et al.* 1993). That is, if a word displays polysemy, its various senses are positioned

[1]Data taken from (Baldwin 1998).

separately in the thesaurus, rather than attempting to maintain a one-to-one isomorphism between lexical form and thesaurus correspondences. This presents us with a dilemna, as we wish to not only classify lexical arguments according to type (i.e. as 'animate', 'locative', etc.), but also to weight the different senses so as to be able to chose between adjunct and complement senses, for example.

The method we use to weight nouns ($W(N)$) on class typicality, is simply to count the total number of occurrences of that noun N in the thesaurus, and the number of occurrences which fall into the particular sub-trees we have designated as classifying a particular type T, and calculate the ratio thereof.

$$W(N) = \frac{freq(N \in T)}{freq(N)} \tag{6.5}$$

For a noun N which never occurs in the extension of T, $W(N)$ thus becomes zero, whereas for an N fully enclosed within T, $W(N)$ is one.

In terms of the application of this weighting scheme to class membership, we stipulate a threshold for **animacy**, such that $W(N)$ must be greater than or equal to 0.5 for N to be judged as animate.

For the **locative** class, on the other hand, we skew the distribution of the produced $W(N)$ to produce preference for highly prototypical locatives, over animacy and other judgements, but penalise less clear-cut examples. The way we do this is to apply the function $Loc(N)$:

$$Loc(N) \quad = \quad \frac{[W(N)+1]^5}{21} \tag{6.6}$$

What this crude and computationally expensive function does is to inflate values closer to one, to a maximum of around 1.52, and penalise anything under a value of around 0.84 (actually, $\sqrt[5]{21} - 1$) by way of a relatively steep parabolic curve (values near zero actually increase slightly).

Chapter 7

Extensions to the basic algorithm

Despite the obvious attractions of the algorithm in the form presented to here, and its ability to weight interpretations, it still lacks in its ability to capture inter-clausal context, in what turns out to provide a surprisingly rich source of restrictions on the interpretation type. Here, we discuss the processing of cosubordinated clauses, coordinated clauses and coordinated heads.

7.1 Relative clause cosubordination

Clause cosubordination in Japanese is indicated by the use of a cosubordinating conjunction of the type *nagara*, *te*, *tutu* and *si*, or through *ren'yo* type inflection (aka. *continuative* (Kuno 1973b)).

(1) [[*pasukaru-ga* t_i *kōan-si,*] t_i *seisaku-si-ta*] *keisan-kikai*$_i$
Pascal-NOM DO design-REN DO make-PAST computing device
'a computing device designed and produced by Pascal'

(2) [[t_i *arubaito-o* *si-nagara*] t_i *gakkō-ni* *kayo-u*] *gakusei*$_i$
SBJ part-time work-ACC to do-WHILE SBJ school-DAT attend-PRES student
'students who work part-time while at school'

Within cosubordinating connectives, Kuno (1973b) observes that *si* and *ren'yo* must be subject coreferential, and Yoshimoto (1986) and Minami (1974) note that all cosubordinating connectives tend to coincide in Subject or Object content. These observations are borne out for (1) and (2).

In terms of relative clause analysis, we wish to suggest (3) as a corollary of the mutual exclusivity of the gapping paradigm:

(3) All component cosubordinated and subordinated clauses within a complex relative clause must agree in relative clause type.

That is, it is not possible to have a relative clause comprised of both gapping and non-gapping clause components. Additionally, we extend the observations of the above researchers to hypothesise that:

(4) For cosubordinated gapping relative clauses, the component clauses must agree in case-role gap type.[1]

In this, we wish to distance ourselves from peripheral adverbial usages of *nagara* and *tutu* (in which the *nagara*/*tutu* suffix is interchangeable with *nagaramo* in the contrastive sense and *toki* in the manner sense) and non-additive usages of *te* (Hasegawa 1996:6). Note that as was the case for subordinate gapping, the scope of gapping is unrestricted between complement and adjunct case slots, and includes, in this case, subordinate gaps.

[1] Note that this coincidence of gap does not apply to anchored clauses.

7.1.1 Processing of clause cosubordination

By way of accepting hypothesis (4) on gap type correspondence, we are able to extend our algorithm to consider case slot *incompatibilities*, in addition to the existing framework of case slot *compatibility* determination. Case slot incompatibilities stem from two sources: (i) directly from the content of the complement valency frame, and (ii) from case slot instantiation. Given a tool set of complement case-role types, it is possible to determine inherent case incompatibilities directly from the valency frame of the verb in question through a simple matching mechanism. This is combined with an analysis of those case slots instantiated in the input, and hence incompatible with that gap through the 'one-case-per-clause' constraint (Fillmore 1968:22).[2] Given that we can expect multiplicity of analysis type due to multiple parses, we take the intersection of gap incompatibilities for each analysis type, and return the resultant set of incompatibilities for the highest scoring analysis type. On the inter-clausal level, the union is taken of the individual incompatibility set for each component clause, in determining the overall incompatibility set.

Determination of the **unique** overall analysis for the relative clause is facilitated through the same process as at the single clause level, in that the weighted outputs for each member clause are summed, and a final sorted list of analysis types determined. However, this is now combined with the incompatibility set to weed out incompatible case-role types, and the highest scoring compatible clause analysis is outputted. In the case that all analysis types are judged to be case incompatible, the overall clause is assumed to be non-gapping.

7.1.2 Gap correspondences

Cosubordination of canonical gapping and subordinated gapping clauses leads to an interesting effect, in that inter-clausal agreement occurs in terms of the gap type, but not as to the clause level from which gapping has occurred (see (5)). It is for this reason that our hypothesis stipulates agreement in **case-role type sub-type**, but makes no mention of **clause level**.

(5) [[[t_i *i-na-i*] *to* *mi-rare*] t_i
SBJ to be-NEG-PRES QUOT consider-PASS-REN SBJ
renraku-sare-na-katta] *hito*$_i$
to contact-PASS-NEG-PAST person
'a person who was assumed not to be in and (hence) not contacted'

(6) [[*kankeisya-o* *nozo-ki*] t_i *pāti-ni* *syussekisi-ta*] *ninzū*$_i$
organiser-ACC exclude-REN SBJ party-DAT attend-PAST number of people
'the number of people who attended the party, excluding organisers'

(7) [*yoru* [*tōkyōwan-o* *watari-nagara*] t_i mi-ru]
night Tokyo Bay-ACC to cross-WHILE DO to see-PRES
reinbōburizzi$_i$
Rainbow Bridge
'Rainbow Bridge as seen at night while crossing Tokyo Bay'

(8) [[*Tarō-ga* *sekininsya-to* *nari,*] *1-gatu-ni*
Taro-NOM person in charge-COM to become-PRES January-DAT
okona-ware-ru] *konkūru*
to hold-PASS-PRES competition
'a competition to be held in January, which Taro is in charge of organising'

[2]Note that application of the one-case-per-clause constraint is restricted to **complement** case slots.

Two verb types which do not contribute to the clause sub-type, and are hence disregarded during the resolution process, are the excluding and including types. Excluding and including clauses are adverbial constructions, and hence exempted from consideration with hypothesis (4). Considering (6), in which the first clause is of the excluding type, the main clause is essentially treated as a simplex clause, and the Subject gapping sub-type can be recovered.

One fact which is clear from the original description of conjunction types is that peripheral subordinating usages exist for all conjunctions except the ren'yo form, suggesting difficulty in correctly predicting the type of clause dependency in a given clause prior to being able to apply the restrictions proposed in section 7.1.1. While this is certainly the case for *te* clauses, complement analysis-based heuristics were found to be productive in correctly analysing *nagara* and *tutu* clauses. These heuristics consist of analysing the complement content of the cosubordinated clause to determine if all non-Subject complement case slots are instantiated. The exceptional treatment of the Subject case is founded in the observation that these are 'small clauses' (Radford 1981), the Subject of these subordinate clause types is inherently coindexed to that of the superordinate clause, through a PRO mechanism, and overt Subject mention within the *nagara* clause is not possible.

If full instantiation is detected, the unit clause in question is therefore discounted from the resolution process, on the grounds of being adverbial. This process can be seen to correctly identify the subordinated *nagara* clause in (7), with the Direct Object gap existing only in the main clause and no Direct Object incompatibility restriction imposed by the *nagara* clause.

One additional qualification which must be made to (4) is that it does not seem to apply to the bounded case-role for bounded relative clauses. To take an example, the Direct Object case-slot is bound in the first clause of (8), but the final clause is a clear instance of Subject case-slot gapping. At the present time, we have no explanation for this effect, and simply disregard bounded relative clauses during cosubordination-based resolution. Note, however, that this does not threaten the applicability of (4), as bounded relative clauses are included under the classification of case-role gapping relative clauses.

7.1.3 The treatment of subordinate clauses

A preliminary study of the relative clause corpus produced for the system suggested that around 6–7% of all relative clauses involve clause cosubordination, pointing to the significance of the above method of analysing cosubordinate clause complexes. Relative clauses containing adverbial clauses, however, seem to account for a much higher proportion, at around 20% of all relative clauses. While the clause sub-type hypothesis proposed above does not apply to adverbial relative clauses, the more general suggestion of coincidence of relative clause type (gapping vs. head restrictive) is suggested to apply to all relative clauses. As is evident in all levels of evaluation, the accuracy of the system for head restrictive clauses is significantly lower than that for gapping clauses, and the application of this basic restriction presents itself as a possible tool in enhancing resolution of the clause type.

One promising starting point for an extension of this type comes from the work of (Okumura and Tamura 1996:874), who suggest that 'subject switching' occurs between adverbial and containing clauses, given a surface Subject in either clause. They then go on to comment that variations in gap type are largely context dependent and not predictable simply from local constraints. The application of their proposed heuristic, and further analysis of the gap switching mechanism, however, remain as outstanding issues in the system development.

7.2 Coordinated relative clauses

Coordinated relative clauses are essentially multiple instances of finite-inflected relative clause bodies, which share a common head. They do not adhere to the stronger-version hypothesis (4), as seen below, but we hypothesise that the weaker hypothesis of (3) is maintained.

