A Grammar for Annotating Syntax, Semantics and Pragmatics of Written and Spoken Japanese for NLP Application Purposes

Melanie Siegel

1 Introduction 1
2 Basic Japanese Phrase Structure 4
2.1 Head-subject phrase 5
2.2 Head-complement Phrase 5
2.3 Head-adjunct Phrase 7
2.4 Coordinated structures 9
2.5 Head-specifier constructions 11
2.6 Head-marker constructions 13
3 The Treatment of Subcategorization 15
3.1 A type system of arguments 19
3.2 Argument scrambling 19
3.3 Zero pronominalization 20
3.4 The status of subject arguments 20
3.5 Generalizations 21
4 Verbal Types, Verbal Inflection and Verbal Structures 22
4.1 Verbal subcategorization types 22
4.1.1 Intransitive verbs 23
4.1.2 Transitive verbs 24
4.1.3 Copula verbs 26
4.1.4 Verbal noun subcategorization 27
4.2 Verbal inflectional types 28
4.3 Auxiliary constructions 34
4.4 The treatment of passive 37
4.5 Causative 41
5 Nominal Structures: Linking Syntax, Semantics and Pragmatics 45
5.1 Ordinary nouns 45
5.2 Pronouns 46
5.2.1 Personal pronouns 48
5.2.2 Locative pronouns 49
5.2.3 Demonstrative pronouns 49
5.2.4 The reflexive *jibun* 49
5.3 Named entities 54
5.4 Nominalizations 55
5.4.1 Data analysis of nominalizations 58

5.5 Temporal expressions ... 63
5.6 Numeral classifiers ... 65
5.6.1 Data: Distribution ... 68
5.6.2 Semantic representations ... 68
5.6.3 The analysis ... 70
5.6.4 Lexical types ... 70
5.6.5 The linker no ... 76
5.6.6 Examples: NumClPs as modifiers ... 77
5.6.7 Unary-branching phrase structure rule ... 78
5.6.8 Examples: NumClPs as nouns ... 79
5.7 Noun modification ... 79
5.8 Relative sentence constructions ... 81
5.9 Pre-nominal adjectives ... 83
6 Particles ... 85
6.1 Co-occurrence of particles ... 86
6.2 The type hierarchy of Japanese particles ... 88
6.3 Case particles ... 89
6.3.1 The case particle *ga* ... 96
6.3.2 The case particle *wo* ... 98
6.3.3 The case particle *ni* ... 101
6.3.4 Other case particles ... 102
6.4 Particles with semantic content ... 102
6.4.1 Complementizers ... 103
6.4.2 Genitive specifying *no* ... 104
6.4.3 Modifying particles ... 105
6.4.3.1 The noun modifying particle *no* ... 105
6.4.3.2 Verb modifying particles ... 108
6.4.3.3 Particles of topicalization ... 113
6.4.4 Noun phrase conjunctions ... 120
6.5 Omitted particles ... 123
6.6 Evaluation of case and modifying particles ... 125
6.7 Sentence particles ... 125
7 Adverbs ... 127
8 Head-Initial Constructions in a Head-Final Language ... 130
8.1 The position of syntactic heads in Japanese ... 130
8.2 Head-initial modification ... 132

8.2.1 Data ... 132
8.2.1.1 Dake ... 132
8.2.1.2 Bakari `only' ... 134
8.2.1.3 *Bakari* and other forms meaning `about' ... 134
8.2.1.4 Numeral classifiers ... 136
8.2.2 Summary ... 136
8.2.3 Analysis ... 136
8.3 Head-initial complementation ... 137
8.3.1 Data ... 137
8.3.2 Analysis ... 140
9 Honorification ... 142
9.1.1 Honorific forms in Japanese ... 143
9.1.2 Interaction of different kinds of honorification in Japanese ... 144
9.1.3 Previous approaches ... 145
9.1.4 Japanese honorification in HPSG ... 146
9.1.5 Effects ... 151
9.1.6 Evaluation ... 152
9.1.7 Honorification in other languages ... 153
10 JACY in Different Application Domains ... 154
10.1 Appointment scheduling in machine translation ... 154
10.2 Emails in the banking domain ... 158
10.3 Parallel multilingual grammar development embedded in hybrid language processing ... 159
10.4 Dictionary definition sentences ... 164
11 Evaluations ... 165
11.1 What is the grammar size? How many rules and lexical entries does it contain? ... 167
11.2 What is the general coverage on what kinds of data? ... 168
11.3 How far is the grammar flexible and useful for applications? Is it domain-adaptable and can be used for different application domains? ... 170
11.4 How far can the grammar be used in multilingual applications? ... 173
11.5 Is the output precise and does it correspond to semantic format and content restrictions? ... 173
12 Conclusion ... 175
References ... 178
Appendix A: Grammar Installation ... 185
Using JACY with itsdb ... 186

Using JACY with PET 186
Using JACY with itsdb and PET 187
Appendix B: Author Index 188

1 Introduction

Natural language processing technology has reached a point where applications that rely on deep linguistic processing are becoming feasible. Such applications (e.g. message extraction systems, machine translation, email categorization and dialogue understanding systems) require natural language understanding, or at least an approximation thereof and the annotation of large amounts of language data. This, in turn, requires rich and highly precise information as the output of an NLP analysis. However, if the technology is to meet the demands of real-world applications, this must not come at the cost of robustness. Robustness requires not only wide coverage by the grammar (in both syntax and semantics), but also large and extensible lexica as well as interfaces to pre-processing systems for named entity recognition, non-linguistic structures such as addresses, etc. Furthermore, applications built on deep NLP technology should be extensible to multiple languages. This requires flexible yet well-defined output structures that can be adapted to grammars of many different languages. Finally, for use in real-world applications, NLP systems meeting the above desiderata must also be efficient.

In this text, we describe the development of a broad coverage grammar for Japanese that has been built for and used in different application contexts. The grammar is based on work done in the *Verbmobil* project (Siegel 2000) on machine translation of spoken dialogues in the domain of travel planning. The second application for JACY was the automatic email response task. Grammar development was described in Oepen et al. (2002a). Third, it was applied to the task of understanding material on mobile phones available on the internet, while embedded in the project DeepThought (Callmeier et al. 2004, Uszkoreit et al. 2004). Currently, it is being used for treebanking and ontology extraction from dictionary definition sentences by the Japanese company NTT (Bond et al. 2004).

The grammar is couched in the theoretical framework of Head-Driven Phrase Structure Grammar (HPSG) (Pollard & Sag 1994, Sag et al. 2003), with semantic representations in Minimal Recursion Semantics (MRS) (Copestake et al. 2001). HPSG is well suited to the task of multilingual development of broad coverage grammars: It is flexible enough (analyses can be shared across languages but also tailored as necessary), and has a rich theoretical literature from which to draw analyzes and inspiration. The characteristic type hierarchy of HPSG also facilitates the development of grammars that are easy to extend. MRS is a flat semantic formalism that works well with typed feature structures and is flexible in that it provides structures that are under-specified for scopal information. These structures give compact representations of ambiguities that are often irrelevant to the task at hand. HPSG and MRS have the further advantage that there are practical and useful open-source tools for writing, testing, and efficiently processing grammars written in these formalisms. The tools we are using include the LKB system (Copestake 2002) for grammar development, [incr tsdb()] (Oepen & Carroll 2000) for testing the grammar and tracking changes, and PET (Callmeier 2000), a very efficient HPSG parser, for processing. We also use the ChaSen tokenizer and POS tagger (Asahara & Matsumoto 2000).

While couched within the same general framework (HPSG), our approach differs from that of Kanayama et al. (2000). The work described there achieves impressive coverage (83.7% on the EDR corpus of newspaper text) with an underspecified grammar consisting of a small number of lexical entries, lexical types associated with parts of speech, and six underspecified grammar rules. In contrast, our grammar is much larger in terms of the number of lexical entries, the number of grammar rules, and the constraints on both, and takes correspondingly more effort to bring up to that level of coverage. The higher level of detail allows us to output

precise semantic representations as well as to use syntactic, semantic and lexical information to reduce ambiguity and rank parses.

Other existing HPSG grammars of Japanese, such as Yoshimoto (1997), Miyata et al. (2001) and Hashimoto and Bond (2005) give very detailed descriptions and implementations of aspects of the Japanese syntax and semantics. Insights from these gave important inspirations for the implementation of our broad-coverage grammar.

For a general introduction and deeper understanding of the HPSG formalism, the reader should consider reading Pollard and Sag (1994) and Sag et al. (2003), while we give a very short overview.

The fundamental notion of an HPSG is the sign. A sign is a complex feature structure representing information of different linguistic levels of a phrase or lexical item. Therefore, it is well suited to represent syntactic, semantic and pragmatic information and the interrelations.

The information is organized in feature-value pairs and is underspecified in a way that only those features that are relevant are expressed. For example, a nominal sign would not contain features denoting tense. The feature structures allow generalizations, as for example all verbs have the general head type **verbal_head**. Sorts, complex feature structures or lists are allowed as values. Grammar rules therefore contain complex information on their nodes, expressed in feature structures.

Most information in HPSG is lexical information and therefore provided with the lexical entries. The grammar implementation is based on a system of types. The lexical types define the syntactic, semantic and pragmatic properties of the words, such that a very important focus is set on the organization of the lexical type hierarchy. The information in a lexicon type therefore contains its head type, its valence type, the type of semantic construction and morphologic information. A lexicon entry contains the lexical type, the orthography and the lexical semantic information, as can be seen in Figure 1.

```
hon-noun := ordinary-nohon-n-lex &          ← lexical type
    [SYNSEM.LKEYS.KEYREL.PRED               ← lexical-semantic information
 '_hon_n_rel,                               ← orthography
      ORTH <! 本 !> ].
```

Figure 1: Lexicon entry for the noun *hon* (book)

The type hierarchy describes the existing signs in a grammar and organizes these in classes with shared peculiarities. Therefore, it describes the general attributes of the signs in these classes. HPSG grammars do not only contain lexical types, but also types that define the properties of phrases and lexical rules for inflectional and derivational morphology. The type hierarchy therefore allows the grammar writer to denote attributes and restrictions for a whole class of signs, leads to generalization and bewares from redundancies and errors. We will see various examples of parts of the type hierarchy in the following, as this is the centre of our grammar. It is not possible to display the whole type hierarchy of JACY on a sheet of paper, as it contains over 2,000 types.

Another important term in HPSG grammar description is *unification*. Phrase structure rules provide some information about how the information in the daughters has to be combined. The daughters of these provide more information. All of this is combined by the process of unification. Unification is a combination of information in attribute-value pairs that ensures the compatibility of the combined information. The operation regards type hierarchies in the compatibility check.

Phrase structure rules in HPSG grammars are restricted to the necessary constructions, as most information is contained in lexical types. We will see a description and motivation of the JACY grammar rules in Chapter 2.

The attribute-value matrix of a sign in the Japanese HPSG is quite similar to a sign in the LinGO English Resource Grammar (henceforth ERG) (Flickinger 2000), with information about the orthographical realization of the lexical sign in PHON, syntactic and semantic information in SYNSEM, information about the lexical status in LEX, nonlocal information in NONLOC, head information that goes up the tree in HEAD and information about subcategorization in SUBCAT.

We assume basic familiarity with the MRS semantic formalism as well and refer to Copestake et al. (2003) and Copestake et al. (2001). The HPSG grammar Matrix contains a documentation that gives a distinct introduction to the practical implementation aspects of MRS, also published as Flickinger et al. (2003).

MRS was designed as a semantic formalism to meet the demands of expressive adequacy, as well as computational tractability. It allows underspecification and is useful (and used) in NLP applications, such as Machine Translation and Information Extraction.

An MRS representation contains three basic components:

1. A bag of predications in RELS, each with a handle, a predication and one or more roles.
2. A set of handle constraints on scoping in HCONS.
3. A group of externally visible attributes in HOOK.

The handle in the predications in RELS is used to express scope relations. For example, modifiers share their handles with the predication they modify. The predication value PRED contains the lexical semantic information, which is a string in most cases and therefore assigned in the lexicon (as for example **_hon_n_rel** in Figure 1), but can be a sort (and thus organized in the type hierarchy) in the case of relations introduced by the grammar or being general, as for example **def_rel** for definite determination. Naming conventions require the lexical relations to be of the structure **_orth_pos_sense_rel** and the grammatical relations to follow the structure **sense_rel**. Part-of-Speech tags are defined in a limited set, and their application to Japanese can be seen in Table 1.

Table 1: MRS pos tags and the application to Japanese

MRS POS tag	Description	Japanese Example
v	verb	食べる (taberu - eat)
n	noun	本 (hon - book)
p	preposition	から (kara - from)
s	verbal noun	勉強 (benkyou - study)
a	adjective, adverb	元気 (genki - healthy)
q	quantifier	この (kono - this)
x	idioms, interjections	今日は (konnichiwa - Good day)
c	conjunction	か (ka - or)

Arguments are assigned by ARG0, ARG1, … ARGn (for the description of Japanese, we need arguments up to ARG3). ARG0 contains the index of the sign itself, in the case of verbs this is of the general type **event**. The value of RELS on the mother of a phrase is the result of

appending the RELS of its daughters. Further, a phrase structure rule can in some cases add its own relations.

The scopal relations in HCONS are relations of type **qeq** ('equality modulo quantifiers'). Each **qeq** relation contains an attribute HARG and an attribute LARG, both taking handles as values. HARG assigns the handle-taking argument position and LARG the outscoped relation. The value of HCONS of a mother is as well the result of appending the HCONS of its daughters.