(9) [*ano yūmeikyoku-o* *sakkyoku-sita*,] [*daremo-ga*
that famous piece-ACC composed everyone-NOM
yoku *sitte-iru*] *Bētōben*
well knows Beethoven
"Beethoven, who you all know well, and wrote *that* famous piece of music"

The potential for coordinated case-role gapping relative clauses to be associated with discrete case-role gaps, is illustrated with (9), for which the first clause in Subject case-role gapping, and the second Direct Object case role gapping. Due to this possibility for non-case-role correspondence, it is necessary to annotate and process coordinated relative clauses independently. Currently, however, relative clauses are attributed a unique overall interpretation, making annotation of case-role switching coordinated relative clauses impossible.

7.3 Noun head coordination

A more approachable problem is the handling of noun head coordination. Unlike relative clause coordination, noun head coordination occurs *within* the noun head, alleviating scope for parallel interpretations for separate noun units. This means that we can safely process each noun head in parallel and *combine* the resultant analyses to determine the analysis type of overall highest plausibility. In this, combination of interpretational scores is performed additively.

Chapter 8

Evaluation

8.1 Evaluation criteria

Evaluation of the system was carried out based on a test set of relative clause complexes extracted from the Japanese EDR corpus (EDR 1995). The test set was classified according to the verb class content of the main verb, to verify the accuracy of each verb class type. Additionally, fixed expressions were identified to compare the overall system performance based on generalised valency frames (Table 8.1) , and that for fixed expressions (Table 8.2).

The system accuracy was analysed according to the overall accuracy for each test set, and also the accuracy on only case-role gapping examples[1]. For each class of rule set outputs given below, the actual number of instances of that class is given in "Total", followed by the overall number of correctly and incorrectly analysed relative clauses. The data is then broken down into the case-role gapping and head restrictive types, following which an analysis is given for each clause sub-classification receiving an output for that verb class. The figures are additionally each analysed in terms of precision, recall and F-measure (β set to 1 throughout evaluation).

$$\text{Precision} = \frac{\textit{\# of instances of that type correctly identified by the algorithm}}{\textit{Total \# of instances categorised as that type by the algorithm}}$$

$$\text{Recall} = \frac{\textit{\# of instances of that type correctly identified by the algorithm}}{\textit{Total \# of instances of that type}}$$

$$\text{F-measure} = \frac{(\beta^2 + 1) \times \text{Recall} \times \text{Precision}}{\beta^2 \times \text{Recall} + \text{Precision}}$$

Cases where a zero denominator has made any of these values incalculable are indicated in the results as '*N.C.*' (Non-Calculable).

Evaluation of total performance for the given data set is calculated only in terms of precision (accuracy), but this figure is identical to that for the system recall on the full input set.

8.1.1 Baseline evaluation

As in any empirical evaluation, it is vital that we establish a baseline figure for the system performance. As summary analysis of the case-role gapping correspondences for the overall data set show, the SUBJECTcase-role alone accounts for 3004/4615 (= 65.09%) of all relative clause instances. Thus, by establishing an algorithm which automotically outputs a SUBJECTanalysis for any input, irrespective

[1]In terms of the tags returned from the algorithm, all instances of EXCLUSIVE, INCLUSIVE, CONTENTand TIME RELATIONALwere excluded from the data.

of head type or verb class membership, we are able to attain an overall accuracy of 65.09%. This forms the true baseline performance for our system (B_1).

An alternative baseline performance figure can be obatined from the implementation of the algorithm proposed in (Baldwin *et al.* 1997a), which constitutes a much simplified version of our final system, but relies on the same basic concepts and methods for non-gapping expressions, Temporal case gapping, time relative constructions, and and case slot instantiation. First, non-gapping expression-headed relative clauses are filtered off as generating head restrictive relative clauses. Next, the system accesses a transitivity judgements for the main verb of the input realtive clause, based on which the system attempts to map the head onto the Direct Object case slot (assumed accusative case marking) for transitive verbs, and the Subject case slot (assumed nominative case marking) failing this. As a default, all relative clauses are assumed to be head restrictive.

This algorithm (B_2) produces significant improvement over B_1 above, with an overall accuracy of 75.1%, and is detailed along with B_1 in Table 8.1.

8.2 Overall evaluation

	# instances	# correct	# incorrect	Precision	Recall	F-measure
Total	4615	4089	526	88.60	–	–
Baseline (B_1)	4615	3004	1725	65.09	–	–
Baseline (B_2)	4615	3466	1149	75.10	–	–
Case-role gapping	3913	3659	481	88.38	93.51	90.87
Head restrictive	601	330	45	88.00	54.91	67.62
Bound Subject	51	33	32	50.77	64.71	56.90
Subject	3004	2890	262	91.69	96.21	93.89
Direct Object	306	270	93	74.38	88.24	80.72
Indirect Object	15	13	4	76.47	86.67	81.25
Co-actor	62	52	8	86.67	83.87	85.25
Co-patient	1	1	0	100.00	100.00	100.00
Medium	18	9	2	81.82	50.00	62.07
Target	7	5	1	83.33	71.43	76.92
Passive Agent	1	0	0	*N.C.*	0.00	*N.C.*
Causee	0	0	1	0.00	*N.C.*	*N.C.*
Instrumental	1	1	1	50.00	100.00	66.67
Locative	99	70	41	63.06	70.71	66.67
Cardinal	16	8	8	50.00	50.00	50.00
Allative	23	17	5	77.27	73.91	75.56
Perlative	4	3	1	75.00	75.00	75.00
Temporal	97	93	7	93.00	95.88	94.42
Durational	49	46	8	85.19	93.88	89.32
Exclusive	114	113	2	98.26	99.12	98.69
Inclusive	28	28	4	87.50	100.00	93.33
Time Relational	19	19	0	100.00	100.00	100.00
Idioms	82	81	0	100.00	98.78	99.39
Subordinate gapping	17	17	1	94.44	100.00	97.14

Table 8.1: Overall analysis

The overall performance of the system on the corpus of 4615 annotated relative clauses is detailed in Table 8.1. The overall system accuracy calculates to around 89%, as compared to 65% for the true baseline Subject case-slot analysis method (B_1) and 75% for the naive transitivity algorithm (B_2).

Within the figure of 89%, the contribution from the case-role gapping and head restrictive relative clause groups is approximately equivalent.

The first major result is the disparity between the recall for these two relative clause types, with case-role gapping clauses far outperforming head restrictive clauses on 95%, as compared to 55%. This points to there being an over-bias towards case-role gapping gapping clause analysis, and overgeneration occurring for this type, an unsurprising result given the core focus of this research on case-role gapping clauses. Within the case-role gapping relative clause type, however, the figures for bounded relative clauses are slightly disappointing, and again there are signs of overgeneration occurring.

An interesting correspondence between case-slot accessibility/immediacy and accuracy, for the complement case-role set, with performance gradually degrading from Subject to Direct Object, Co-actor and Indirect Object. It must be said, however, that this trend is probably not entirely coincidental, due to the focus placed on the more accessible case-roles during the verb class production phase.

The worst figures are seen for the Local case-role set, a sign of the frequent ambiguity between the locality and animacy/autonomy senses, as occurs for country name references. Additionally, the context-independence of locative detection leads to the system occassionally missing the local sense altogether. Having said this, it is reassuring to note that the lowest F-measure values are at least comparable to the accuracy for the true baseline of B_1, with the Allative and Perlative gapping analyses roughly equivalent in degree to B_2.

The treatment of the time case-slots and time relatives was, if anything, better than expected, and the main source of noise between the Temporal and Durational case-slots was mistaken mapping between the two. That is, the system is generally able to ascertain the time-relatedness of time case-gapping, but has slight difficulties in differentiating between the two sub-types.

Similarly, the system performed remarkably well for the well-defined head restrictive sub-classifications, with only the Inclusive sense falling below a 95% F-measure value. One conclusion which could be drawn from this is that the successful handling of other well-defined head restricting phenomena could well be the most efficient method of further improving system performance, and eliminating overgeneration of case-slot gapping interpretations.

Finally, full clause-based idiom detection was predictably excellent, as was the identification of subordinate gapping instances.

8.2.1 Evaluation of fixed expressions

Fixed expressions were singled out for separate evaluation due to their high presence in the system dictionary. Due to the slightly lower average number of generalised case slots within a fixed valency frame, as compared to a (fully) generalised valency frame, one would expect them to perform significantly better than their generalised counterparts. The figures in Table 8.2 do not support this supposition, however, principally because of the surprisingly low precision and recall for head restrictive relatice clauses.

As a means of comparison, we decided to compare the performance of fixed expressions-based analysis for fixed expressions, with a generalised valency analysis, to test whether the relatively low accuracy represents some inherent complexity in the data set, or simply an erroneous analysis procedure. The results for this generalised analysis are given in the bracketed “Generalised” row. Happily, there is a significant performance gap between the teo analysis methods, with fixed expression-based processing gaining a 4 percentage point advantage over generalised analysis.

One additional statistic extracted for the fixed expression data set was the number of occurrences of head displacement, and the accuracy of fixed argument matching. As indicated in the “Head displacement” row, the system was able to detect fixed argument displacement with 100% precision and recall. Given that the difference between the generalised and fixed expression analyses was within the order of these 16 occurrences of head gapping, we then verified the accuracy of generalised analysis on these 16 instances. This revealed that the generalised analysis method was able to correctly reproduce

	# instances	# correct	# incorrect	Precision	Recall	F-measure
Total	212	183	29	86.32	–	–
(Generalised)	212	(174)	(38)	(82.08)	–	–
Head displacement	16	16	0	100.00	100.00	100.00
Case-role gapping	182	173	26	86.93	95.05	90.81
Head restrictive	29	9	3	75.00	31.03	43.90
Bound Subject	2	0	1	0.00	0.00	*N.C.*
Subject	136	133	20	86.93	97.79	92.04
Direct Object	29	28	3	90.32	96.55	93.33
Indirect Object	0	0	1	0.00	*N.C.*	*N.C.*
Medium	3	3	0	100.00	100.00	100.00
Locative	1	1	0	100.00	100.00	100.00
Cardinal	2	0	0	*N.C.*	0.00	*N.C.*
Allative	4	4	0	100.00	100.00	100.00
Temporal	1	1	1	50.00	100.00	66.67
Durational	1	0	0	*N.C.*	0.00	*N.C.*
Inclusive	1	1	0	100.00	100.00	100.00
Time Relational	1	1	0	100.00	100.00	100.00

Table 8.2: Analysis of fixed expressions

	# instances	# correct	# incorrect	Precision	Recall	F-measure
Total	50	45	5	90.00	–	–
		(26)	(24)	(52.00)	–	–
Case-role gapping	44	44	5	89.80	100.00	94.62
		(25)	(24)	(51.02)	(56.82)	(53.76)
Head restricting	6	1	0	100.00	16.67	28.57
Subject	37	37	2	94.87	100.00	97.37
Direct Object	4	4	2	66.67	100.00	80.00
Locative	1	1	1	50.00	100.00	66.67
Temporal	2	2	0	100.00	100.00	100.00
Subordinate gapping	15	15	1	93.75	100.00	96.77

Table 8.3: Results of subordinate gapping analysis

14 of the 16 displacement instances, and that the difference between the overall performance for the two methods was well beyond the scope of this localised phenomenon.

8.2.2 Evaluation of subordinate gapping

Basic evaluation of the subordinate gapping detection algorithm was carried out on a set of 50 relative clauses containing a quotative main verb. The system was then further tested on the component subset of case-role gapping relative clauses, with successful gap detection requiring the additional correct identification of the level of embedding of the gap. This derivative test set of 45 gapping relative clauses included 15 *subordinate* gapping clauses. The same data sets were analysed with a simplex-type relative clause analyser which did not have access to the subordinate-superordinate case slot correspondences nor the procedure to detect subordinate case-slot gapping. Results are given in Table 8.3, with the bracketed figures under the "Total" and "Case-role gapping" rows indicating the results for the simplex analyser on the respective data sets.

It is perhaps unrealistic to directly compare the results of the two analysis types for gapping clauses,

	# instances	# correct	# incorrect	Precision	Recall	F-measure
Total	184	142	42	77.17	–	–
		(123)	(61)	(66.85)	–	–
Case-role gapping	105	95	41	69.85	90.48	78.84
		(87)	(60)	(59.18)	(82.86)	(69.05)
Head restricting	79	47	1	97.92	59.49	74.02
BOUND SUBJECT	2	1	1	50.00	50.00	50.00
SUBJECT	83	82	34	70.69	98.80	82.41
DIRECT OBJECT	8	3	3	50.00	37.50	42.86
MEDIUM	1	0	0	*N.C.*	0.00	*N.C.*
LOCATIVE	4	2	2	50.00	50.00	50.00
TEMPORAL	7	7	1	87.50	100.00	93.33

Table 8.4: Results of cosubordinated clause analysis

in that the simplex algorithm is uncapable of correctly analysing the 15 gapping subordinate-type clauses. Having said this, the degree to which the subordinate gapping rule set outperformed the simplex algorithm goes beyond the scope of these 15 examples, particularly as a result of gap incompatibility judgements. Perhaps more important, however, is that the subordinate gapping rule set returned higher figures than the overall averages calculated during overall evaluation (see Table 8.1).

8.2.3 Evaluation of clause cosubordination

The basic cosubordinated clause interface method outlined in Section 7.1 was tested on a set of 185 relative clauses containing multiple clause instances marked with the *nagara*, *si* and *tutu* conjunctions, or ren'yo inflection. As a means of comparison, the basic simplex algorithm was used to analyse the same test set, and the accuracy on gapping relative clauses contained in the original test set was calculated. The results for the evaluation are given in Table 8.4.

Clearly, the expanded method of handling inter-clausal dependency outperforms the original algorithm, although the disparity between the respective results is perhaps not as marked as could have been expected. One of the main sources of error was that coindexed zero Subjects tended to be mistaken as Subject gaps, which accounted for around 75% of the errors in both cases. Perhaps more important, though, is the fact that hypothesis (4) was upheld for all observed cosubordinated relative clauses, and that the heuristic for distinguishing between cosubordinated and adverbial relative clauses worked successfully on all applications (see Section 7.1).

8.2.4 Evaluation of inter-personal relational verbs

As touched on in Section 4.3, relational verbs are involved with reciprocity of sense between the source and target case-roles, for target Co-actors and Co-patients. Here, we use the test case of inter-personal relational verbs to test system outputs for reciprocity of sense with the annotated case-role, to evaluate the severity of system disagreement for incorrect outputs.

The basic results for inter-personal relational verbs, ignoring sense reciprocity, are given in Table 8.5. In fact, inter-personal relational verbs performed the worst of all the test sets presented here, suggesting that verb sense reciprocity may be biasing the performance for the verb class. Additionally, there is a clear case of overgeneration for the Subject case-role, some of which may be misanalysed target slots.

However, on repeating the experiment with the relaxed evaluation criterion that either of the source or target case slots will be accepted for relational verb sense, the performance gain is not as signficant as could have been expected, although still significant. Nonetheless, our intuition that a proportion

	# instances	# correct	# incorrect	Precision	Recall	F-measure
Total	361	305	56	84.49	–	–
Case-role gapping	279	254	55	82.20	91.04	86.39
Head restrictive	78	47	1	97.92	60.26	74.60
Bound Subject	2	0	1	0.00	0.00	*N.C.*
Subject	154	150	34	81.52	97.40	88.76
Direct Object	22	18	8	69.23	81.82	75.00
Indirect Object	12	11	3	78.57	91.67	84.62
Co-actor	60	52	5	91.23	86.67	88.89
Target	5	3	0	100.00	60.00	75.00
Locative	11	9	3	75.00	81.82	78.26
Temporal	8	8	1	88.89	100.00	94.12
cdur	4	3	0	100.00	75.00	85.71
Time Relational	4	4	0	100.00	100.00	100.00

Table 8.5: Analysis of inter-personal relational verbs (without reciprocity)

	# instances	# correct	# incorrect	Precision	Recall	F-measure
Total	361	313	48	86.70	–	–
Case-role gapping	279	262	47	84.79	93.91	89.12
Head restricting	78	47	1	97.92	60.26	74.60
Bound Subject	2	0	1	0.00	0.00	-100.00
Subject	154	150	26	85.23	97.40	90.91
Direct Object	22	18	8	69.23	81.82	75.00
Indirect Object	12	11	3	78.57	91.67	84.62
Co-actor	60	60	5	92.31	100.00	96.00
Target	5	3	0	100.00	60.00	75.00
Locative	11	9	3	75.00	81.82	78.26
Temporal	8	8	1	88.89	100.00	94.12
Durational	4	3	0	100.00	75.00	85.71
Time Relational	4	4	0	100.00	100.00	100.00

Table 8.6: Analysis of inter-personal relational verbs (with reciprocity)

	# instances	# correct	# incorrect	Precision	Recall	F-measure
Total	92	87	5	94.57	–	–
Case-role gapping	79	78	5	93.98	98.73	96.30
Head restrictive	12	8	0	100.00	66.67	80.00
Bound Subject	1	1	0	100.00	100.00	100.00
Subject	72	72	4	94.74	100.00	97.30
Locative	1	1	0	100.00	100.00	100.00
Cardinal	0	0	1	0.00	*N.C.*	*N.C.*
Temporal	4	4	0	100.00	100.00	100.00
Durational	1	0	0	*N.C.*	0.00	*N.C.*
Time Relational	1	1	0	100.00	100.00	100.00

Table 8.7: Analysis of conflated ergative verbs

	# instances	# correct	# incorrect	Precision	Recall	F-measure
Total	290	268	22	92.41	–	–
Case-role gapping	278	262	10	96.32	94.24	95.27
Head restrictive	12	6	12	33.33	50.00	40.00
Bound Subject	10	8	2	80.00	80.00	80.00
Subject	241	227	4	98.27	94.19	96.19
Inclusive	27	27	4	87.10	100.00	93.10

Table 8.8: Analysis of Conjoining verbs

of the overgeneration in the Subject case slot may have been detracting from the results for Co-actors (there were no Co-patients in the test set) is proven correct, as the number of correctly analysed Subject case-roles remains unchanged, but an additional 8 Co-actors analyses are reproduced.

On mimicking the above test for reciprocity on generic relational verbs, some gain in performance resulted, but not to the same extext as seen here for inter-personal relational verbs. The reciprocity data for generic verbs is ommited from this thesis for reasons of space.

8.3 Verb class-specific evaluation

The remainder of this chapter is devoted to results for the individual verb classes developed in this research. As general trends, the movement and locational verbs produced lower overall accuracies than most other classes, dipping below the mean accuracy seen in the overall analysis, in the case of distal movement verbs. The verb class which stands out as requiring further attention is that of tool-based actions, although it is important to realise that the deflated results are no fault of the verb class characterisation, as there was only one Implement occurrence in the entire corpus.