The HOOK contains the information on a (composed) sign that is externally visible, i.e. accessible for semantic composition. The INDEX is necessary to bind arguments, LTOP contains the link to the top handle of a sign (for scoping restrictions) and XARG contains an externally visible argument of the sign, mostly the first argument in verbal constructions. The XARG is the locus of control. The value of HOOK of a mother in a phrase structure is identified with the value of its semantic head daughter.

In the next chapter, we start with the basic Japanese phrase structure and give an overview of how this relates to the basic phrase structure schemata in HPSG theory. This immediately leads to the problem of subcategorization in Japanese, which needs distinct methods for the treatment of optionality and scrambling in Chapter two. Verbal constructions, verbal valence, morphology, auxiliary constructions, passive and causative and the organization of the verbal type hierarchy are the topics of Chapter three. The description of nominal constructions and the nominal type hierarchy in Chapter five shows the interrelation between the information on different linguistic levels. Particles play a central role in syntactic and semantic information, and are described in Chapter six. To finalize the overview from a lexical type hierarchy viewpoint, we describe adverbs in Chapter seven. Chapter eight shows that in treating real-world data, we can find constructions that surprisingly contradict theoretical assumptions in literature. Honorification (Chapter nine) links the information of syntax, semantics and pragmatics and asks for an extension of the formalism. Chapter ten describes the application of the grammar to different application domains. Chapter eleven sets up questions that are relevant for the evaluation of this kind of grammars and gives answers concerning JACY.

2 Basic Japanese Phrase Structure

Pollard and Sag (1994) describe six basic schemata used in HPSG grammars:

1. Head-Subject Schema
2. Head-Complement Schema
3. Head-Subject-Complement Schema
4. Head-Marker Schema
5. Head-Adjunct Schema
6. Head-Filler Schema

Three of these (Head-Subject Schema, Head-Complement Schema and Head-Adjunct Schema) can also be found in JACY. The Head-Filler Schema in Pollard and Sag (1994) concerns empty structures and the usage of SLASH, for which we found different solutions. The Head-Subject-Complement Schema is replaced by a mechanism for scrambling, which is significant for parsing a language like Japanese.

This chapter will show how the schemata are used and modified for parsing Japanese, the instances of these general rule schemata, and necessary additions. Further, it will introduce a schema for coordinated structures, head-specifier constructions and verbal noun and light verb constructions (which we will call head-marker constructions).

2.1 Head-subject phrase

Pollard and Sag (1994:402) describe the head-subject-phrase as follows:

"The SYNSEM | LOCAL | CATEGORY | SUBCAT value is < >, and the DAUGHTERS value is an object of sort *head-comp-struc* whose HEAD-DAUGHTER value is a phrase whose SYNSEM | NONLOCAL | TO-BIND | SLASH value is { }, and whose COMPLEMENT-DAUGHTERS value is a list of length one."

Figure 2 shows the basic head-subject phrase in JACY[1].

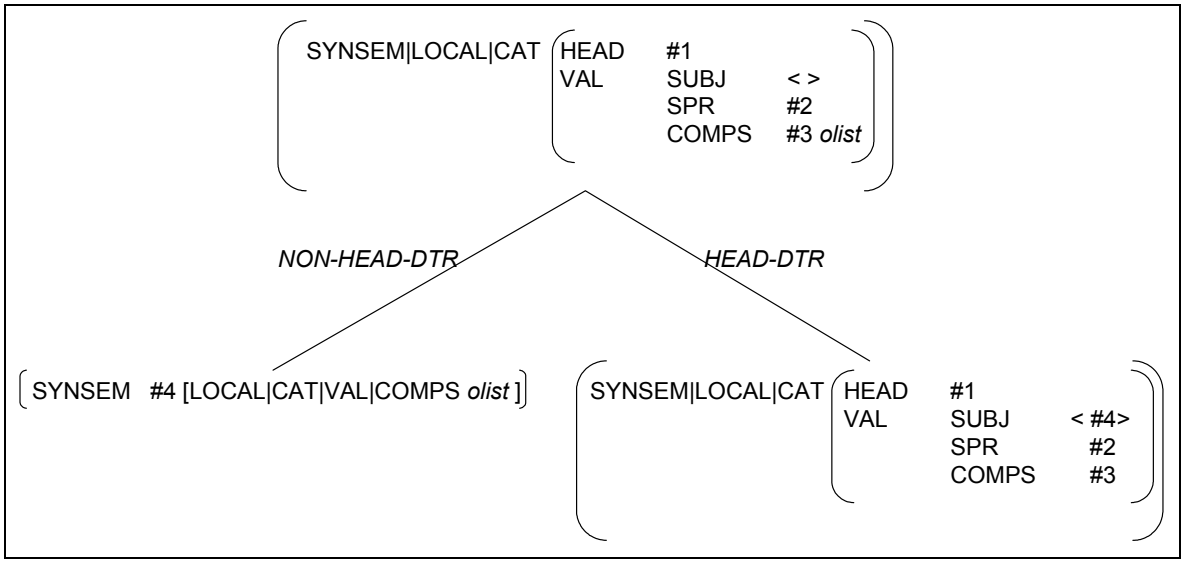


Figure 2: Head-subject type in JACY

This JACY phrase structure differs from Pollard/Sag's in prominent places:

- The subcategorization value (VAL in JACY) is not a single list, but a complex structure containing three different kinds of lists.
- The complement list is not necessarily empty when binding the subject.
- SLASH is not used in the Japanese structure.

These differences illustrate the fact that the basic treatment of subcategorization in JACY differs from the treatment of subcategorization in Pollard/Sag (1994). We explain and motivate these differences in Chapter 3: "The Treatment of Subcategorization".

2.2 Head-complement Phrase

This is the definition of the head-complement schema in Pollard/Sag (1994: 402):

"The SYNSEM | LOCAL | CATEGORY | SUBCAT value is a list of length one, and the DAUGHTERS value is an object of sort *head-comp-struc* whose HEAD-DAUGHTER value is a word."

The formalism allows the organization of rule schemata in types, in a similar way as lexical types. Further, instances of the rule types represent the rules themselves. The head-complement-types can be seen in Figure 4.

[1] Co-references are indicated with "#" and a number.

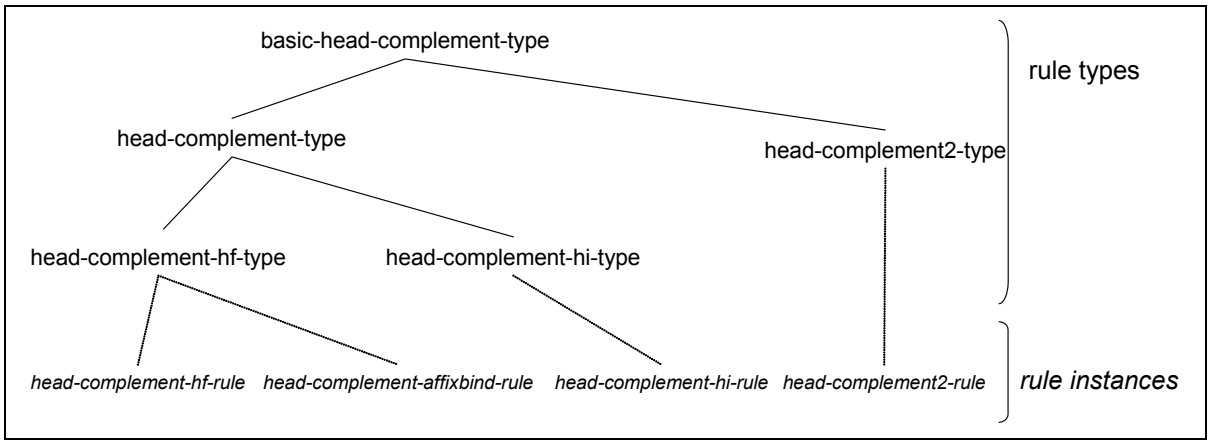


Figure 3: Types and instances of the basic head complement type

JACY's basic head-complement rule type does not constrain its Head-Daughter value to be a word, as the notion of word is in principle problematic in Japanese language processing.

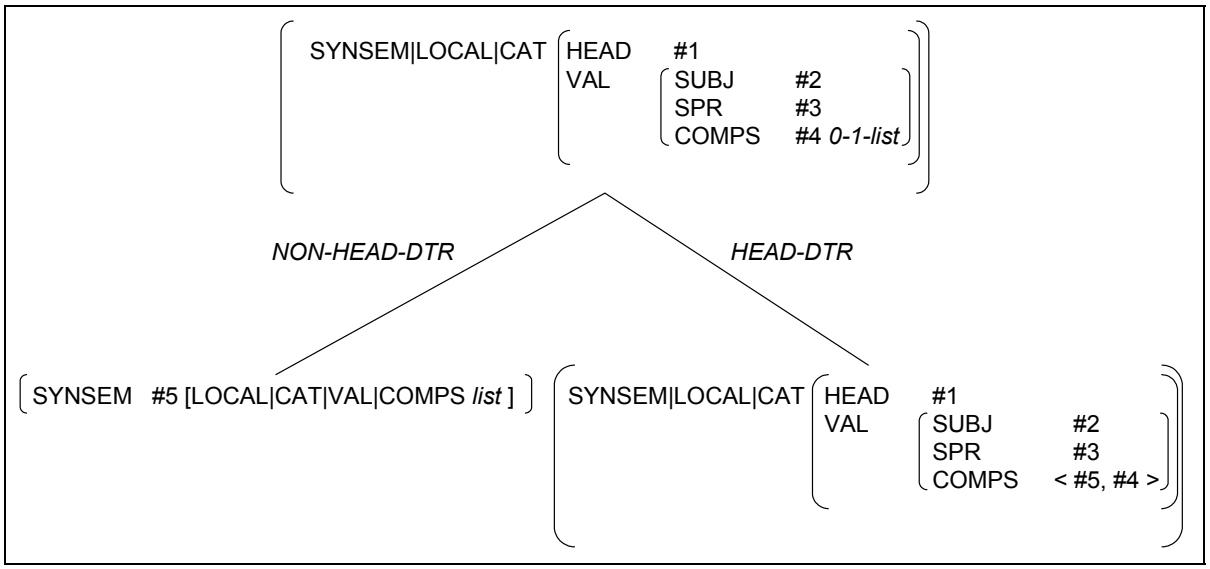


Figure 4: Head-complement type in JACY

Two rule types inherit from the basic **head-complement-type**: **head-complement-hf-type** and **head-complement-hi-type** for head-final and head-initial head-complementation (for motivation of the usage of two rules see Chapter 8: "Head-Initial Constructions in a Head-Final Language").

The **head-complement-affixbind-rule** inherits from the **head-complement-hf-type** and binds affixes required by words like *doko* (where) in *doko…mo* (wherever) constructions.

For the treatment of sentences with ditransitives, there is the additional **head-complement2-type**. This is needed for the treatment of argument scrambling, to ensure the possibility to bind the second argument on the COMPS list first (see Chapter 2: "The Treatment of Subcategorization")

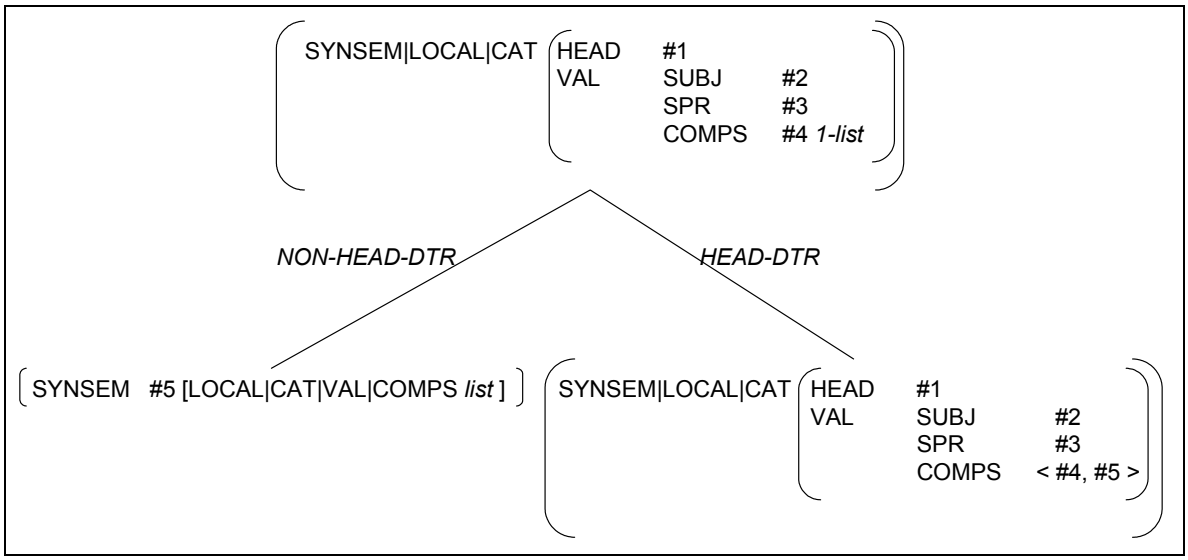


Figure 5: Head-complement2 type

2.3 Head-adjunct Phrase

Pollard/Sag (1994: 403) define the Head-Adjunct-Schema as follows:

"The DAUGHTERS value is an object of sort *head-adjunct-struc* whose HEAD-DAUGHTER | SYNSEM value is token-identical to its ADJUNCT-DAUGHTER | SYNSEM | LOCAL | CATEGORY | HEAD | MOD value and whose HEAD-DAUGHTERS | SYNSEM | NONLOCAL | TO-BIND | SLASH value is { }."