	# instances	# correct	# incorrect	Precision	Recall	F-measure
Total	743	729	14	98.12	–	–
Case-role gapping	725	724	10	98.64	99.86	99.25
Head restrictive	18	5	4	55.56	27.78	37.04
Bound Subject	2	2	2	50.00	100.00	66.67
Subject	722	721	8	98.90	99.86	99.38
Temporal	1	1	0	100.00	100.00	100.00

Table 8.9: Analysis of the copula

	# instances	# correct	# incorrect	Precision	Recall	F-measure
Total	82	70	12	85.37	–	–
Case-role gapping	74	66	11	85.71	89.19	87.42
Head restrictive	8	4	1	80.00	50.00	61.54
Bound Subject	0	0	1	0.00	*N.C.*	*N.C.*
Subject	44	43	4	91.49	97.73	94.51
Direct Object	2	2	2	50.00	100.00	66.67
Medium	2	1	0	100.00	50.00	66.67
Target	1	1	0	100.00	100.00	100.00
Instrumental	0	0	1	0.00	*N.C.*	*N.C.*
Locative	4	4	1	80.00	100.00	88.89
Allative	14	9	1	90.00	64.29	75.00
Temporal	4	3	0	100.00	75.00	85.71
Durational	3	3	1	75.00	100.00	85.71

Table 8.10: Analysis of distal movement verbs

	# instances	# correct	# incorrect	Precision	Recall	F-measure
Total	156	139	17	89.10	–	–
Case-role gapping	124	113	17	86.92	91.13	88.98
Head restrictive	28	22	0	100.00	78.57	88.00
Subject	47	45	12	78.95	95.74	86.54
Direct Object	7	7	0	100.00	100.00	100.00
Indirect Object	0	0	1	0.00	*N.C.*	*N.C.*
Co-actor	49	42	3	93.33	85.71	89.36
Target	5	3	0	100.00	60.00	75.00
Locative	9	9	1	90.00	100.00	94.74
Temporal	4	4	0	100.00	100.00	100.00
Durational	3	3	0	100.00	100.00	100.00
Time Relational	4	4	0	100.00	100.00	100.00

Table 8.11: Analysis of empathy verbs

	# instances	# correct	# incorrect	Precision	Recall	F-measure
Total	119	115	4	96.64	–	–
Case-role gapping	116	114	4	96.61	98.28	97.44
Head restrictive	3	1	0	100.00	33.33	50.00
Bound Subject	1	1	0	100.00	100.00	100.00
Subject	2	1	2	33.33	50.00	40.00
Exclusive	112	112	2	98.25	100.00	99.12

Table 8.12: Analysis of excluding verbs

	# instances	# correct	# incorrect	Precision	Recall	F-measure
Total	119	111	8	93.28	–	–
Case-role gapping	115	108	7	93.91	93.91	93.91
Head restrictive	4	3	1	75.00	75.00	75.00
Subject	73	70	1	98.59	95.89	97.22
Direct Object	5	3	0	100.00	60.00	75.00
Co-actor	0	0	2	0.00	*N.C.*	*N.C.*
Target	7	5	0	100.00	71.43	83.33
Locative	24	24	4	85.71	100.00	92.31
Temporal	4	4	0	100.00	100.00	100.00
Durational	2	2	0	100.00	100.00	100.00

Table 8.13: Analysis of exisential verbs

	# instances	# correct	# incorrect	Precision	Recall	F-measure
Total	87	81	6	93.10	–	–
Case-role gapping	86	80	6	93.02	93.02	93.02
Head restrictive	1	1	0	100.00	100.00	100.00
Subject	58	55	0	100.00	94.83	97.35
Direct Object	3	2	0	100.00	66.67	80.00
Co-actor	0	0	2	0.00	*N.C.*	*N.C.*
Target	7	5	0	100.00	71.43	83.33
Locative	14	14	4	77.78	100.00	87.50
Temporal	3	3	0	100.00	100.00	100.00
Durational	1	1	0	100.00	100.00	100.00

Table 8.14: Analysis of experiential verbs

	# instances	# correct	# incorrect	Precision	Recall	F-measure
Total	336	300	36	89.29	–	–
Case-role gapping	306	284	24	92.21	92.81	92.51
Head restrictive	30	16	12	57.14	53.33	55.17
Bound Subject	10	8	1	88.89	80.00	84.21
Subject	266	254	12	95.49	95.49	95.49
Direct Object	5	5	3	62.50	100.00	76.92
Indirect Object	1	1	0	100.00	100.00	100.00
Co-actor	2	0	6	0.00	0.00	*N.C.*
Co-patient	1	1	0	100.00	100.00	100.00
Target	5	3	0	100.00	60.00	75.00
Locative	10	7	2	77.78	70.00	73.68
Cardinal	1	0	0	*N.C.*	0.00	*N.C.*
Temporal	4	4	0	100.00	100.00	100.00
Durational	1	1	0	100.00	100.00	100.00

Table 8.15: Analysis of generic relational verbs

	# instances	# correct	# incorrect	Precision	Recall	F-measure
Total	81	81	0	100.00	–	–
Case-role gapping	0	0	0	*N.C.*	*N.C.*	*N.C.*
Idioms	81	81	0	100.00	100.00	100.00

Table 8.16: Analysis of idioms

	# instances	# correct	# incorrect	Precision	Recall	F-measure
Total	42	37	5	88.10	–	–
Case-role gapping	40	36	5	87.80	90.00	88.89
Head restrictive	2	1	0	100.00	50.00	66.67
Subject	12	8	1	88.89	66.67	76.19
Inclusive	28	28	4	87.50	100.00	93.33

Table 8.17: Analysis of including verbs

	# instances	# correct	# incorrect	Precision	Recall	F-measure
Total	2661	2305	356	86.62	–	–
Case-role gapping	2219	2037	337	85.80	91.80	88.70
Head restrictive	426	253	19	93.01	59.39	72.49
Bound Subject	31	18	16	52.94	58.06	55.38
Subject	1613	1536	171	89.98	95.23	92.53
Direct Object	251	219	74	74.74	87.25	80.51
Indirect Object	8	6	4	60.00	75.00	66.67
Co-actor	53	43	7	86.00	81.13	83.50
Medium	13	4	1	80.00	30.77	44.44
Target	7	5	1	83.33	71.43	76.92
Passive Agent	1	0	0	*N.C.*	0.00	*N.C.*
Causee	0	0	1	0.00	*N.C.*	*N.C.*
Instrumental	1	1	1	50.00	100.00	66.67
Locative	74	56	38	59.57	75.68	66.67
Cardinal	10	6	5	54.55	60.00	57.14
Allative	22	16	5	76.19	72.73	74.42
Perlative	2	1	0	100.00	50.00	66.67
Temporal	84	80	6	93.02	95.24	94.12
Durational	44	43	7	86.00	97.73	91.49
Time Relational	15	15	0	100.00	100.00	100.00
Idioms	1	0	0	*N.C.*	0.00	*N.C.*
Subordinate gapping	17	17	1	94.44	100.00	97.14

Table 8.18: Analysis of locational action verbs

	# instances	# correct	# incorrect	Precision	Recall	F-measure
Total	95	81	14	85.26	–	–
Case-role gapping	83	75	14	84.27	90.36	87.21
Head restrictive	12	6	0	100.00	50.00	66.67
Bound Subject	1	1	4	20.00	100.00	33.33
Subject	60	59	5	92.19	98.33	95.16
Direct Object	12	12	5	70.59	100.00	82.76
Indirect Object	1	0	0	*N.C.*	0.00	*N.C.*
Medium	4	3	0	100.00	75.00	85.71
Locative	3	0	0	*N.C.*	0.00	*N.C.*
Allative	1	0	0	*N.C.*	0.00	*N.C.*
Durational	1	0	0	*N.C.*	0.00	*N.C.*

Table 8.19: Analysis of non-quantative verbs

	# instances	# correct	# incorrect	Precision	Recall	F-measure
Total	255	232	23	90.98	–	–
Case-role gapping	232	212	12	94.64	91.38	92.98
Head restrictive	23	20	11	64.52	86.96	74.07
Bound Subject	16	13	2	86.67	81.25	83.87
Subject	196	186	4	97.89	94.90	96.37
Direct Object	1	1	1	50.00	100.00	66.67
Co-actor	0	0	1	0.00	*N.C.*	*N.C.*
Locative	4	2	2	50.00	50.00	50.00
Cardinal	2	1	0	100.00	50.00	66.67
Allative	4	1	1	50.00	25.00	33.33
Temporal	5	4	0	100.00	80.00	88.89
Durational	3	3	1	75.00	100.00	85.71

Table 8.20: Analysis of partitive verbs

	# instances	# correct	# incorrect	Precision	Recall	F-measure
Total	27	24	3	88.89	–	–
Case-role gapping	19	18	3	85.71	94.74	90.00
Head restrictive	6	4	0	100.00	66.67	80.00
Subject	16	16	2	88.89	100.00	94.12
Allative	1	1	0	100.00	100.00	100.00
Time Relational	2	2	0	100.00	100.00	100.00

Table 8.21: Analysis of proximal movement verbs

	# instances	# correct	# incorrect	Precision	Recall	F-measure
Total	690	595	95	86.23	–	–
Case-role gapping	595	538	91	85.53	90.42	87.91
Head restrictive	92	55	4	93.22	59.78	72.85
Bound Subject	7	3	3	50.00	42.86	46.15
Subject	441	414	43	90.59	93.88	92.20
Direct Object	98	83	25	76.85	84.69	80.58
Indirect Object	2	1	1	50.00	50.00	50.00
Co-actor	2	1	1	50.00	50.00	50.00
Medium	4	0	0	*N.C.*	0.00	*N.C.*
Locative	10	7	11	38.89	70.00	50.00
Cardinal	1	0	0	*N.C.*	0.00	*N.C.*
Temporal	13	12	3	80.00	92.31	85.71
Durational	16	16	4	80.00	100.00	88.89
Time Relational	2	2	0	100.00	100.00	100.00
Idioms	1	0	0	*N.C.*	0.00	*N.C.*
Subordinate gapping	17	17	1	94.44	100.00	97.14

Table 8.22: Analysis of quotative verbs

	# instances	# correct	# incorrect	Precision	Recall	F-measure
Total	47	30	17	63.83	–	–
Case-role gapping	38	27	17	61.36	71.05	65.85
Head restrictive	9	3	0	100.00	33.33	50.00
Bound Subject	1	1	1	50.00	100.00	66.67
Subject	21	19	8	70.37	90.48	79.17
Direct Object	4	4	7	36.36	100.00	53.33
Indirect Object	1	0	0	*N.C.*	0.00	*N.C.*
Medium	5	0	0	*N.C.*	0.00	*N.C.*
Instrumental	1	1	1	50.00	100.00	66.67
Locative	2	1	0	100.00	50.00	66.67
Cardinal	1	0	0	*N.C.*	0.00	*N.C.*
Allative	1	0	0	*N.C.*	0.00	*N.C.*
Durational	1	1	0	100.00	100.00	100.00

Table 8.23: Analysis of tool-based action verbs

	# instances	# correct	# incorrect	Precision	Recall	F-measure
Total	97	80	17	82.47	–	–
Case-role gapping	72	69	17	80.23	95.83	87.34
Head restrictive	25	11	0	100.00	44.00	61.11
Subject	53	53	15	77.94	100.00	87.60
Direct Object	2	1	0	100.00	50.00	66.67
Indirect Object	1	1	0	100.00	100.00	100.00
Medium	3	3	0	100.00	100.00	100.00
Locative	0	0	1	0.00	*N.C.*	*N.C.*
Cardinal	1	1	0	100.00	100.00	100.00
Allative	3	3	0	100.00	100.00	100.00
Perlative	4	3	1	75.00	75.00	75.00
Temporal	2	2	0	100.00	100.00	100.00
Durational	2	2	0	100.00	100.00	100.00

Table 8.24: Analysis of travelling verbs

Chapter 9

Conclusions

In this thesis, we have proposed an account of relative clause-hood, focusing on the precepts of case-role construal and boundedness of case-roles within the relative clause body. Analysis of case-role gapping was then described based along the lines of three main paradigms:

- First, we introduced an **argument status hierarchy**, defined according to affinity with the predicate, which assigns behavioural properties to the different argument types according to rank. This argument type hierarchy was applied to define a case-role schema, and predict the syntactic nature of each case-role type based on the inherent features of each argument type. Additionally, the basic ranking of argument types was used to apply preferences to argument types to accommodate the case-role gap, for which purpose the verb class hierarchy was linked in closely with argument stasis, and the valency frame dictionary referenced by the resolution system tagged accordingly.
- Next, we defined a 17-way **case-role schema** with which to analyse case-role gapping in the relative clause context. All effort was made to provide tests for the less intuitive case-role types, and document their syntactic and semantic behvioural patterns, so as to make them reproduceable in an alternate system/theoretical context. Case-roles were allocated a unique argument status, which was used to predict basic accessibility to case-role gapping for the various verb classes proposed. One factor which set the core component of the case-role schema apart from traditional Case accounts was its grammatical dependence, with the central case-roles as the Subject, Direct Object and Indirect Object, and case-role transformation occurring readily under modal transformation.
- Last, we proposed a **verb class hierarchy** with which to map links between case-roles intrinsic to that verb class, and predict adjunct (in)compatibility. As an account of the basic semantic of each class, a simple conditional-based rule set was described for each rule set, which can be applied in case-role gapping resolution as a realisation of the interdependence between case-roles and to the propotypical verb sense.

The individual verb class rule-sets were complemented with a set of adjunct-based filters, which interface with the verb class rule-sets through sortal preferences and relative weighting. Basic weighting mechanisms were then described, which are combined additively in the context of multiple analyses from the different rule-sets. Discussion was next made of clausal cosubordination, and clausal and noun head coordination, and the roles they can play in restricting the scope of interpretation and offsetting local sortal preferences.

Finally, a full account was given of the system resources in carrying out the resolution task, including the means used to extract and structure the system dictionary.

Under evaluation, the proposed system produced a mean accuracy of around 89% on the test set of 4615 relative clause instances. Separate experimentation was also documented to indicate the

effectivity of fixed expressions in the system dictionary, and applicability of inter-clausal relations to narrow the scope of case-slot interpretations and identify alternate host clauses for the case-slot gapping process.

9.0.1 Future research

The major area for improvement lies in the closer analysis of head restrictive relative clause types, and providing an account thereof. This point comes across clearly from the system evaluation, for which the performance on case-slot gapping was inevitably higher than that on head restrictive clauses.

The very nature of the verb class hierarchy suggests virtually unlimited scope for expansion, additionally, either depth-wise in increasing the granularity for currently modelled semantic types, or breadth-wise in identifying new typologies of verbs/new parameters with which to classify verbs orthogonally to the existing framework.

In terms of additional applications, the combination of verb classes, argument status and case-roles could be tried out on a borader-ranging task, such as discourse processing or semantic-based parsing.

Appendix A

The Relative Clause Resolution System

The valency dictionary utilised by the system was extracted from the NTT valency dictionary (Ikehara *et al.* 1997) and subsequently modified/expanded. First, description will be given of the original structure of the NTT valency dictionary, followed by discussion of the extraction method and data incorporated/excluded from the system valency dictionary.

A.1 The system valency dictionary

A.1.1 Valency frame types

The system valency dictionary is based on a maximally generalised default valency frame for each verbal stem, complemented with full clause-based idioms, fixed expressions and 'conditional instantiated' entries.

Default valency frame

Every verbal stem compatible with non-fixed expression usage (see below) is attributed with a (unique) default valency frame. This simply stipulates the obligatory and optional complement content of the hypothetical 'default sense' of that verb, including candidate case marking for each case slot. Case slots are given in their unmarked ordering, and are supplied with a case-role index. In terms of case, the unmarked ordering is generally defined by:

SUBJECT > *DIRECT OBJECT* > *INDIRECT OBJECT*

One inevitable failing of any attempt to attribute a unique valency frame to an essentially lexical verb representation, is the inability to combine verb senses involving distinct case marking/valency. Assuming identical **complement** valency, however, it is often possible to merge such multiple valency frames simply by combining corresponding case slots. Naturally, this also assumes case slot correspondence, a sometimes unrealistic expectation.

One method of overcoming valency variation in cases where subsumption of valency frame content occurs, is to mark coordinated/optional case slots within the valency frame. This approach is generally applicable to the Indirect Object, Co-actor and Co-patient case slots for relational verbs.

An alternative method is to take the intersection of conflicting candidate valency frames, and generate expanded derivative valency frames, as required through the use of verb classes. This is the method applied for conflated ergative and partitive verbs.

A more serious dilemna results from overlap in case marking between distinct case slots, such that the basic marking produces a one-to-many mapping onto the valency frame. In these cases, ambiguous case marking is retained in the valency frame and various heuristics are applied to resolve case slot ambiguities at run time.

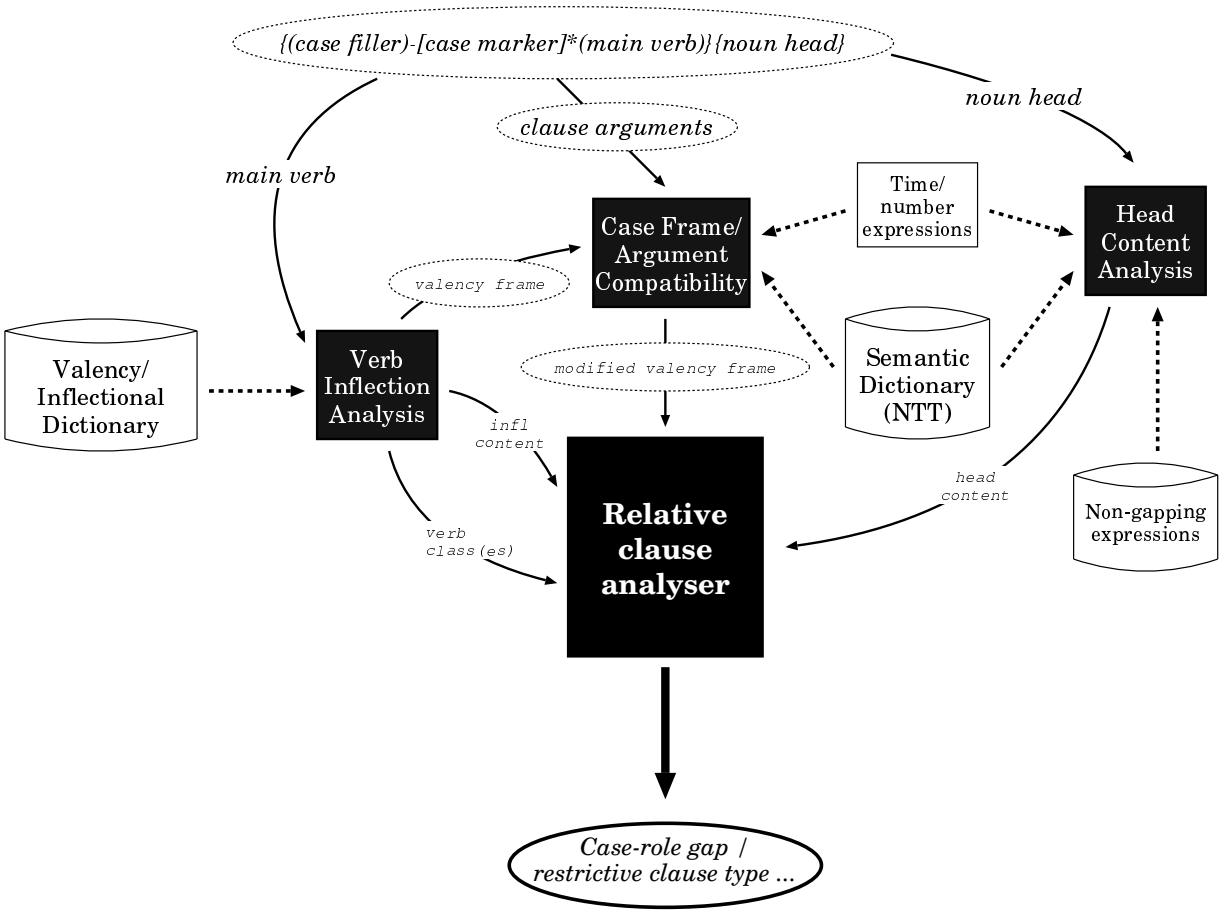

Figure A.1: The full system hierarchy

Additionally, in merging case slots and case marking, we potentially create ungrammatical case marker combinations. Consider the case of stative verbs (i.e. *deki(-ru)* "can do/be done", *ar(-u)* "to be/have", *et al.*), which exhibit the case marking characteristics described in the following connection matrix:

		DObject	
		NOM	ACC
	TOP	yes	yes
Subject	NOM	yes	yes
	DAT	yes	no

That is, the Subject/Direct Object case marking patterns of TOP/NOM, TOP/ACC, NOM/NOM, and so on are acceptable, but DAT/ACC produces ungrammaticality.

In collapsing these combinations into a single valency frame representation, we produce the following, inferring the existence of this ungrammatical DAT/ACC marking type:

(1) *Subject-* { DAT / TOP/NOM / NOM } *DObject-* { NOM / NOM / ACC } *deki-ru*

While recognising this overmodelling characteristic of the valency frame representation used, I suggest that this does not pose problems for pure analysis purposes.