The JACY **head-adjunct-rule-type** identifies the LOCAL | CAT of the head daughter with the one of the single argument in the SYNSEM| LOCAL | CAT | HEAD | MOD list. As the SLASH mechanism is not used for scrambling in JACY, the COMPS list in VAL is restricted to *olist*, a list of optional arguments. This ensures that no adjacent arguments are allowed to be on the valence list, when adjuncts are found, such that for example no particle modification can intervene between a noun and its case particle.

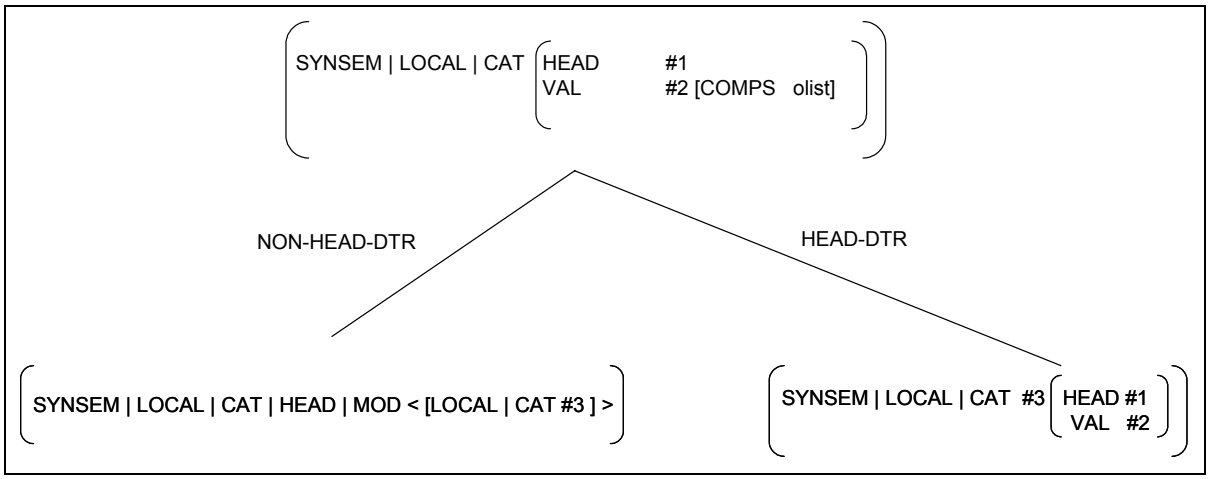


Figure 6: Head-adjunct-rule-type

As for the head-complement-types, there exist subtypes and instances to the **head-adjunct-rule-type** for head-initial and head-final adjunct structures (see Figure 8). Intersective and scopal modification need their own subtypes to enable correct semantic structures. These are cross-classified with head-final and head-initial adjunct structures. The subtypes make reference to a HEAD feature of their arguments, which we call POSTHEAD. It is firstly divided for relative and non-relative structures. Non-relative structures can be head-initial (right), head-final (left), or coordinative structures as well as compounds (see Figure 7).

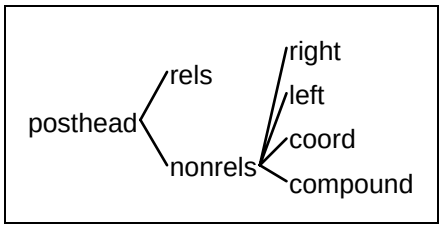


Figure 7: Posthead type hierarchy

An i-adjective has the POSTHEAD value **rels**, such that it can undergo relative sentence constructions. Pre-nominal modifiers like *purasu* (plus) have the POSTHEAD value *left*, such that they can undergo head-initial modification, while posthead noun modifiers like *nado* (and so on) have the POSTHEAD value *right*. The verbal *te*-form adds POSTHEAD *coord* to the verb's HEAD, such that it can undergo sentence coordination rules. Nominal compounds are constructed by the **compounds-rule**, which is an instance of the rule type for head-adjunct rules with head-final intersective modification (**hadj-final-i**). The compounds-rule adds a *compound-rel* to the MRS using C-CONT. The information in C-CONT is the semantic information that is added by rules, as defined by the Grammar Matrix. The POSTHEAD value of nouns, which is accessed by this rule, is *compound*.

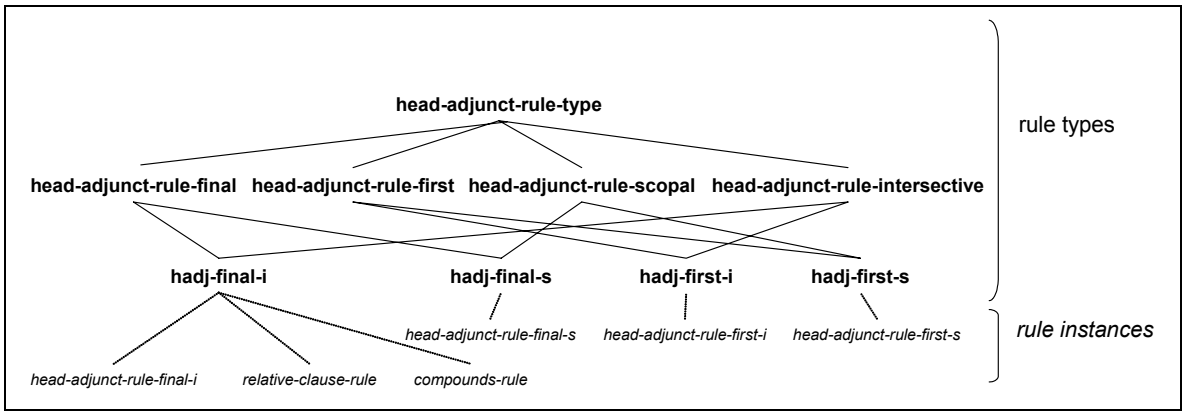


Figure 8: The head-adjunct rule type

2.4 Coordinated structures

Kurohashi and Nagao (1994) identify three types of coordination structures in Japanese:

1. Conjunctive noun phrases. Noun phrases that can include adjectival or clausal modifiers can be conjoined (see Example 1 and Example 2).
2. Conjunctive predicative clauses. Predicative clauses can be conjoined (see Example 3 and Example 4).
3. Incomplete conjunctive predicative clauses. Conjunction of verbal arguments (see Example 5).

Example 1: From Kurohashi/Nagao (1994)

```
gen-gengo            no   kaiseki   to    aite-gengo
source language text GEN  analysis CONJ  target language text
no   seisei      wo
GEN  generation  ACC
```

(The Analysis of the source language text and the generation of the target language text)

Example 2: From Kurohashi/Nagao (1994)

```
gen-gengo                  no          kaiseki-suru      shori       to
source language text       GEN         analyzing         processing  CONJ

aite-gengo           seisei-suru       shori          wo
target language text generation        processing     ACC
```

(The processing of analyzing the source language text and generating the target language text)

Example 3: From Kurohashi/Nagao (1994)

```
gen-gengo            no   kaiseki-shi,  aite-gengo
source language text GEN  analyzing     target language text


no   seisei-suru
GEN  generating
```

(Analyzing the source language text, generating the target language text)

Example 4: From Kurohashi/Nagao (1994)

kaiseki	dewa	riyou-suru	ga,	seisei	dewa	riyou-shinai
analysis	for	use	but	generation	for	do-not-use

(Use for analysis, but do not use for generation)

Example 5: From Kurohashi/Nagao (1994)

zensha	wo	kaiseki	ni,	kousha	wo	seisei	ni
the former	ACC	analysis	for	the latter	ACC	generation	for

(The former for analysis, the latter for generation)

We give analyses for the first and the second type of conjoined structures, but not for the third one, which needs further investigation.

Coordinative structures are handled by binary tree structures, as all our structures are binary (see Figure 9). The **binary-type-conj** inherits from the general type for (binary) modification **binary-modification-type**. The conjunction rule type (**conj-rule-type**) with its rule instance conj-rule makes use of a Head feature that we introduced to control coordination: C-MOD. C-MOD takes a list value of a list with no or one item on it. Coordinative inflections or particles get the information about the type they combine with in C-MOD. Therefore, the conjunction rule can access this information and unify CAT, CONT, BAR, NUCL, LEX and NON-LOCAL in C-MOD of the conjunction with the next conjunct. A conjunction therefore determines the type of the left conjunct in its valence (it is its complement) and the type of the right conjunct in C-MOD. The conjunction rule takes Head and Valence of the right conjunct and Index and LTOP of the first conjunct, combines the CONTEXT information and restricts both conjuncts to be saturated.

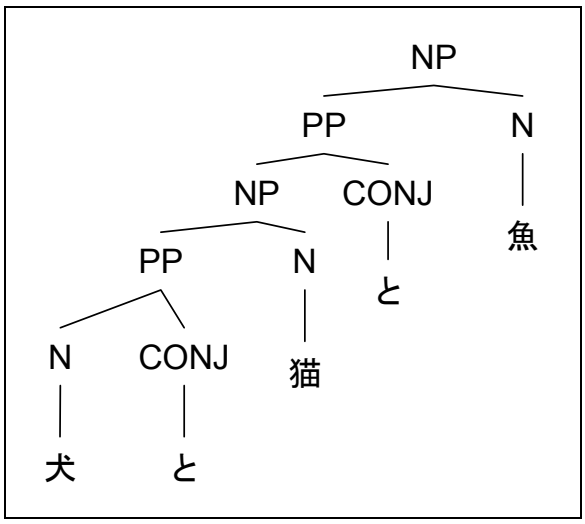


Figure 9

The same structure is used for conjunctive predicative phrases that contain a predicative conjunction like *ga*, *keredomo* or *node*.

Furthermore, we have implemented a type of sentence coordination without predicative conjunctions, containing verbs in te-form, such as in:

Example 6

花子	が	ご飯	を	食べて、	早く	寝た
Hanako	ga	gohan	wo	tabete,	hayaku	neta
Hanako	NOM	rice	ACC	eat,	quickly	slept

(Hanako ate rice and went to bed quickly)

For this kind of coordinative structures, we used the feature C-MOD in HEAD as well. This enables us to state possible coordinations on words or inflections that allow them. The *te* inflection of verbs is a good example.

The **sentence-te-coordination-rule** inherits from the **sentence-coord-type**. It states that the value of C-MOD of the non-head daughter contains one element whose value of LOCAL | CAT is identical with the value of SYNSEM | LOCAL | CAT of the head-daughter. The VAL is of type *saturated*, such that only sentences can undergo this rule.

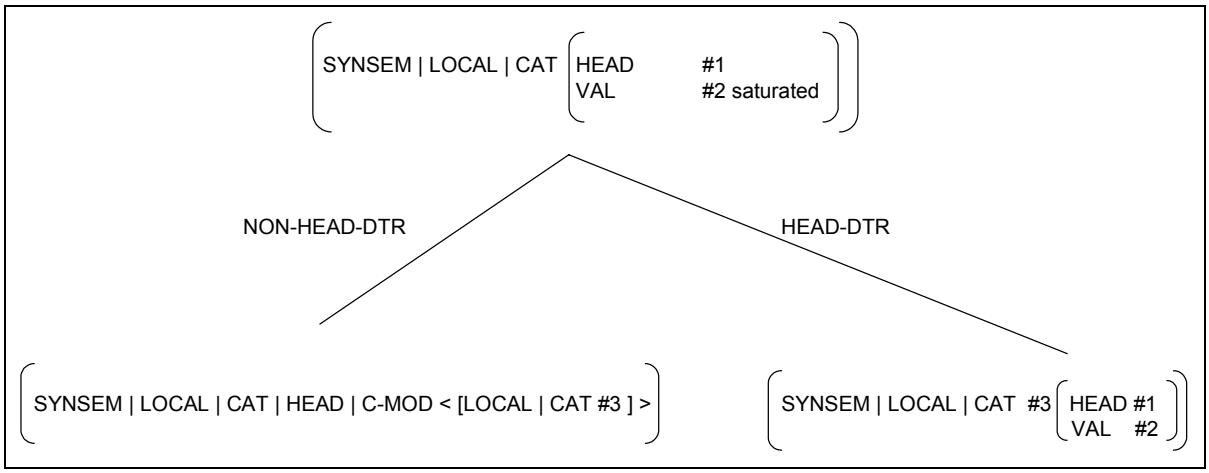


Figure 10: Sentence coordination

Another rule was needed for parsing real-world data, where sometimes two sentences occur without sentence segmentation or coordinations: The **runon_s** in JACY. The **runon_s** rule inherits from the **sentence-coord-type** as well, but is not restricted to *te*-form verbs. It is marked to be a robust rule, for robust parsing of real-world text.

In the case of more than two conjuncts, we keep the general binary construction policy and semantically let the second conjunction refer to the C-ARG of the first conjunct in its L-INDEX.

2.5 Head-specifier constructions

Some constructions in Japanese require the combination of head and valence information of the daughters in a way that is different from the one in head-complement schemas. Head-specifier constructions are introduced to HPSG to account for determiner-noun combinations (see for example Sag et al. 2003). We use them for these as well, but also for a couple of constructions that combine the information on the daughters in a similar way.

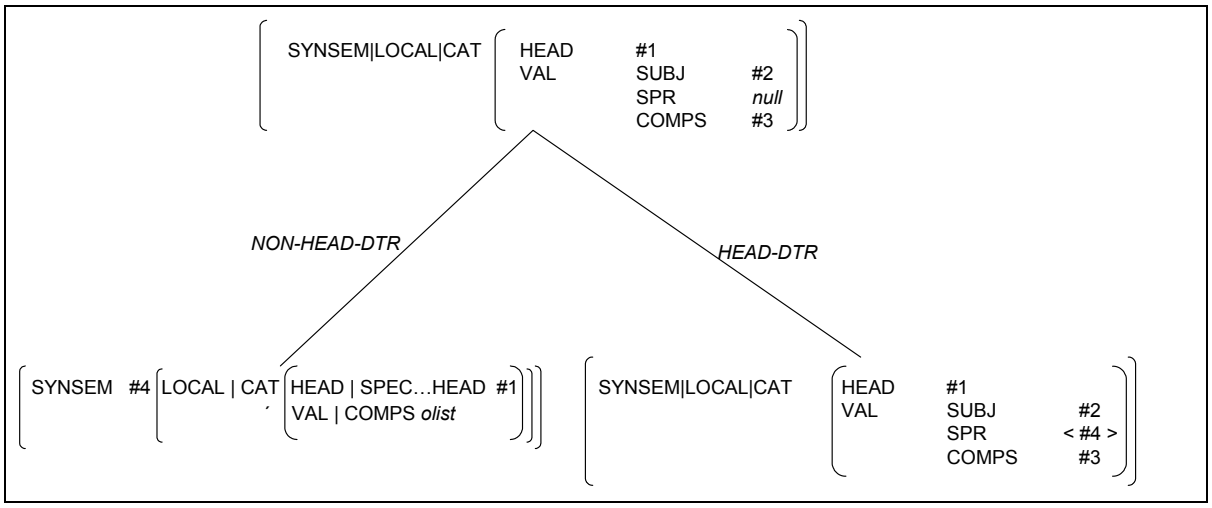


Figure 11: head-specifier-rule-type

Figure 11 shows the basics of the **head-specifier rule**. The head-daughter subcategorizes for a specifier, whose SYNSEM value is the non-head daughter's SYNSEM value. Therefore, the VAL.SPR on the mother is saturated, i.e. null. The HEAD information on the mother comes from the head-dtr. The subject and comps lists come from the head-daughter as well. The COMPS list on the nonhead-daughter must be a list of optional arguments, to assure that no adjacent arguments are left there. The non-head daughter selects for the head of the head-daughter via SPEC.

In a determiner-noun construction like Example 7, the determiner specifies a **noun_head** in HEAD.SPEC and the noun subcategorizes for an optional determiner.

Example 7

その　人
sono　hito
that　person

Surnames as well as some nouns specify for a title, while a title subcategorizes for a noun. Therefore, both Example 8 and Example 9 can be accounted for, but we don't get ordinary nouns with titles.

Example 8

田中　　さん
Tanaka　san
Tanaka　Ms/Mr

Example 9

学生　　さん
gakusei　san
student　Ms/Mr

The **head-specifier-rule** is used by nominalizing constructions as well. A predicative nominalization subcategorizes for a verb, while the verbal endings on the other hand determine the SPEC behaviour of the verb. A negative ending for example states that it specifies for a noun. This can be a regular noun, as in Example 10 or a nominalization, as in Example 11. The same is valid for the plain *ru* ending or the *tai* ending (*want to*), but not for polite endings like *masu*, as in Example 12.

Example 10

ご飯　を　食べられない　人
gohan　wo　taberarenai　hito
rice　ACC　cannot eat　person
(The person that cannot eat rice)

Example 11

ご飯　を　食べられない　こと
gohan　wo　taberarenai　koto
rice　ACC　cannot eat　fact
(The fact that someone cannot eat rice)

Example 12

*ご飯　を　食べます　こと
gohan　wo　tabemasu　koto
rice　ACC　eat (hon)　fact
(The fact that someone eats rice)

Adjectives and nominalizers combine using the same rule.

Japanese auxiliaries combine with verbs and provide either aspectual or perspective information or information about honorification. In a verb-auxiliary construction, the information about subcategorization is a combination of the VALENCE information of verb and auxiliary, depending on the type of auxiliary. The rule responsible for the information combination in these cases is the **head-specifier-rule** as well.[2]

Verbal endings are separated and therefore are attached with a binary rule. As the attachment of verbal stems and verbal endings requires a special treatment of argument combination as well, the **head-specifier-rule** type is responsible here as well. In JACY, the responsible rule instance is called **vstem-vend**. Verbal endings add various information about (addressee) honorification, tense, mood, etc., while the argument structure of the stem-ending complex comes from the stem. The ending subcategorizes for the stem, being its specifier.[3]

2.6 Head-marker constructions

A special treatment is needed for Japanese verbal noun + light verb constructions. In these cases, a word that combines the qualities of a noun with those of a verb occurs in a construction with a verb that has only marginal semantic information. The syntactic, semantic and pragmatic information on the complex is a combination of the information of the two. The verbal noun does not inflect. However, it subcategorizes, can be intransitive, transitive or ditransitive and gives sortal restrictions for its arguments. It is adjacent and obligatory to the light verb[4]. The predicate is formed by the complex.

Consider Example 13.

[2] For auxiliary constructions see section 4.3: “Auxiliary Constructions”.

[3] For verbal constructions see chapter 4.

[4] Dubinsky (1997) explains the atypical behaviour of verbal nouns.

Example 13

花子　　が　勉強　　した
Hanako　ga　benkyou　shita
Hanako　NOM　study　　light verb

(Hanako studied)

The verbal noun *benkyou* contains subcategorization information (transitive), as well as semantic information (the *benkyou*-relation and its semantic arguments). The light verb *shita* supplies tense information (*past*). Pragmatic information can be supplied by both parts of the construction, as in the formal form *o-benkyou shi-mashi-ta*. Research literature (e.g., Grimshaw and Mester 1988) talks about so-called "argument-transfer", where the arguments of the verbal noun are transferred to the light verb. Our analysis is based on the viewpoint that the verbal-noun – light verb complex is sub-syntactic, i.e., at the boarder of morphology and syntax. It needs a special rule that allows the combination of the information from both components. The rule that licenses this type of combination is the **vn-light-rule** (see Figure 12), an instance of the **head-marker-rule-type**. The **head-marker-rule-type** combines the HEAD information from the head daughter with the valence and semantic information from the non-head daughter under the viewpoint that the construction is sub-syntactic.

We use the specifier position for the verbal noun. The rule unifies the valence of the daughters and passes the head information to the mother. The non-head daughter selects the head daughter by the MARK feature in its head feature. The head daughter selects the non-head daughter by the SPR list in VAL. The verbal noun has sub-syntactic status, realized by the [BAR –] feature.

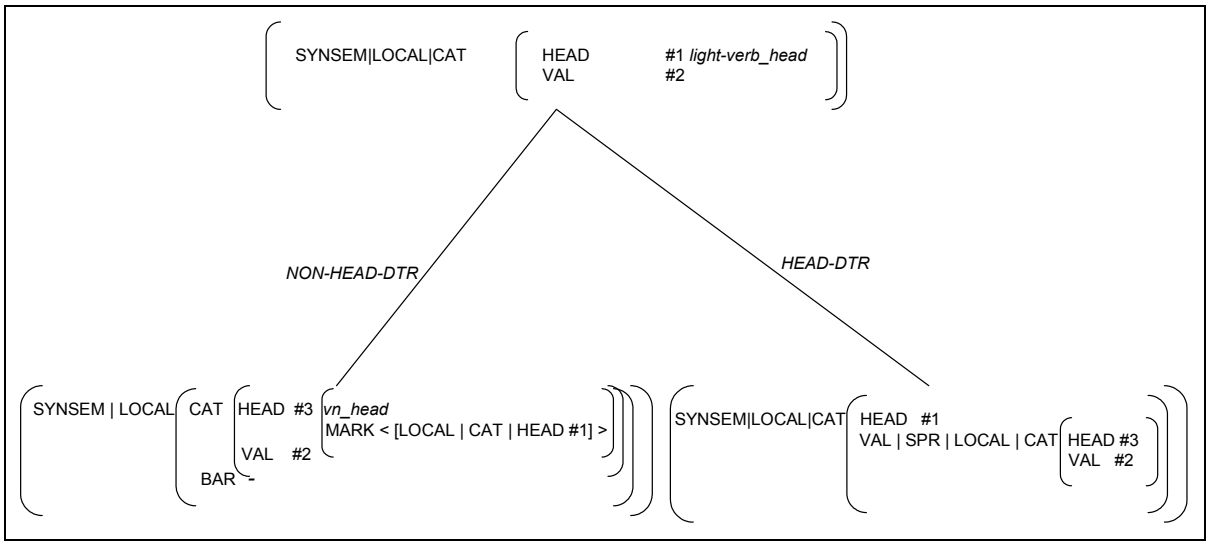


Figure 12: vn-light-rule

The **head-marker-rule-type** is used for the combination of verbal endings as well. This as well is a sub-syntactic construction, where the information is combined in an unusual way, just like the verbal-noun – light verb constructions.

3 The Treatment of Subcategorization

The subcategorizational peculiarities in Japanese differ from English subcategorization in the aspects of scrambling, zero pronominalization and peculiarities of the Japanese subject (never obligatory, restricting subject honorification and reflexive binding).

A fundamental difference between Japanese grammar and English or German grammar is the fact that verbal arguments can be optional and are omitted quite regularly. See Example 14, which are grammatical and (embedded in a proper context) understandable sentences in Japanese. Subjects and objects that refer to the speaker are omitted in most cases in spoken Japanese language. The predicate arguments can freely scramble as can be seen in Example 15.

Example 14

a) 花子　が　ご飯　を　食べた
Hanako ga gohan wo tabeta
Hanako NOM rice AKK eat-past
(Hanako ate rice)

b) ご飯　を　食べた
Gohan wo tabeta
rice AKK eat-past
(ate rice)

c) 食べた
Tabeta
eat-past
(ate)

Example 15

ご飯　を　花子　が　食べた
Gohan wo Hanako ga tabeta
rice AKK Hanako NOM eat-past
(Hanako ate rice)

On the other hand, there exist obligatory and adjacent verbal arguments, as the following example from the Verbmobil corpus and its ungrammatical reductions and scramblings show:

Example 16

a) 会議　は　二時間　ぐらい　です
Kaigi wa nijikan gurai desu
Meeting TOP two hours about COP
(The meeting is about two hours.)

b) *会議　　は　です
Kaigi　wa　desu
Meeting　TOP　COP

c) *です
desu
COP

d) *二時間　　ぐらい　会議　　は　です
Nijikan　gurai　kaigi　wa　desu
Two hours　about　meeting　TOP　COP

The Japanese subject has a special status, as it is the entity that is referred to by honorific agreement and reflexive binding (as will be shown in Chapter 9: “Honorification”).

It is not trivial to categorize arguments and adjuncts in Japanese, due to these facts of optionality and scrambling. Some hypothesis about the division:

1) Subjects are arguments.
    - They are always optional.
    - They are the goal of subject honorification.
    - They are nominative case.
2) Entities that are obligatory are always arguments.
3) Entities that are marked by *wo* (accusative) are arguments.
    - They can be optional or obligatory/adjacent.
4) Entities that can be passivized are arguments.
    - It has to be shown, whether this is valid in the other direction as well, such that things that cannot be passivized are adjuncts.
5) Things that get a semantic restriction from the head are arguments.

HPSG, as outlined in Pollard/Sag (1994), but also in Sag et al. (2003) is insufficient for the treatment of Japanese subcategorization. First, it cannot account for the division of adjacent/obligatory and optional arguments. Second, it cannot account for scrambling, because the lists are sorted and the head-argument-rules must be applicable in various orders for the treatment of Japanese. The SLASH mechanism does not help, either, because it entails the supposition that arguments are at least satisfied somewhere in the sentence, which is not true for all arguments in Japanese. It is moreover the fact that arguments that can scramble are also optional. Third, it does not account for the special role of subjects in Japanese in the areas of empathy and honorific agreement.

(Gunji 87) uses a set instead of a list of categories as the value of SUBCAT to account for argument scrambling. This approach did not provide the necessary mechanism for the distinction between optional and obligatory arguments and the special status of subjects.

(Gunji 91) adds the feature ADJACENT that contains a set of adjacent complements. The ordering in this list is irrelevant to the application of grammar rules. This idea does not divide between the status of adjacency and obliqueness of arguments, which is correct for Japanese, as Gunji describes. As we will show in Chapter 10: “JACY in Different Application

Domains", we have set the JACY grammar in a multilingual context, which requires a careful distinction; although in Japanese, adjacency and obliqueness correspond.

Sirai (1996) describes two attributes that account for subcategorization in the Japanese Phrase Structure Grammar. SUBCAT takes a set of local categories that are optional and scrambable arguments, while ADJACENT takes the category of one obligatory and not scrambable argument. The problem with having two lists of this kind is that generalizations about, e.g. object control, cannot be easily stated. Furthermore, the approach is not applicable to languages where arguments can be optional and adjacent and thus lacks multilingual generality.

Another idea of handling Japanese subcategorization (which was implemented in an earlier version of the grammar) was to use grammatical functions to clearly divide (and access) the verbal arguments. This approach could not be followed further, when setting the grammar in a multilingual context: while the naming was not obvious, the approach could not easily be applied to other languages.

We could build several lexicon entries for each lexeme, where each entry represents one possible argument structure. This was a strategy adopted in an earlier version of the German HPSG in Verbmobil (Müller/Kasper 2000). This approach has two disadvantages: the lexicon explodes, especially for a language like Japanese, and the approach lacks generality and modularity.