Full clause-based idioms

Full clause-based idioms are idioms defined at the full relative clause complex level. Here, the full range of a predetermined valency frame must be realised within the relative clause, and main verb inflection is often highly restricted in scope. Additionally, the head is commonly restricted to a single lexical candidate.

Representation of full clause-based idiom entries is made by way of a valency frame, including any lexical fillers, explicit descripition of the full lexical content of the main verb and a surface description of compatible heads; unrestricted head compatibility is indicated by an asterisk. Recognition of full clause-based idioms is based on a full match in terms of the case slot and main verb content, and also the head if so required. Unlike fixed expressions, case marking variation is disallowed, and incompatibility is assumed with all adjunct case slots.

For analysis purposes, full clause-based idioms are simply tagged as being an "IDIOM", with no attempt made to classify the relationship between the head and clause body. This is justified by the fact that the full relative clause complex construes a single idiomatic unit.

Fixed expressions

Fixed expressions are defined as entries which contain at least one integral complement in their valency frame; any such complement case slots are provided with a set of lexical fillers. To be triggered, all lexically instantiated case slots must match with the system input. As a means of ensuring no overlap between fixed expressions, the system valency dictionary has been designed such that lexical fillers are mutually exclusive in surface content for a given verbal stem and case slot. That is, for a given verbal stem, no two fixed expressions share any subset of the fixed case element content. However, this guarantee of mutual exclusivity is not sufficient in itself to guarantee at most one fixed expression for an arbitrary system input, as case element correspondence can potentially occur between distinct case slots in the input for separate fixed expression entries.

One issue to arise from the inclusion of fixed expressions in the case dictionary is whether instantiated case fillers can be gapped to become the noun head ('*displaceability*'). Here, there is a distinct division

between 'idiomatic-type' fixed expressions and 'case element-defined' fixed expressions. In general, the semantics of idiomatic-type fixed expressions are not intuitively accessible from the independent meanings of the verbal stem and instantiated case fillers, and displacement of any one of the case fillers removes the idiomatic sense. For case element-defined fixed expressions, on the other hand, the fixed case element(s) tend to be intrinsically limited in default sense, and simply restrict rather than modify the semantic content of the root verb. Indeed, some of the case element-defined fixed expressions extracted from the NTT dictionary were questionable as to their true idiomatic status, but for reasons of economy and consistency, no attempt was made to filter off such usages.

In terms of argument status, most integral complements produce idiomatic-type fixed expressions.

The NTT valency dictionary does not contain information on the displaceability of fixed case elements from within relative clauses, and as such, all fixed expressions were manually analysed, and displaceable fixed case elements marked explicitly as being such. The method used to judge the displaceability of each fixed case slot was a threefold one.

Firstly, the fixed case element in question was moved from within the valency frame to assume the role of the head of the derived relative clause. Any non-fixed case slots were then instantiated with appropriately non-specific arguments before comparing the semantics of the resulting relative clause with the original semantics of the source sense to check for coincidence of sense.

Next, the relative clause complex produced above was considered with all non-fixed case slots elipted. To fulfil the displaceability requirement, the gapped case slot had to be uniquely identifiable as having been gapped from its original case slot, within the scope of the original sense of that entry.

Finally, the freedom of case slot order was tested within a matrix clause context ("*scrambling*"), with appropriate non-specific arguments used as fillers for the non-fixed case slots. Despite Japanese being well known as a free word order language, different permutations of case slots can produce variable acceptability. Thus, freedom of case slot order was measured by the existence of at least one case frame permutation which involved the fixed case slot in question, and not by requiring that all such permutations lead to an acceptable case slot order. Idiomatic fixed expressions are characterised by fixed case slots generally being final in the case slot ordering, and adverbs or adverbial arguments not being insertable between the fixed case slot and predicate. In the case that either of these conditions can be violated for a given fixed case slot, that case slot is generally displaceable in the relative clause context.

Case slots which fulfilled all of the above requirements were judged to be fully displaceable, and marked as such.

Conditional instantiated expressions

Similarly to fixed expressions, conditional instantiated entries have a given set of case slots which must be realised in the system input for that valency frame to be triggered. However, with conditional instantiated entries, the case slots requiring instantiation are unrestricted as to their lexical content (for our purposes at least). Rather, the given conditional case slots are assumed to be uniquely identifiable from their associated case markers. Naturally, the conditional case markers must be disjunctive with the content of the default valency frame, as well as conditional case slots in other conditional instantiated entries for that same root verb.

A.1.2 Valency frame type preferences

Despite inbuilt guarantees about the representational integrity of fixed expressions, and between the case marker content of conditional instantiated and default valency frame entries, potential overlap exists between other entry types. This problem is overcome by establishing valency frame type preferences such that full clause-based idioms are preferred over fixed expressions, which are in turn preferred over conditional instantiated entries, with conditional instantiated expressions overriding the default

valency frame.

A.1.3 Dictionary entry content

All dictionary entries are supplied with their verbal stem, inflectional type, a string of verb classes, and a description of head compatibility. The full range of verb classes is described in Figure 4.1, whereas the notions of verbal stem representation, inflectional types and head compatibility are described below.

Verbal stem representation

The verbal stem is the uninflecting "head" of the verb, and must have a phonological content equivalent to at least one kana character. As described in Section 6.1, representational variation commonly arises for a given verb, partly due to the co-existing lexicographic systems (*hiragana*, *katakana*, and *kanji*), and partly due to inconsistency in the phonological content attributed to kanji characters.

The issue of lexical inconsistency is confused further by the fact that component characters in kanji compounds can usually be rewritten as their kana equivalents. To take an example, the stem of the Sino-Japanese verb tyōtatu-suru [調達する] "to procure" is made up of the kanji compound 調達, with the prefix of '調' corresponding to the phonological content of *tyō* and the suffix of '達' corresponding to *tatu*. Clearly defined phonological correspondences of this type result from coincidence with the "basic" reading(s) of each of the component kanji (see (Shibatani 1990:130))[1] such that *tyō* and *tatu* are basic readings of 調 and 達, respectively. We will term the component-rewritable kanji compounds of this type 'decomposable'. Decomposable kanji compounds potentially have at least 3^n distinct lexical forms, where n is the number of component kanji contained in the compound.[2] Realistically, the full extent of these representations would not be realised, but this figure gives an indication of the extent of lexical ambiguity inherent in the Japanese writing system. Instances where the kanji-wise decomposable nature of kanji compound verbs becomes an issue arise when one or more of the component kanji have fallen into disuse or are particularly complex, and the kanji in question are individually replaced with kana. One verb for which this phenomenon is commonly observed is ussekisuru [鬱積する] "bottle up", and the initial kanji of '鬱'. A common mechanism to mark such usages is to use katakana instead of the default hiragana script, although hiragana can equally be employed.

Note that component kanji-kana equivalence does not occur for all verbs, and that there are rare cases of compound kanji representations forming a single undecomposable unit ('unit' kanji compounds). One such example is kaze [風邪] "the common cold", where the component characters of '風' and '邪' do not correspond to the kana content of *ka* and *ze* respectively, but rather kana correspondence occurs on a compound basis.

Potential lexical inconsistency for decomposable kanji compounds is represented within the system dictionary through the use of regular expressions, in that the hiragana equivalent is given for each unit. These regular expressions are expanded to generate all possible kanji/hiragana verbal stem combinations in the run-time system valency dictionary. The expanded dictionary entries are combined into equivalence classes within the run-time valency dictionary, based on lexical equivalence of the expanded verbal stem. Entries sharing a common stem are further partitioned into subclasses, according to inflectional type.

This does not solve the hiragana/katakana ambiguity, however, and if a match for a string containing katakana characters is not found at any stage by the system, all katakana characters are converted to

[1] Note that "compounding" effects resulting from combining kanji can produce predictable phonological variation from the base forms, instances of which are included in the notion of 'coincidence of basic reading'. See (McCawley 1968; Ohno and Shibata 1977; Tsujimura 1996; Vance 1987) for further details.

[2] This assumes each kanji character is replaced either by its full hiragana or katakana equivalent, with no mixed use of hiragana/katakana character replacements for any one kanji character. Additionally, kana stem inconsistencies are not included in this figure.

their hiragana equivalents in rematching the string. The reason that katakana-hiragana conversion is carried out only after failure to detect a match, is that there are significant numbers of verbs for which a substring of the stem **must** be in katakana. This occurs particularly for words which originated from non-Japanese sources (commonly European languages), such as *dabur-u* "to double up/overlap".

Inflectional types

Inflectional types are given as an index to the inflectional class that verbal stem belongs to, through which the full inflectional paradigm for that verb is defined.

Head compatibility

Each verb sense is provided with a head type. The head type is used in the detection of full relative clause-based idioms (see above), by way of describing the lexical head for each full idiom entry. For other dictionary entry types, this head type is necessarily given as an asterisk ("don't care").

A.2 NTT valency dictionary

The NTT valency dictionary (Shirai *et al.* 1997; Ikehara *et al.* 1997; Bond and Shirai 1997) was designed for use with the NTT ALT-J/E machine translation system, and relies primarily on valency frames constrained by semantic class data interfacing with the NTT thesaurus.[3]

Verb sense disambiguation is carried out based on semantic constraints applied to each case slot, for which purposes combinations of thesaurus classes and explicit lexical entries are employed. Two main types of case element entries exist: explicit case fillers provided within the valency frame, and external constraint sets (lexical and/or thesaurus based) indexed to their corresponding case slot. This differentiation is made to distinguish between, respectively, lexical case elements which must match in full with the content of that case slot in the input (including the surface case marker), and those for which only the semantic head of that case slot is considered. Within this second type of head-based representation there are a significant number of case slots where only lexical case filler candidates are given in the constraint set.

Based on these case slot types, dictionary entries can be classified as being either 'fixed expressions' or 'generalised'. Fixed expressions are defined as having at least one case slot for which only explicit lexical categories are supplied, and include both internal and external case frame representation of the case element content. Unlike generalised expressions, which can occur with an arbitrary set of case slots ellipted, fixed elements require that their full case element content be instantiated to be realised.