The idea to use a scrambling lexical rule that takes a lexicon entry and produces COMPS lists in various orders might be a possibility too, but bears the danger of explosion of grammar processing.

The Grammar Matrix (see Bender et al. 2002) contains a treatment of subcategorization that is partly drawn from the insights of Japanese HPSG construction. The idea is to stay with the HPSG COMPS list and add a SUBJ list. Constraints about optionality and adjacency are added. Types of possible argument structures are ordered in a type hierarchy, as can be seen in Figure 13.

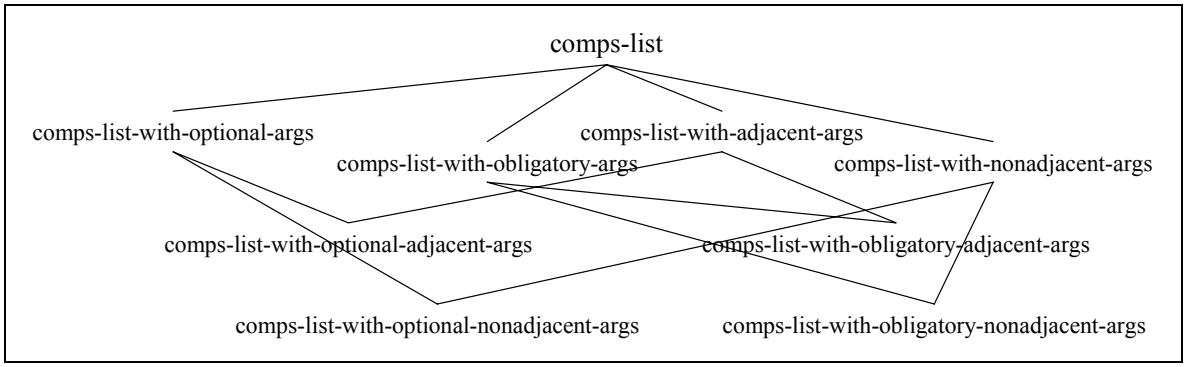


Figure 13: Type hierarchy of complement lists in Grammar Matrix

In order to give a direct encoding to the division of optional and obligatory arguments, as well as scrambling and adjacent arguments, the argument status is explicitly stated in an attribute OPT. This contains information about the saturation status of subcategorized arguments. It is an advantage of this approach that it provides a straightforward and easy-to-process way of dealing with scrambling and optionality of arguments. There are no lexical rules necessary that move arguments from valence to adjacency or slash lists, there is no need for traces and slash's. The grammar uses different head-complement structures that pick up the first, second or third argument of the COMPS list and are not ordered in their application.

The Matrix Grammar idea of subcategorization is applied to Japanese. We adopt the idea to divide between the notions of adjacency and optionality. Verbal subcategorization frames

contain the information about the status of arguments, as these notions are concerned. Upon this, we add a subject list to be able to access the subcategorized subject for the treatment of subject-related phenomena.

Arguments in subcat frames are lists in our approach, rather than sets (as in Gunji's proposal). The main reason for this is technical: The TDL formalism that underlies our grammar writing in the LKB system (Copestake 2001) does not allow the usage of sets.

Scrambling is resolved by using different head-complement structures that pick up the first, second or third argument of the COMPS list and are not ordered in their possible application.

We tested on the Japanese grammar that the treatment is still adequate for the phenomena associated with Japanese subcategorization.

The default values for optionality would be [OPT +] for Japanese, while it would be [OPT -] for English. The different types of possible argument structures can be ordered in a type hierarchy. We do not sort arguments into different lists (or sets), but rather note the status on the argument itself, because we would like to restrict a complement in a lexical type, underspecified whether it is adjacent or optional in the types or lexical entries that inheres from this type.

The value of the features SUBJ and COMPS therefore is a list of signs of type **synsem** with the additional feature OPT.

Adjacency must be checked in every rule that combines heads and arguments or adjuncts. This is stated in the principle of adjacency, formulated as follows:

- *In a headed phrase, the VALENCE of the non–head daughter must contain only arguments of the type* **comps-list-with-nonadjacent-arguments**.
- *In a head–complement structure, the VALENCE of the head daughter must contain only arguments of the type* **comps-list-with-nonadjacent-arguments** *besides the non–head daughter.*
- *In a head–adjunct structure, the VALENCE of the head daughter must contain only arguments of the type* **comps-list-with-nonadjacent-arguments**.

This principle is realized in all phrasal types. The **head-subject-phrase**, for example, requires its nonhead-daughter and its head-daughter to have **olist** as the value of COMPS. **Olist** is a list of optional arguments, as introduced by the Grammar Matrix (see Bender et al. 2002).

Figure 14: The valence of a typical transitive verb

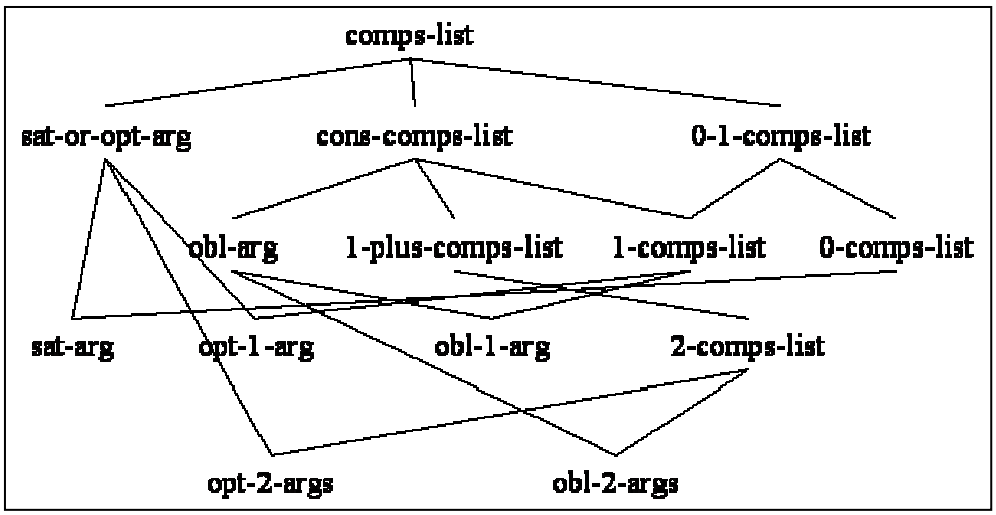


Figure 15: Type hierarchy of complement lists in JACY

The approach solves the basic problems of Japanese subcategorization:

1. The necessary distinction of argument types for Japanese.
2. The scrambling phenomena.
3. Zero pronominalization.
4. The special status of subjects.
5. The necessity to express generalizations.

3.1 A type system of arguments

The necessary distinction of argument types for Japanese, optional versus obligatory/adjacent, can be described in a type system. The verb *taberu - eat*, for example, contains an optional subject and an optional complement, which is expressed in the types of the SUBJECT and COMPS values (**opt-1-arg**). A light verb contains a SPR that is obligatory and therefore, the value of SPR is of type **obl-1-arg**. Figure 15 shows the type system of argument structures in JACY. Figure 14 shows the value of VALENCE of a typical transitive verb like *taberu* – eat. The values of SUBJ, COMPS and SPR are typed lists. Figure 16 shows the value of VALENCE of a case particle containing an adjacent and obligatory complement, which is again reflected in the argument type of the COMPS list.

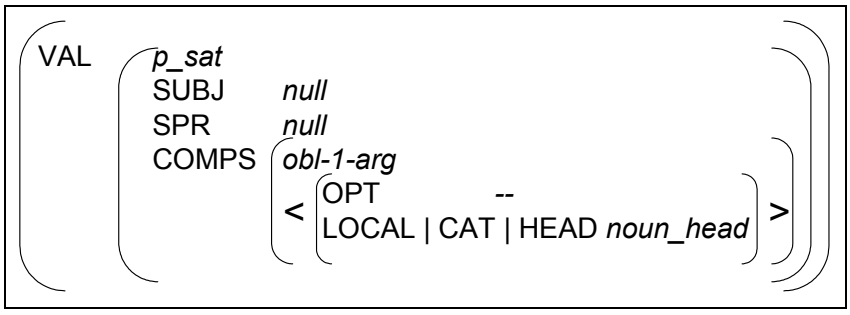


Figure 16: Valence of a case particle

3.2 Argument scrambling

The problem of scrambling of verbal arguments is solved by this approach. The **head-subject-rule** does not contain a restriction on the saturation status of the head, besides the restrictions stated by the principle of adjacency (COMPS values must be of type **olist**). There are two head-complement-rules to account for scrambling, such that the two possible complements

can be saturated in each order, just being checked for adjacent arguments (see Section 2.2 for details). Figure 17 shows the tree structure of sentences with scrambled and non-scrambled arguments.

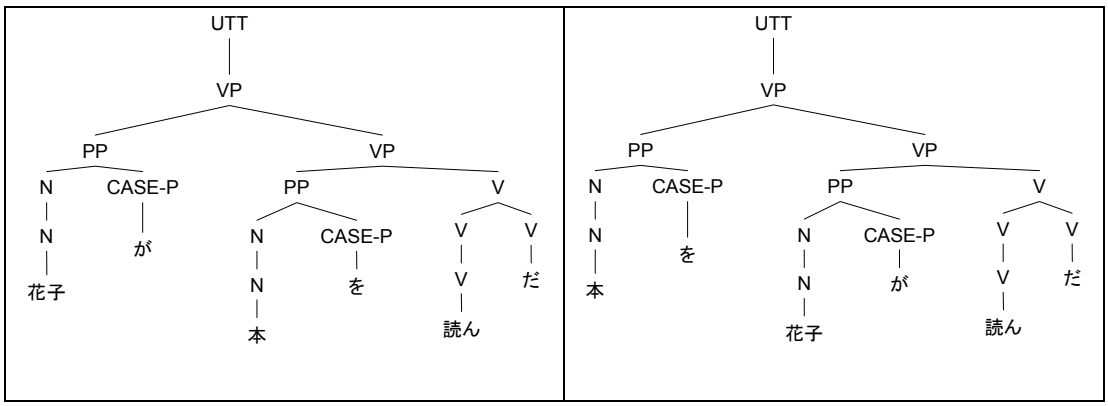


Figure 17: Tree structure and scrambling

3.3 Zero pronominalization

Japanese verbal arguments cannot only scramble, but also be omitted. This is very often the case with subjects, but also objects can be affected (examples are given above). The solution to this problem in the suggested approach is the introduction of a lexical rule that inserts semantic information to the argument structure of a verb and marks the argument in the valence structure as saturated, without inserting empty categories into the syntactic structure. The lexical rule can apply to arguments that are optional and are therefore annotated with [OPT +]. These insert semantic information to the MRS and empty the valence list. Figure 18 shows the lexical rule that applies to unexpressed subjects.

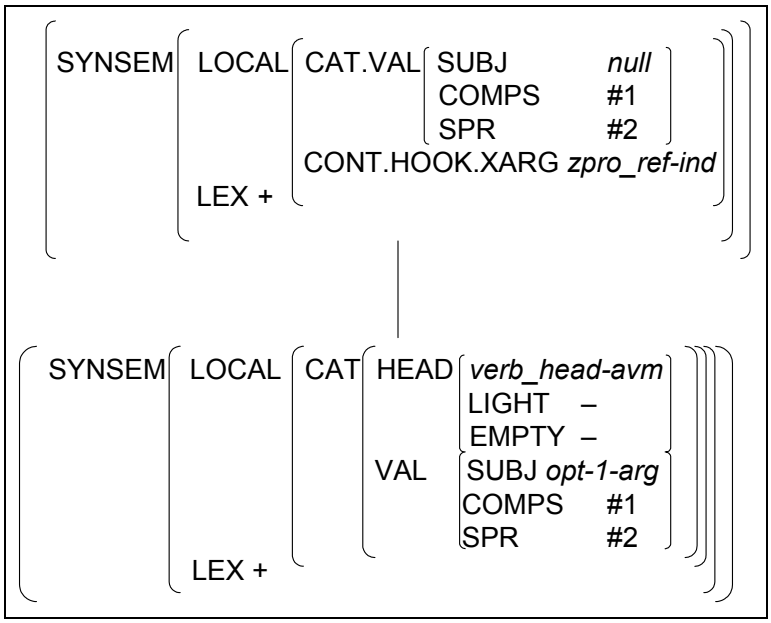


Figure 18: subject zero pronoun insertion rule

3.4 The status of subject arguments

Japanese has a rich set of possibilities to express social relations between speaker, addressee, subjects and objects of an utterance. The social relation between speaker and subject that is not referring to the speaker is expressed by the lexical choice of verbs, by the expression *o-VERB-ni-naru* , by the honorific prefix *o/go* at nouns referring to entities belonging to the

subject and by the lexical choice of pronouns. A relation of distance between speaker and subject (where the subject is the addressee or a third person) can be - for example - expressed by the verb *irassharu* (to go), while in a familiar situation the verb *iku* with the same semantic content is used. Possible referring expressions for the second and third person can be, for example, *sochira* and *X-san* in relations of distance and *kimi* or *X-kun* in relations of familiarity. A clear reference point to the sentence subject is therefore necessary, because subject honorification sets syntactic, semantic and pragmatic restrictions on the verbal subject (see Chapter 9: "Honorification"). Our subcategorization approach delivers this. The head-subject rule can easily test for honorification agreement restrictions between subject and verb and the lexical types for verbs and nouns can be restricted to honorification.

Reflexive binding is highly determined by the subject arguments of verbs as well (see Gunji 83 and section 5.2.4: "The reflexive jibun"). Our approach to subcategorization makes the formulation of empathy and honorification relating the subject argument possible by providing direct access to the subject.

3.5 Generalizations

There are generalizations on the argument status of verbal arguments that should be expressed. An example is verbal subjects, which are always optional in Japanese. This generalization is accounted for by the fact that the SUBJ list of all verbs contains the type **opt-1-arg**.

4 Verbal Types, Verbal Inflection and Verbal Structures

The lexical types combine (via cross-classification in a type hierarchy) peculiarities of words on different levels. Verbal types therefore contain information about morphology (stemtype), syntax (head), valence (subcategorization types), semantics (semtypes, linking of syntactic arguments to semantic ones) and pragmatics (honorification).

4.1 Verbal subcategorization types

There are intransitive (**subj-arg**), transitive (**subj-comps-arg**) and ditransitive subcategorization types. Derived from the subcategorization facts that are outlined above, we have identified a set of verbal stem types, classified by their subcategorization behaviour, as displayed in Table 2. Subjects are always optional in Japanese, and can therefore be omitted, which is stated for all subject arguments in the type **sbj-arg**. The table notes the part-of-speech (POS) and case of complements, and whether they are optional (**opt**) or obligatory (**obl**).

Table 2

Verbal type	**Subcategorization pattern**		**Example**
	SBJ	**COMPS**	
intrans-stem-lex	P-ga		太る
to_intrans-stem-lex		P-to (opt)	いう
v1-stem-lex	P-ga	P-wo (opt)	見守る
v2-stem-lex	P-ga	P-ni	乗る
v2a-stem-lex	P-ga	ADV (obl)	なる
v2b-stem-lex	P-ga	P-ni-or-to (obl)	なる
v3-stem-lex	P-ga	P-to (obl)	言う
v4-stem-lex	P-ga	P-wo (opt), P-ni (opt)	置く
v5-stem-lex	P-ga	P-to (opt)	付き合う
v5a-stem-lex5	P-ga	P-to (obl)	思い出す
v6-stem-lex6	P-ga	P-ni-or-to (opt)	入れ替わる
v8-stem-lex	P-ga	N (obl)	書き送る
cop-id-stem-lex	P-ga-or-coparg7	N (obl)	です

[5] **v5a** and **v3** are different in the semantic type of arguments they take: **v5a** takes a complement sentence and therefore semantically links its hook, while **v3** takes a nominal complement, which is linked via its index.

[6] **v6** differs from **v2b** semantically: **v2b** inserts a **cop-id_rel**.

[7] *coparg* is the case given to colons, when these are used in the same way as case particles and to the topic particle *wa* in copula constructions.

4.1.1 Intransitive verbs

Intransitive verbs of the type **intrans-stem-lex** take a subject (which is optional by definition) and no complements. The subject is headed by a particle, whose case is **ga**. An example for an intransitive verb is the verb *futoru* (to become fat), as in Example 17.

Example 17

猫　が 太 っ た
neko ga futotta
cat GA became fat

(the cat became fat)

The subject is typically linked to ARG1 of the verb in the MRS semantics.

```
h1,e2:PAST:INDICATIVE,
h1:proposition_m(e2, h3),
h4:_neko_n(x5:THREE),
h6:u(x5, h7, h8),
h9:_futoru_v(e2, x5),
h3 qeq h9,
h7 qeq h4
```

Figure 19: MRS for *neko ga futotta*

The Japanese verb *iu* in certain contexts takes only one complement, but no subject. This is a complement sentence, which is marked by the particle *to* in Japanese, as in Example 18.

Example 18

花子　が 元気　だ　と いう こと を 聞いた

Hanako ga genki da to iu koto wo kiita
Hanako GA healthy Copula TO IU NOM ACC heard

(I heard that Hanako is healthy)

The lexical type links the handle of the proposition of the complement sentence to ARG1 of the verb in the MRS semantics.

```
h1,e2:PAST:INDICATIVE,
h1:proposition_m(e2, h3),
h4:named(x5, "hanako"),
h6:(x5, h7, h8),
h9:_genki_a(e10:PRESENT, x5),
h11:proposition_m(e10, h12),
h13:_iu_v_3a(e14:INDICATIVE:PRESENT, h11),
h15:_koto_n(x16, h17),
h18:u(x16, h19, h20),
h17:proposition_m(e14, h21),
h22:_kiku_v(e2, u23, x16),
h3 qeq h22,
h7 qeq h4,
h12 qeq h9,
h19 qeq h15,
h21 qeq h13
```

Figure 20: MRS for *Hanako ga genki da to iu koto wo kiita*

4.1.2 Transitive verbs

The most frequent type of transitive verbs is those that take a subject, which is marked by *ga* and an optional complement, which is marked by *wo*. They are given the type **v1-stem-lex**. An example is the verb 食べ, as in Example 19.

Example 19

花子　　が　ご飯　　を　食べた
Hanako　ga　gohan　wo　tabeta
Hanako　GA　rice　WO　ate

(Hanako ate rice)

```
h1,e2:PAST:INDICATIVE,
h1:proposition_m(e2, h3),
h4:named(x5, "hanako"),
h6:(x5, h7, h8),
h9:_gohan_n(x10:THREE),
h11:u(x10, h12, h13),
h14:_taberu_v(e2, x5, x10),
h3 qeq h14,
h7 qeq h4,
h12 qeq h9
```

Figure 21: MRS for *Hanako ga gohan wo tabeta*.

The MRS semantics contains a linking of the subject to ARG1 and the complement to ARG2.

Another class of transitive verbs, *v2-stem-lex*, takes complements marked by *ni*. These are always optional, as the complement of *noru* in Example 20.

Example 20

花子　　が　（バス　に）　乗る
Hanako　ga　（basu　ni）　noru
Hanako　GA　bus　NI　ride

(Hanako rides on the bus)

They are linked to the semantics the same way as the complements in the **v1-stem-lex** type described above.

Verbs like *au* (meet) take a *ga* marked subject and a complement, which can be marked by *ni* or *to*. They are given the type **v6-stem-lex**:

Example 21

私　　が　花子　　に/と　会いました
watashi　ga　Hanako　ni/to　aimashita
I　　GA　Hanako　NI/TO　met

(I met Hanako)

The only verb in type **v2b-stem-lex**, *naru*, allows the complement to be marked by either *ni* or *to*. This complement is obligatory. Copula semantics is added to the MRS in this usage of *naru*.

Example 22

花子　　が　大人　　に/と　なった
Hanako　ga　otona　ni/to　natta
Hanako　GA　adult　NI　　became

(Hanako became adult)

naru, on the other hand, can take an adverbial (obligatory) complement, as in Example 23, where copula semantics is added as well. This gets the type **v2a-stem-lex**.

Example 23

辺り　　　が　明るく　　なった
atari　　ga　akaruku　natta
neighborhood　GA　bright　　became

(The neighborhood became bright)

Verbs in the class of **v3-stem-lex** require a *ga* marked subject and an obligatory sentence complement, which is marked by *to*. To this class belong verbs like *zonjiru* (know), *kataru* (relate) and *iu* (say).

Example 24

花子　　が　［良い］　と　言いました
Hanako　ga　“yoi”　　to　iimashita
Hanako　GA　“good”　TO　said

(Hanako said “good”)

The MRS semantics of these contains a link to the handle of the proposition of the complement sentence to ARG2 of the verb in the MRS semantics. ARG1 is linked to the subject.

```
h1,e2:PAST:INDICATIVE,
h1:proposition_m(e2, h3),
h4:named(x5, "hanako"),
h6:(x5, h7, h8),
h9:_ii_a(e11:PRESENT, u10),
h12:proposition_m(e11, h13),
h14:_iu_v(e2, x5, h12),
h3 qeq h14,
h7 qeq h4,
h13 qeq h9
```

Figure 22: MRS for *Hanako ga ‘yoi’ to iimashita*

Transitive verbs of type **v5-stem-lex** take a subject and an optional *to* marked sentence complement, such as the verb *omou* (think). Here as well, ARG1 of the verbal semantics is linked to the subject and ARG2 is linked to the handle of the complement sentence’s proposition.

Example 25

花子　　が　元気だ　　と　思います
Hanako　ga　genkida　to　omoimasu
Hanako　GA　healthy　TO　think

(I think that Hanako is healthy)

In the case of **v5a-stem-lex**, the *to*-marked sentence complement is obligatory. Although the subcategorization patterns of **v5a** and **v3** look very much alike, **v5a** and **v3** are different in the

semantic type of arguments they take: **v5a** takes a complement sentence and therefore semantically links its hook, while **v3** takes a nominal complement, which is linked via its index. *wakaru* (understand) is an example of these:

Example 26

この 映画 が 面白い と 分かった
kono eiga ga omoshiroi to wakatta
this film GA interesting TO understood

(I understood that this film is interesting)

4.1.3 Copula verbs

The Japanese copula verbs belong to the type **cop-id-stem-lex**. They take a subject and a complement. The subject is marked by *ga* or *wa*, the complement is a noun (without a case particle). In the MRS, the copula adds a **cop-id-rel**, with an ARG1 and an ARG2, linked to the subject and the complement, respectively.

Ordinary copula verbs do not modify, different to copula verbs like *dearu* or *degozaru*. This type contains words like *da, desu, nandesu* or *nanodesu*, as can be seen in Example 27.

Example 27

これ は はな です
kore wa hana desu
this TOP flour COP

(This is a flower)

```
h1,e2:PRESENT,
h1:proposition_m(e2, h3),
h4:generic-nom(x5:NEUT),
h6:dem(x5, h7, h8),
h9:_wa_p(e10:NO_TENSE, x5, e2),
h11:_hana_n(x12:THREE),
h13:u(x12, h14, h15),
h16:cop_id(e2, u17, x12),
h3 qeq h16,
h7 qeq h4,
h14 qeq h11
```

Figure 23: MRS for *kore wa hana desu*

Copula verbs can be conditionals, such as *nara*. These add conditional semantics with the relation **if_then_rel**. They can contain question marking, such as *kai* or *kanaa*, such that the sentence MRS contains a **question_m_rel**. *dai* requires its complement to contain a wh-word. Verbs like *dewanai* or *janai* contain a negation, which is added to the MRS as a **neg_relation** with an ARG1 pointing to a handle that outscopes the label of the **cop-id-rel**. The head of these behaves (syntactically) like a negative adjective, such that the type **cop-id-neg-stem-lex** is hooked higher in the verbal type hierarchy. Still, the valence is the typical copula valence and the MRS contains a **cop-id-rel**. Figure 24 shows the type structure of copula verbs and examples.

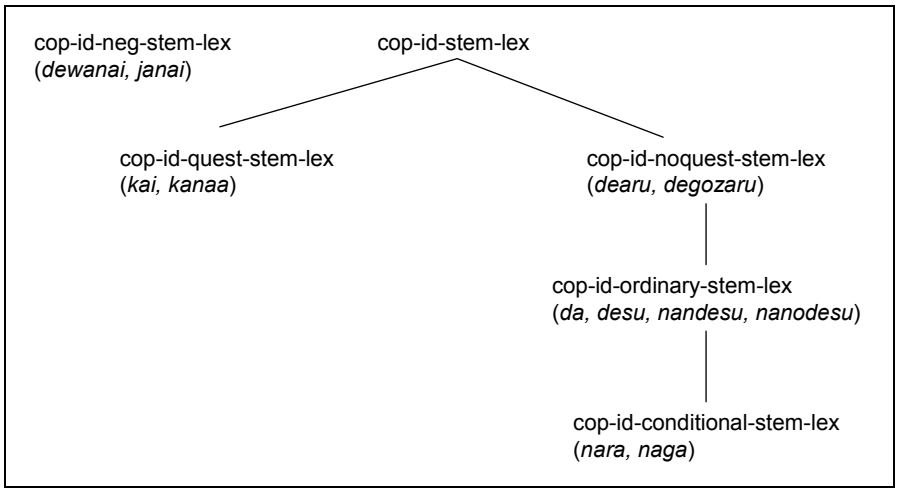


Figure 24: Copula verb types

```
h1,e2:PRESENT:MOOD,
h1:proposition_m(e2, h3),
h4:generic-nom(x5:NEUT),
h6:dem(x5, h7, h8),
h9:_wa_p(e10:NO_TENSE, x5, e2),
h11:_hana_n(x12:THREE),
h13:u(x12, h14, h15),
h16:neg(e2, h17),
h18:cop_id(e20, u19, x12),
h3 qeq h16,
h7 qeq h4,
h14 qeq h11,
h17 qeq h18
```

Figure 25: MRS for *kore wa hana dewanai*

There is another class of copula verbs that do not add any semantic relations. These occur together with so-called na-adjectives and add information about tense and honorification. They can be called light copula, as their function is similar to light verbs in the VN – light verb constructions (*cop-light-lex*). Example 28 shows the usage of the light copula *desu*.

Example 28

花子	は	元気	です
Hanako	wa	genki	desu
Hanako	TOP	healthy	COP-LIGHT

(Hanako is healthy)

4.1.4 Verbal noun subcategorization

Japanese verbal nouns basically follow the same subcategorization principles. Table 3 shows the subcategorization patterns of verbal nouns in JACY.