Due to the NTT valency dictionary having been developed specifically to produce interlingual sense correspondence, the granularity of sense division is generally finer than is the case for a monolingual Japanese case frame dictionary such as the IPAL verb dictionary (IPA 1987).

A unique English translation is supplied for each verb sense, including mappings between corresponding case slots. Additionally, each dictionary entry is supplied with a unique verb aspectual class, a set of verbal semantic classes, and details of inflectional types incompatible with that particular verb sense.

A.3 System valency dictionary extraction

The system valency dictionary was extracted from the NTT valency dictionary by taking the first sense for each verbal stem as the default valency frame and comparing this with remaining generalised

[3] In this paper, I describe the use of deep case markers, which are contained in the full version of the valency dictionary as employed by NTT, but not provided within (Ikehara *et al.* 1997). Details of the range of deep case markers and the derivation process can be found in (Bond and Shirai 1997).

(non-fixed expression) verb senses of that same verbal stem. Comparison was based on (a) argument status, (b) deep case correspondence, and (c) case marker correspondence for matching case slots.

Case slots were first evaluated in terms of their adjunct/complement status. This was achieved through direct application of the deep case analysis given within each valency frame. Based on this disambiguated argument status, any adjunct case slots were tentatively removed from the valency frame.

Deep case case comparison was facilitated by superimposing complement valency frame 'skeletons' onto the default valency frame skeleton. In the case of a full match, corresponding case markers were simply merged, whereas instances of the current valency frame constituting a past of the default valency frame were treated by similarly merging the case markers for the set of matched case slots. Any instances of the current valency frame being a superset of the default valency frame were marked for manual analysis. Likewise, occurrences of partial overlap/full disjunction was marked for later analysis.

The verb class content of the default valency frame was determined through the union of the verb class content of other generalised (non-fixed expression) verb senses of that same verbal stem.

Following extraction of the default valency frame for each verbal stem, fixed expressions were extracted individually, and simply converted into the desired format for inclusion into the system valency dictionary. Within the fixed expressions contained in the NTT dictionary, there is a small proportion of lexical overlap resulting from multiple senses being attributed to a single valency frame. In general, the valency frame content of these overlapping entries coincides fully, and subsequent occurrences of a given valency frame can simply be ignored. For the limited number of overlapping fixed-expression valency frames for which this was not the case, the same merging process as applied for generalised entries was employed.

All full clause-based idioms included in the system valency dictionary were manually added, as clause-based expressions are not included in the NTT valency dictionary.

A.3.1 Manual correction/modification of the system valency dictionary

Subsequent to the extraction of the system valency dictionary, 'displaceable' fixed case elements were identified and marked as being such. Additionally, any remaining adjunctive case slots were manually discluded, and complement case slots attributed with a deep case ID.

Ambiguities arising from deep case analysis of the NTT dictionary case frames were resolved wherever possible. In cases where genuine incompatibility existed, consideration was given to the justification/tenability for a verb class-based analysis, or generation of alternative entry types. Close analysis of commonly occurring instances of valency variation revealed a number of trends in the data.

Firstly, there was a significant number of instances of ergative parallelism for a single verbal stem. In this, the accusatively marked Direct Object case slot for the transitive sense is transformed into the nominatively marked Subject case slot in the intransitive sense; this phenomenon is detectable through the sortal restrictions on the respective case slots. This effect was particularly noticeable for Sino-Japanese verbs, and these two senses were frequently the only interpretations listed for that verbal stem. Given the well-defined nature of this parallelism and relatively high frequency of occurrence, a verb class was established for entries of this type, and a consistent intransitive valency frame representation enforced (i.e. the Direct Object is not contained in the valency frame representation).

One primary design decision involved with the use of verb classes was to incorporate adjunct case representation implicitly in the verb class designation, and retain only complement case slots in the valency frame. In terms of the system algorithm, this amounts to verb-specific handling of adjuncts, with general complement resolution being standardised.

To aid in the evaluation process and enhance tractability of the resolution process, a deep case marking paradigm was introduced. This involved analysing each complement case slot in a given valency frame and classifying that case as being one of the following types:

Base entry #	Derived entry #	Verb root/ inflection type combinations	Average entries per combination
7767	26752	10601	2.52

Table A.1: Case valence dictionary statistics

Essentially, this equates to labelling each case slot with a unique ID, and identifying the case slot trace through the use of this ID. However, by way of establishing semantic equivalence classes of deep class IDs, evaluation can be made of parallelism between analysis of different verb stem and within verb classes.

While these are strong correspondences between typical surface case markers and deep case IDs, significant deviation exists.

In particular, explicit deep case identification is required to distinguish between distinct case slots marked with an identical surface case marker, and also between valency frame-defined case slots and verb class/algorithm-defined case slots.

A.3.2 Verb inflectional analyser

The verb inflectional analyser operates in a deterministic, bottom-up manner, and attempts to generate the input stem verb from an arbitrary verb root and inflectional type. For successful parses, the analyser returns an inflectional analysis of the stem verb. Although any given combination of a verb root and inflectional type will lead to a unique output, multiple successful parses can potentially exist for one stem verb, derived from distinct verb root/inflectional type combinations. This is one drawback from deriving all possible *hiragana/kanji* combinations of each verb root, and at present, the system cannot choose between multiple successful parses simply from the lexical representation of the stem verb.

Inputs to the verb inflectional analyser correspond to entries in the case valence dictionary for which the verb root is a prefix of the stem verb under analysis. As is evidenced by table A.1, the current verb valence dictionary has an average of 2.52 entries for each distinct verb root/inflectional type combination, as a result of both the generation of all lexical derivatives of the verb root, and the existence of multiple verb senses through fixed and conditional expressions. To avoid duplicating inflectional processing, entries with common verb root and inflectional types are merged together for inflectional analysis purposes. For successful parses, these merged entries are then decomposed into their constituent parts, and outputted to the valency frame compatibility analyser.

A.3.3 Valency frame compatibility analyser

The principal role of the valency frame compatibility analyser is to match relative clause arguments with fixed/conditional expression arguments to ascertain whether each such candidate verb sense is compatible with the contents of the relative clause input. An additional role is in the determination of the solution for a given valency frame, by using the order of the case slots to test each for surface realisation (including parallel candidate markers for each case slot), using the inherent reverse-order of the frame.

An equally important role performed at this level is the decomposition of valency frames for relational verb senses, and detection of target and source case elements in the input (see below for details).

A.3.4 Semantic dictionary / lexical analysers

The semantic dictionary utilised in our research is that developed by NTT (Ikehara *et al.* 1993), which is then combined with pronoun, number and temporal analysers, devised based on regular expressions

to maximise efficiency and dynamicism. The principle use made of the semantic dictionary is in classifying noun heads as being animate (person/organisation), locative or abstract. This information is used to override the inherent preferential ordering of the valency frame, according to the verb class of the stem verb.

Bibliography

Baldwin, T. 1998. Relative clause coordination and subordination in Japanese. In *Proc. of the Australian Natural Language Processing Postgraduate Workshop*, 1–10.

——, H. Tanaka, and T. Tokunaga. 1997a. Analysis of head gapping in Japanese relative clauses. In *Information Processing Society of Japan SIG Notes*, volume 97, no. 4, 1–8.

——, T. Tokunaga, and H. Tanaka. 1997b. Semantic verb classes in the analysis of head gapping in Japanese relative clauses. In *Proc. of the 4th Natural Language Processing Pacific Rim Symposium 1997 (NLPRS'97)*, 409–14.

Blake, B. 1990. *Relational Grammar*. Routledge.

—— 1994. *Case*. CUP.

Bond, F., and S. Shirai, 1997. *Practical and Efficient Organization of a Large Valency Dictionary*. Handout at the *Workshop on Multilingual Information Processing, held in conjunction with NLPRS'97*.

Brinker, K. 1972. *Konstituentenstrukturgrammatik und operationale Satzgliedanalyse: Methodenkritische Untersuchungen zur Syntax des einfachen Satzes im Deutschen*. Frankfurt am Main: Athenäum Verlag.

EDR, 1995. *EDR Electronic Dictionary Technical Guide*. Japan Electronic Dictionary Research Institute, Ltd. (In Japanese).

Emons, R. 1974. *Valenzen englischer Prädikatsverben*. Tübingen: Niemeyer.

Fillmore, C.J. 1968. The case for case. In *Universals in linguistic theory*, ed. by E. Bach and R.T. Harms, 1–88. New York: Holt, Rinehart and Winston.

Foley, W., and R. Van Valin. 1984. *Functional Syntax and Universal Grammar*. CUP.

Good, I.J. 1965. *The Estimation of Probabilities*. MIT Press, Cambridge MA.

Halliday, M. A. K. 1970. Language structure and language function. In *New Horizons in Linguistics*, ed. by J. Lyons, 140–65. Harmondsworth, Middlesex: Penguin.

—— 1994. *An Introduction to Functional Grammar*. Edward Arnold, 2nd edition.

Hasegawa, Y. 1996. *A Study of Japanese Clause Linkage*. CSLI.

Heinz, H. 1978. On the possibility of distinguishing between complements and adjuncts. In *Valence, Semantic Case, and Grammatical Relations*, ed. by W. Abraham, 22–45. Amsterdam: John Benjamins.

Helbig, G., and W. Schenkel. 1973. *Wörterbuch zur Valenz und Distribution deutscher Verben*. Leipzig: VEB Verlag Enzyklopädie.

Hoshi, H., 1994. *Passive, Causive, and Light Verbs: A Study of Theta Role Assignment*. University of Connecticut dissertation.

Huddleston, R. 1976. *An Introduction to English Transformational Grammar*. Longman.

Iida, M. 1996. *Context and Binding in Japanese*. CSLI Publications.

Ikehara, S., M. Miyazaki, and A. Yokoo. 1993. Classification of language knowledge for meaning analysis in machine translation. *Transactions of the Information Processing Society of Japan* 34.1692–1704. (In Japanese).