Table 3: Subcategorization patterns of verbal nouns

Verbal noun type	**Subcategorization pattern**		**Example**
	SBJ	**COMPS**	
vn-intrans-lex	P-ga		発生
vn-trans1-lex	P-ga	P-wo (opt)	アレンジ
vn-trans2-lex	P-ga	P-ni	電話

vn-trans3-lex	P-ga	P-to (opt)	結婚
vn-trans8-lex	P-ga	N (obl)	よろしくお願い
vn-ditrans-lex	P-ga	P-wo (opt), P-ni (opt)	掲載
vn-ditrans-toni-lex	P-ga	P-wo (opt), P-ni-or-to (opt)	略

4.2 Verbal inflectional types

Japanese verbs inflect in different ways due to their stem types and the verbal ending they combine with. Therefore, the stem types must be classified and assigned to verbal types. A classification of stem types can be seen in Figure 26.

Verbs combine with endings, which assign tense, addressee honorification, or mood, as can be seen in Example 29.

Example 29

a.	食べる	b.	食べた	c.	食べました	d.	食べられた
	taberu		tabeta		tabemashita		tabe rareta
	eat		ate		ate (hon)		was eaten

Verb stems inflect due to their inflectional type and combine with verbal endings due to their inflection. Figure 26 gives the type hierarchy of stem types.

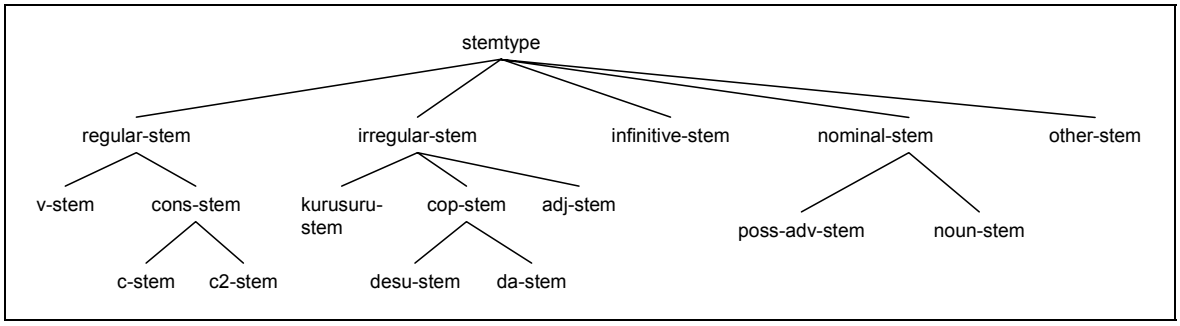


Figure 26: Type hierarchy of stemtypes

Verbs must be sorted into these types in order to be applicable for the correct inflectional rule. Table 4 shows the differences in inflection of the types for the combination with the past tense ending.

Table 4: Inflection types and combination with past tense

Stem	**Inflection when combined with past tense ending**	**Example**
v-stem	delete る	食べる → 食べ (*taberu – tabe*)
c-stem	く→い、る→っ、う→っ、す→し、つ→っ、む→ん、ぐ→い、ぶ→ん、	聞く→聞い (*kiku – kii*)

	ぬ→ん	
c2-stem	く→っ、る→っ、う→う	行く→行っ (*iku – ikk*)
kurusuru-stem	来る→来、する→し、くる→き	来る→来 (*kuru – ki*)
cop-stem[8]	だ→だっ、す→し	です→でし (*desu – deshi*)
adj-stem	い→かっ	高 い → 高 か っ (*takai – takakatt*)

Inflectional rules apply to verbal stems. They make morphologic changes and give the result a morphological type in the type hierarchy under **morphbindtype**.

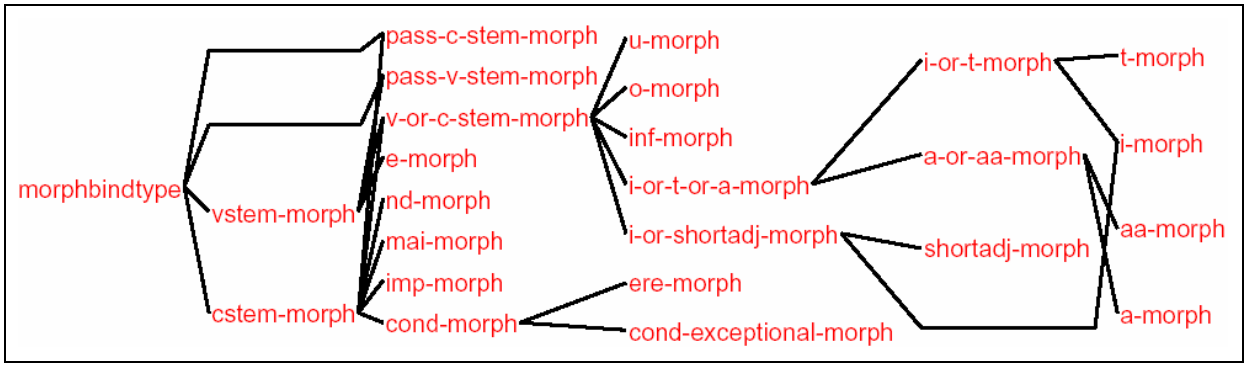


Figure 27: Type hierarchy of morphological binding types

First of all, there are rules that pipe stems to words, such that no verbal endings are needed. These are listed in the following table:

Name of inflection rule	**Change of morphological type**	**Example**
`ru-lexeme-infl-rule`	`(no change to the morphology)`	食べる (*taberu*)
`eru-lexeme-infl-rule`[9]	`(regular-stem -> u-morph)`	学べる (*manaberu*)
`infinitive-lexeme-1-infl-rule`	`(regular-stem -> inf-morph)`	食べ (*tabe*)
`imperative-c2-stem-infl-rule`	`(c2-stem -> imp-morph)`	下さい (*kudasai*)
`desu-lexeme-infl-rule`	`(cop-stem -> u-morph)`	です (*desu*)
`de-lexeme-infl-rule`	`(desu-stem -> u-morph)`	で (*de*)
`ra-lexeme-infl-rule`	`(da-stem -> u-morph)`	なら (*nara*)
`kuru-lexeme-infl-rule`	`(kurusuru-stem -> u-morph)`	来る (*kuru*)
`infinitive-lexeme-2-infl-rule`	`(kurusuru-stem -> inf-morph)`	来 (*ki*)
`adj-i-lexeme-infl-rule`	`(adj-stem -> u-morph)`	ではない

[8] **cop-stem** is the mother node of **desu-stem** and **da-stem** in the type hierarchy of stemtypes (as can be seen in Figure 26). The distinction is not needed for past tense marking, but for other inflections.

[9] This rule transforms to potential form.

		(*dewanai*)

The inflectional rules that apply to verbs, which then need verbal endings are the following:

Name of inflectional rule	**Change of morphological type**	**Attachement to ending**	**Example**
i-lexeme-c-stem-infl-rule	(c-stem -> i-morph)	ます、ました、まして、ません、ませんでした、はじめる、ましたら、ましたらば、ながら、ましょう、よう、たい、たがる、たく、られる、なさい	読みます (*yomimasu*)
i-lexeme-c2-stem-infl-rule	(c2-stem -> i-morph)	ます、ました、まして、ません、ませんでした、はじめる、ましたら、ましたらば、ながら、ましょう、よう、たい、たがる、たく、られる、なさい	行きません (*ikimasen*)
i-lexeme-v-stem-infl-rule	(v-stem -> vstem-morph[10])	ます、ました、まして、ません、ませんでした、はじめる、ましたら、ましたらば、ながら、ましょう、よう、たい、たがる、たく、られる、なさい, **た、たり、て、たら、たらば、てる、ちゃう, ありませんでした、ない、ありません、なさ、ぬ、ないで、ずに、なる、ざるをえません, なさ, う, させる、さす、られる**	食べます (*tabemasu*)
a-lexeme-negative-cons-stem-infl-rule	(cons-stem[11] -> a-or-aa-morph)	**ず、ありませんでした、ない、ありません、なさ、ぬ、ないで、ずに、なる、ざるをえません**	読まない (*yomanai*)
pass-lexeme-stem-infl-rule	(cons-stem -> pass-c-stem-morph)	せる、れる	読ませる (*yomaseru*)
t-lexeme-c-stem-infl-rule	(c-stem -> t-morph)	**た、たり、て、たら、たらば、てる、ちゃう**	聞いた (*kiita*)
tt-lexeme-c2-stem-infl-rule	(c2-stem -> t-morph)	**た、たり、て、たら、たらば、てる、ちゃう**	行った (*itta*)
cond-lexeme-infl-rule	(regular-stem -> cond-morph)	ば、る	読めば (*yomeba*)
cond-spoken-lexeme-infl-rule	(v-stem -> cond-	ば、る	食べれば (*tabereba*)

[10] **vstem-morph** is a supertype to a variety of morphological binding types, as can be seen in Figure 27: Type hierarchy of morphological binding types. The result of applying this inflectional rule is therefore useful to a variety of verbal endings.

[11] As can be seen in Figure 26: Type hierarchy of stemtypes, cons-stem is a supertype to **c-stem** and **c2-stem**. This inflectional rule can therefore be applied to both.

	exceptional-morph[12]		
o-lexeme-c-stem-infl-rule	(cons-stem -> o-morph)	う	読もう（*yomou*）
o-lexeme-v-stem-infl-rule	(v-stem -> o-morph)	う	食べよう（*tabeyou*）
nd-lexeme-infl-rule	(c-stem -> nd-morph)	だ、だり、で、だら、だらば、じゃう	読んだ（*yonda*）
tt-cop-lexeme-infl-rule	(cop-stem -> t-morph)	**た、たり、て、たら、たらば、てる、ちゃう**	だった（*datta*）
o-cop-lexeme-infl-rule	(cop-stem -> o-morph)	う	でしょう（*deshou*）
ki-lexeme-infl-rule	(kurusuru-stem -> i-morph)	ます、ました、まして、ません、 ませんでした、 はじめる、ましたら、ましたらば、ながら、ましょう、よう、たい、たがる、たく、られる、なさい	来ません（*kimasen*）
ka-lexeme-infl-rule	(kurusuru-stem -> a-morph)	**ない、なさ、ぬ、ないで、ずに、なる、ざるをえません**	来ない（*konai*）
kaa-lexeme-infl-rule	(kurusuru-stem -> aa-morph)	ず	来ず（*kozu*）
kit-lexeme-c-stem-infl-rule	(kurusuru-stem -> t-morph)	**た、たり、て、たら、たらば、てる、ちゃう**	来た（*kita*）
ke-lexeme-infl-rule	(kurusuru-stem -> cond-morph)	ば、る	来れば（*kureba*）
ko-lexeme-infl-rule	(kurusuru-stem -> o-morph)	う	来う（*kou*）
sa-lexeme-infl-rule	(kurusuru-stem -> pass-c-stem-morph)	せる、れる	来させる（*kosaseru*）
mai-lexeme-infl-rule	(kurusuru-stem -> mai-morph)	まい、だけ	来まい（*komai*）
adj-te-t-lexeme-c-stem-infl-rule	(adj-stem -> t-morph)	**て**	ではなくて（*dewanakute*）
adj-past-t-lexeme-c-stem-infl-rule	(adj-stem -> t-morph)	**た**	ではなかった（*dewanakatta*）
adj-kere-lexeme-infl-rule	(adj-stem -> cond-morph)	ば、る	ではなければ（*dewanakereba*）

[12] This is a rule that allows the so-called Ranuki form, which is used in spoken language.

The following is an example of information that an inflectional rule adds to the verb stem:

RMORPH-BIND-TYPE	*i-morph*
SYNSEM.LOCAL.CAT.HEAD.MODUS	*indicative*

Figure 28: Information added by an inflectional rule

The inflectional rule itself contains a restriction on the stemtype of the verbal stem it applies to, such as: ARGS.FIRST.STEMTYPE *c-stem.*

Derivational rules inflect verb stems and change their syntactic category. There are two derivational rules that apply to verbs (as can be seen in Table 5): The *v2vn-infl-rule* changes the syntactic category to a verbal noun, while the *v2n-infl-rule* changes the syntactic category to a noun by attaching *kata*.

Table 5: Derivational rules that apply to verb stems

Rule	**Derivation**	**Inflection**	**Example**
v2vn-infl-rule	regular-stem -> vn	く → き、 す → し (and others)	食べる → 食べ
v2n-infl-rule	regular-stem -> n	む → み方、 く → き方 (and others)	食べる → 食べ方

Figure 29 shows, how the information on a verbal stem changes when going through an inflectional rule and then combining with an ending. Stem, ending and the stem-ending complex are lexical, i.e., [LEX +].

This is the information a verbal stem contains:

- LEX +
- STEMTYPE
- J-NEEDS-AFFIX +
- INFLECTED –
- Valence information

The verbal stem contains a stem type, which is the main point of selection for the inflection rule. [J-NEEDS-AFFIX +] means that the stem needs an affix (an ending) to be a valid argument in a phrase structure. Going through inflectional rules, the information on the inflected stem is:

- LEX +
- **RMORPH-BIND-TYPE**
- J-NEEDS-AFFIX +
- **INFLECTED +**
- Valence information

The inflected verbal stems can now attach to verbal endings, such as *ta*, *masu*, or *mashite*. The inflectional rule changes the value of INFLECTED to '+' and adds an RMORPH-BIND-

TYPE, which is referred to by the verbal ending. Verbal endings are separated in ChaSen and are therefore attached with a binary rule (*vstem-vend*, an instance of *head-specifier*). They add various information about (addressee) honorification, tense, mood, etc. The argument structure of the stem-ending complex comes from the stem. The ending subcategorizes for the stem (SPR).