——, ——, ——, S. Shirai, H. Nakaiwa, K. Ogura, Y. Ooyama, and Y. Hayashi. 1997. *Nihongo Goi Taikei – A Japanese Lexicon*. Iwanami Shoten. 5 volumes. (In Japanese).

Inoue, K. 1976. *Henkei Bunpo to Nihongo*. Tokyo: Taishukan.

IPA, 1987. *IPA Lexicon of the Japanese Language for Computers*. (In Japanese).

Jacobsen, W.M. 1992. *The Transitive Structure of Events in Japanese*. Kurosio Publishers.

Kameyama, M. 1995. The syntax and semantics of the Japanese Language Engine. In *Japanese Sentence Processing*, ed. by R. Mazuka and N. Nagai, 153–76. Lawrence Erlbaum Associates.

Kanzaki, K. 1997. Lexical semantic relations between adnominal constituents and their head nouns. *Mathematical Linguistics* 21.53–68. (In Japanese).

——, and H. Isahara. 1997. Lexical semantics for adnominal constituents in Japanese. In *Proc. of the 4th Natural Language Processing Pacific Rim Symposium 1997 (NLPRS'97)*, 573–6.

Keenan, E. L., and B. Comrie. 1977. Noun phrase accessibility and universal grammar. *Linguistic Inquiry* 8.63–99.

Kuno, S. 1973a. *Nihon Bunpō Kenkyū*. Taishukan. (In Japanese).

—— 1973b. *The Structure of the Japanese Language*. MIT Press, Cambridge MA.

—— 1976. Subject, theme, and the speaker's empathy [A reexamination of relativization phenomena]. In *Subject and Topic*, ed. by C. Li, 419–44. New York: Academic Press.

—— 1978. *Danwa no Bunpō*. Tokyo: Taishukan. (In Japanese).

—— 1987. *Functional Syntax*. Chicago: University of Chicago Press.

——, and E. Kaburaki. 1977. Empathy and syntax. *Linguistic Inquiry* 8.627–72.

Kuroda, S-Y. 1979. On Japanese passives. In *Explorations in Linguistics: Papers in Honor of Kazuko Inoue*, ed. by G. Bedell, E. Kobayashi, and M. Muraki, 305–47. Tokyo: Kenkyusha.

—— 1980. Bun no kozo. In *Nichi-ei Hikaku Koza II: Bunpo*, ed. by T. Kunihiro, 23–62. Tokyo: Taishukan. (In Japanese).

Levin, B. 1993. *English Verb Classes and Alterations*. University of Chicago Press.

Machine Translation System Laboratory (MTSL). 1995. *Expressing in Interlingua Ver. 2*. Technical report, Center of the International Cooperation for Computerization (CICC).

MARTIN, S. E. 1975. *A Reference Grammar of Japanese*. Tuttle.

MATSUMOTO, YOSHIKO. 1990. Role of pragmatics in japanese relative clauses. *Lingua* 82.111–29.

——. 1996. Interaction of factors in construal: Japanese relative clauses. In (Shibatani and Thompson 1996).

——. 1997. *Noun Modifying Constructions in Japanese*. John Benjamins.

MCCAWLEY, J. D. 1968. *The Phonological Component of a Grammar of Japanese*. The Hague: Mouton.

—— 1976. Relativization. In *Syntax and Semantics 5: Japanese Generative Grammar*, ed. by M. Shibatani, 295–306. New York: Academic Press.

MINAMI, F. 1974. *Gendai Nihongo no Kōzō*. Taishukan. (In Japanese).

MIYAGAWA, S. 1988. Predication and numeral quantifier. In *Papers from the Second International Workshop on Japanese Syntax*, ed. by W. Poser, 157–92. Stanford: CSLI.

—— 1989a. Light verbs and the ergative hypothesis. *Linguistic Inquiry* 20.659–68.

—— 1989b. *Structure and Case Marking in Japanese*. San Diego: Academic Press.

MURAKI, M., 1970. *Presupposition, Pseudo-clefting, and Thematization*. University of Texas, Austin dissertation.

NAKAIWA, H., and S. IKEHARA. 1994. Japanese zero pronoun resolution in a machine translation system using verbal semantic attributes. In *Proc. of SPICIS '94*, B341–6.

——, and ——. 1995. Intrasentential resolution of Japanese zero pronouns in a machine translation system using semantic and pragmatic constraints. In *Proc. of the 6th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-95)*, 95–105.

——, and ——. 1996. Anaphora resolution of Japanese zero pronouns with deictic reference. In *Proc. of the 16th International Conference on Computational Linguistics (COLING '96)*, 812–7.

——, and ——. 1997. A system of verbal semantic attributes in Japanese focused on syntactic correspondence between Japanese and English. *Journal of the Information Processing Society of Japan* 38.215–25. (In Japanese).

——, S. SHIRAI, S. IKEHARA, and T. KAWAOKA. 1995. Extrasentential resolution of Japanese zero pronouns using semantic and pragmatic constraints. In *Proc. of the AAAI Spring Symposium Series, Empirical Methods in Discourse Interpretation and Generation*, 99–105.

——, A. YOKOO, and S. IKEHARA. 1994. A system of verbal semantic attributes focused on the syntactic correspondence between Japanese and English. In *Proc. of the 15th International Conference on Computational Linguistics (COLING '94)*, 672–8.

NAKAU, M. 1971. *Sentential Complementation in Japanese*. Tokyo: Kaitakusha.

NEUMANN, C., 1994. La construction relative en franconien et en français et les deux fonctions de la relativation. Master's thesis, Université Michel de Montaigne Bordeaux III.

NISHIDA, F., S. TAKAMATSU, and H. KUROKI. 1980. A proper treatment of syntax and semantics in machine translation. In *Proc. of the 8th International Conference on Computational Linguistics (COLING '80)*, 447–54.

Nomura, N., and K. Muraki. 1996. An empirical architecture for verb categorization frame. In *Proc. of the 16th International Conference on Computational Linguistics (COLING '96)*, 640–5.

Ohno, S., and T. Shibata. 1977. *On'in*. Iwanami Shoten. (In Japanese).

Okumura, M., and K. Tamura. 1996. Zero pronoun resolution in Japanese discourse based on centering theory. In *Proc. of the 16th International Conference on Computational Linguistics (COLING '96)*, 871–6.

Okutsu, K. 1974. *Seisei Nihongo Bunpo Ron: Meisi-ku no kozo*. Tokyo: Taishukan.

Perlmutter, D. M. 1980. Relational grammar. In *Current Approaches to Syntax*, ed. by E. A. Moravcsik and J. R. Wirth, volume XIII of *Syntax and Semantics*, 195–229. New York: Academic Press.

Platt, J. T. 1971. *Grammatical form and grammatical meaning: a tagmemic view of Fillmore's Deep Structure Case concepts*. Amsterdam: North-Holland.

Pustejovsky, J. 1995. *The Generative Lexicon*. MIT Press.

Radford, A. 1981. *Transformational Syntax: A Student's Guide to Chomsky's Extended Standard Theory*. CUP.

Saito, M., 1985. *Some Asymmetries in Japanese and Their Theoretical Consequences*. MIT dissertation.

Sato, R., 1989. *Nihongo no Rentai-shūshoku-setsu no Imi-kaiseki ni-kansuru Kenkyū*. Master's thesis, Tokyo Institute of Technology. (In Japanese).

Shibatani, M. 1978. *Nihongo no Bunseki*. Taishukan. (In Japanese).

—— 1990. *The Languages of Japan*. CUP.

——, and S.A. Thompson (eds.) 1996. *Grammatical Constructions – Their Form and Meaning*. Clarendon Press: Oxford.

Shirai, S., S. Yokoo, H. Inoue, H. Nakaiwa, S. Ikehara, and A. Yagi. 1997. Nichi-ei kikai-hon'yaku ni-okeru imi-kaiseki no tame no kōbun jisho [A structural dictionary for semantic analysis in Japanese-English machine translation]. In *Proc. of the Third Annual Meeting of the Japanese Association for Natural Language Processing*, 153–6. (In Japanese).

Smith, H. 1996. *Restrictiveness in Case Theory*. CUP.

Somers, H. L. 1984. On the validity of the complement-adjunct distinction in Valency grammar. *Linguistics* 22.507–30.

—— 1987. *Valency and Case in Computational Lingustics*. Edinburgh University Press.

Starosta, S. 1988. *The case for lexicase*. London: Pinter.

Talmy, L. 1996. The windowing of attention in language. In (Shibatani and Thompson 1996), chapter 10, 235–87.

Tateishi, K. 1994. *The Syntax of 'Subjects'*. CSLI.

Teramura, H. 1970. The syntax of noun modification in Japanese. *The Journal of the Association of Teacher of Japanese* VI.23–55.

—— 1975–78. Rentai-shushoku no shintakusu to imi Nos. 1–4. In *Nihongo Nihonbunka 4–7*, 71–119, 29–78, 1–35, 1–24. Osaka: Osaka Gaikokugo Daigaku. (In Japanese).

—— 1978, 1981. *Nihongo no bunpo 1 & 2: Nihongo kyoiku shido sankosko 4 & 5*. Tokyo: Research Institute of the National Language. (In Japanese).

—— 1980. Meishi shushoku-bu no hikaku. In *Nichieigo Hikaku Koza: Bunpo 2*, ed. by T. Kunihiro, 221–60. Tokyo: Taishukan. (In Japanese).

Tsujimura, N. 1996. *An Introduction to Japanese Linguistics*. Blackwell.

Uszkoreit, H. 1987. Linear precedence in discontinuous constituents: Complex fronting in German. In *Discourse Continuity*, ed. by G. J. Huck and A. E. Ojeda, Syntax and Semantics 20, 405–20. Orlando: Academic Press.

Van Valin, R. 1984. A typology of syntactic relations in clause linkage. In *Proc. of the Tenth Annual Meeting of the Berkeley Linguistics Society*, 542–58.

Vance, T.J. 1987. *An Introduction to Japanese Phonology*. New York: SUNY Press.

Yoshimoto, K. 1986. Study of Japanese zero pronouns in discourse processing. In *Information Processing Society of Japan SIG Notes*, volume 86, no. 52. (In Japanese).