This is information on the verbal ending:

- BAR –
- LEX +
- LMORPH-BIND-TYPE
- J-NEEDS-AFFIX –
- Head information, such as honorification, tense, fin, cop-arg, modus

The verbal ending adds an LMORPH-BIND-TYPE, which is used for unification in the combining with the verbal stem. [BAR -] assures that the stem has to combine with an ending to be able to participate in a phrase structure rule. Head information is the main information the ending contributes in the stem-ending complex.

The verb stem – ending complex contains the following information:

- BAR +
- LEX +
- J-NEEDS-AFFIX –
- Head information from the ending.
- Valence information from the stem.

It can now be part of phrasal structures ([BAR +] and [J-NEEDS-AFFIX -] and contains a combination of the relevant information from its daughters.

The INDEX information on tense and mood is unified between the two parts of the verb. The semantic relations are combined; in the case of simple endings they come from the verbal stem alone.

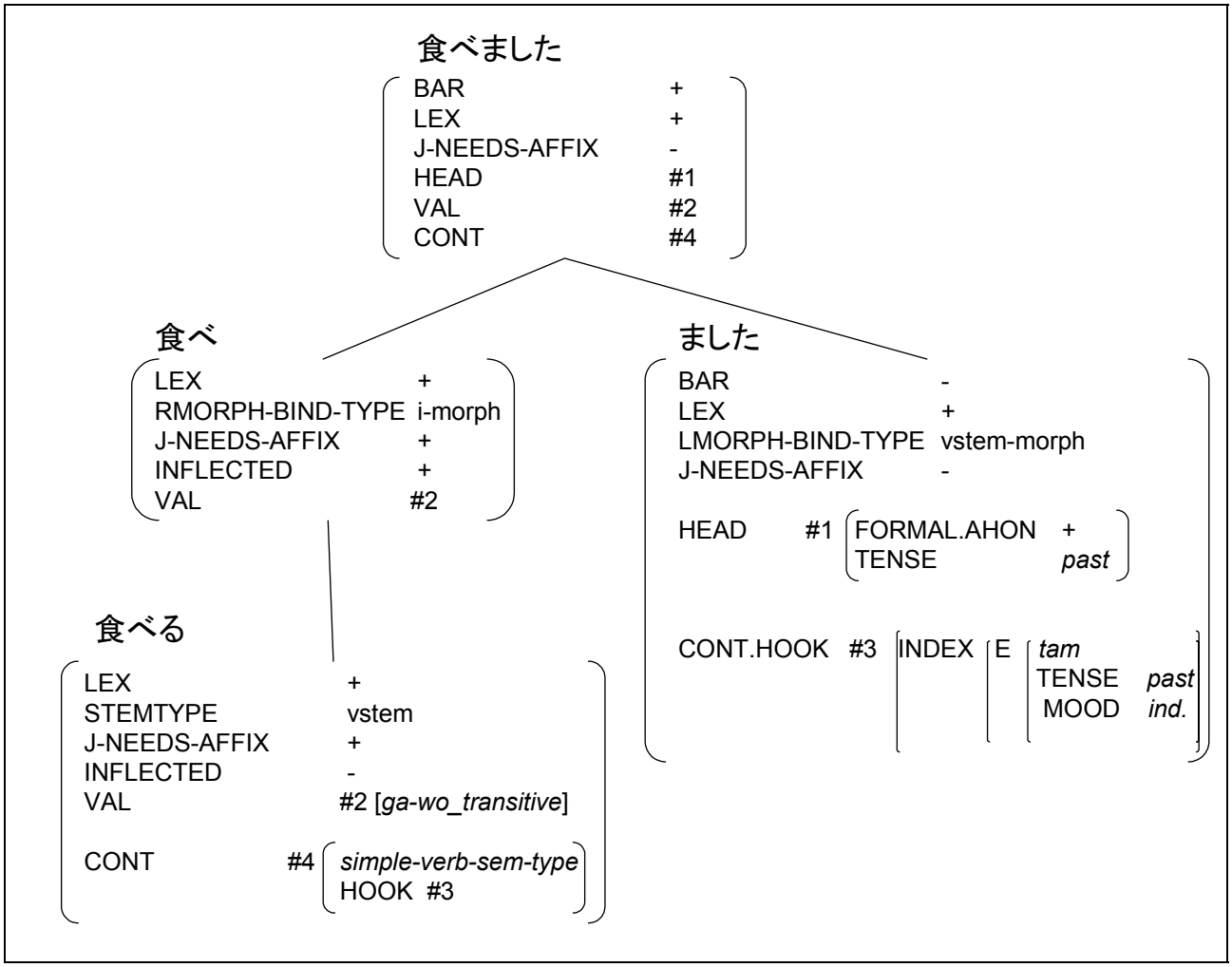


Figure 29: An example for verbal inflection and stem-end combination

4.3 Auxiliary constructions

Japanese auxiliaries combine with verbs and provide either aspectual or perspective information or information about honorification. In a verb-auxiliary construction, the information about subcategorization is a combination of the SUBCAT information of verb and auxiliary, depending on the type of auxiliary. The rule responsible for the information combination in these cases is the **head-specifier-rule**.

We have three basic types of auxiliaries. The first type is aspect auxiliaries. These are treated as raising verbs, and include such elements as *iru* (roughly, progressive) and *aru* (roughly, perfective), as can be seen in Example 30.

The other two classes of auxiliaries provide information about perspective or the point of view from which a situation is being described. Both classes of auxiliaries add a *ni* (dative) marked argument to the argument structure of the whole predicate. The classes differ in how they relate their arguments to the arguments of the verb. Verbs in one class (including *kureru* 'give'; see Example 33 are treated as subject control verbs. The other class (including *morau* 'receive', see Example 34) establishes a control relation between the *ni*-marked argument and the embedded subject.

Example 30: pure aspect

ケ-キ　を　食べて　いる
keeki　wo　tabete　iru
cake　ACC　eat　AUX (progressive)

(*Someone is eating cake.)*

Example 31: aspect[13]

ソ-ジ　が　して　ある
sooji　ga　shite　aru
cleaning　NOM　do　AUX (perfective))

(*The cleaning has been done.)*

Example 32: complex aspect

花子　が　ケ-キ　を　食べて　見る
Hanako　ga　keeki　wo　tabete　miru
Hanako　NOM　cake　ACC　eat　AUX (modal: try to)

(*Hanako tries to eat the cake..)*

Example 33: perspective

先生　が　私　に　本　を　買って　くれた
sensei　ga　watashi　ni　hon　wo　katte　kureta
teacher　NOM　I　DAT　book　ACC　buy　AUX (subj-control)

(*The teacher bought me a book.)*

Example 34: obj-id

私　が　先生　に　本　を　買って　もらった
watashi　ga　sensei　ni　hon　wo　katte　moratta
I　NOM　teacher　DAT　book　ACC　buy　AUX (obj-control)

(I got a book bought by the teacher..)

The pure aspect auxiliary *iru* in Example 30 adds tense and aspect information to the MRS of the sentence. The types of aspect and auxiliary examples can be seen in Table 6.

Table 6: Aspect types

Example	**Aspect**
おる (oru), いる (iru), いらっしゃる (irassharu)	progressive
おく(oku)	prospective
いく(iku)	inceptive
しまう (shimau)	terminative
ある (aru), ござる (gozaru)	perfective
くる (kuru)	perfect_progressive
みる (miru)	modal

[13] Thanks to Shigeko Nariyama for this example.

Some of the aspect auxiliaries add only the aspect information to the MRS semantics of the sentence. An example is *iru*, which adds aspect (see Example 30). The MRS contains the aspectual information in the INDEX. These are therefore called **pure-aspect auxiliaries**[14].

There are different types of honorificational information that can be added by aspect auxiliaries. For example, *oru* and *oku* add subject honorification with negative polarity, while *irassharu* adds subject honorification with positive polarity. This information is added to the CONTEXT part of the sign.

Other **aspect** auxiliaries make changes to the valence of the verbal complex, as can be seen in Example 31. These attach to a transitive verb. The verb's ARG1 is always a zero pronoun. The subject of the verb-aux complex is the ARG2 of the verb. The MRS for Example 31 is the following:

```
h1,e2:INDICATIVE:PRESENT:PERFECTIVE,
h1:proposition_m(e2, h3),
h4:_keeki_n(x5:THREE),
h6:u(x5, h7, h8),
h9:_taberu_v(e2, u10, x5),
h3 qeq h9,
h7 qeq h4
```

Figure 30: MRS for keeki ga tabete aru

Complex aspect auxiliaries add a relation to the MRS. Their ARG1 is identical to the ARG1 of the verb and their ARG2 is the handle of a proposition that outscopes the verbal relation:

```
h1,e2:INDICATIVE:PRESENT:MODAL,
h1:proposition_m(e2, h3),
h4:named(x5:PNG, "hanako"),
h6:(x5, h7, h8),
h9:_keeki_n(x10:THREE),
h11:u(x10, h12, h13),
h14:_taberu_v(e15:TENSED:INDICATIVE, x5, x10),
h16:_miru_aux(e2, x5, h17),
h17:proposition_m(e15, h18),
h3 qeq h16,
h7 qeq h4,
h12 qeq h9,
h18 qeq h14
```

Figure 31: MRS for Hanako ga keeki wo tabete miru

Perspective (subject-control) auxiliaries make their ARG1 identical to the ARG1 of the verb, add a ni-OBJ as ARG2 and link the handle of the proposition on top of the verb to their ARG3. Examples for perspective auxiliaries are: あげる (*ageru*), くれる (*kureru*), やる (*yaru*), さしあげる (*sashiageru*), くださる (*kudasaru*). The MRS for Example 33 is the following:

[14] Actually, the progressive aspect can be further classified using the semantic context and is therefore a bit more complex than the English progressive. See Yoshimoto (1997) for a more detailed discussion and further classification of the progressive aspect of *iru*.

```
h1,e2:PAST:INDICATIVE,
h1:proposition_m(e2, h3),
h4:_sensei_n(x5:THREE),
h6:u(x5, h7, h8),
h9:pron(x10:ONESG),
h11:(x10, h12, h13),
h14:_hon_n(x15:THREE),
h16:u(x15, h17, h18),
h19:_kau_v(e20:TENSED:INDICATIVE, x5, x15),
h21:_kureru_v(e2, x5, x10, h22),
h22:proposition_m(e20, h23),
h3 qeq h21,
h7 qeq h4,
h12 qeq h9,
h17 qeq h14,
h23 qeq h19
```

Figure 32: MRS for *sensei ga watashi ni hon wo katte kureta*

Perspective auxiliaries as well can add honorificational information. *kudasaru* adds subject honorification with positive polarity, while *sashiageru* adds subject honorification with negative polarity. As subject honorification needs a syntactic statement about the subject (namely, in SYNSEM.LOCAL.HEAD.FORMAL.SHON), the subject "belongs" to the auxiliary. Therefore, zero pronouns and complements have to be bound by the auxiliary after it is attached to the verb.

They can also add empathy information to the CONTEXT of the sentence. The empathy is set to ARG1 in the cases of *ageru*, *sashiageru* and *yaru* and to ARG2 in the cases of *kureru* and *kudasaru*.

Obj-id auxiliaries add an ARG1 and set the speaker's empathy to it. Their ARG2 is the ARG1 of the main verb and their ARG3 is the handle of a proposition that outscopes the verb. These can add honorificational information as well, as e.g. *itadaku* adds subject honorification with negative polarity. This is the MRS for Example 34:

```
h1,e2:PAST:INDICATIVE,
h1:proposition_m(e2, h3),
h4:pron(x5:ONESG),
h6:(x5, h7, h8),
h9:_sensei_n(x10:THREE),
h11:u(x10, h12, h13),
h14:_hon_n(x15:THREE),
h16:u(x15, h17, h18),
h19:_kau_v(e20:TENSED:INDICATIVE, x10, x15),
h21:_morau_v(e2, x5, x10, h22),
h22:proposition_m(e20, h23),
h3 qeq h21,
h7 qeq h4,
h12 qeq h9,
h17 qeq h14,
h23 qeq h19
```

Figure 33: MRS for *watashi ga sensei ni hon wo katte moratta*

4.4 The treatment of passive

The Japanese passive is morphologically built by attaching *reru* to a c-stem verb stem with a-inflection or *rareru* to a v-stem verb stem (see Example 35 and Example 36).

Example 35: c-stem passive

話さ　　れる
hana-sa　reru
speak　　PASSIVE

Example 36: v-stem passive

食べ　られる
tabe　rareru
eat　　PASSIVE

There are two types of Japanese passives: The simple and the adversative passive (or, direct and indirect passive, as it is called by Uda 1996). The simple passive is (parallel to other languages, such as English or German) only available for transitive and ditransitive verbs. An example for a simple transitive passive is Example 37.

Example 37: transitive passive

ご飯　が　井上　に　食べ　られた
gohan　ga　inoue　ni　tabe　rareta
rice　NOM　Inoue　DAT　eat　PASSIVE

(*The rice was eaten by Inoue*.)

The verb that gets a passive ending changes its *ga*-marked subject into a complement that is marked by *ni* or *kara*. The *wo*- or *ni*-marked complement is changed to a subject that is marked by *ga*.

Oshima (2003) proposes to add a relation *lack-conrol-rel* to the MRS in all cases of passive. We rather leave the representation of the direct passive parallel to the analysis in other languages (as for example Pollard and Sag 1994 do as well) and add a relation only in the case of the adversative passive. There is therefore no relation added to the MRS in the case of simple direct passive, such that the semantics of the passive sentence looks just alike the semantics of the active sentence. Only in the index, there is information about the passivization in E.PASS + (see Figure 34).

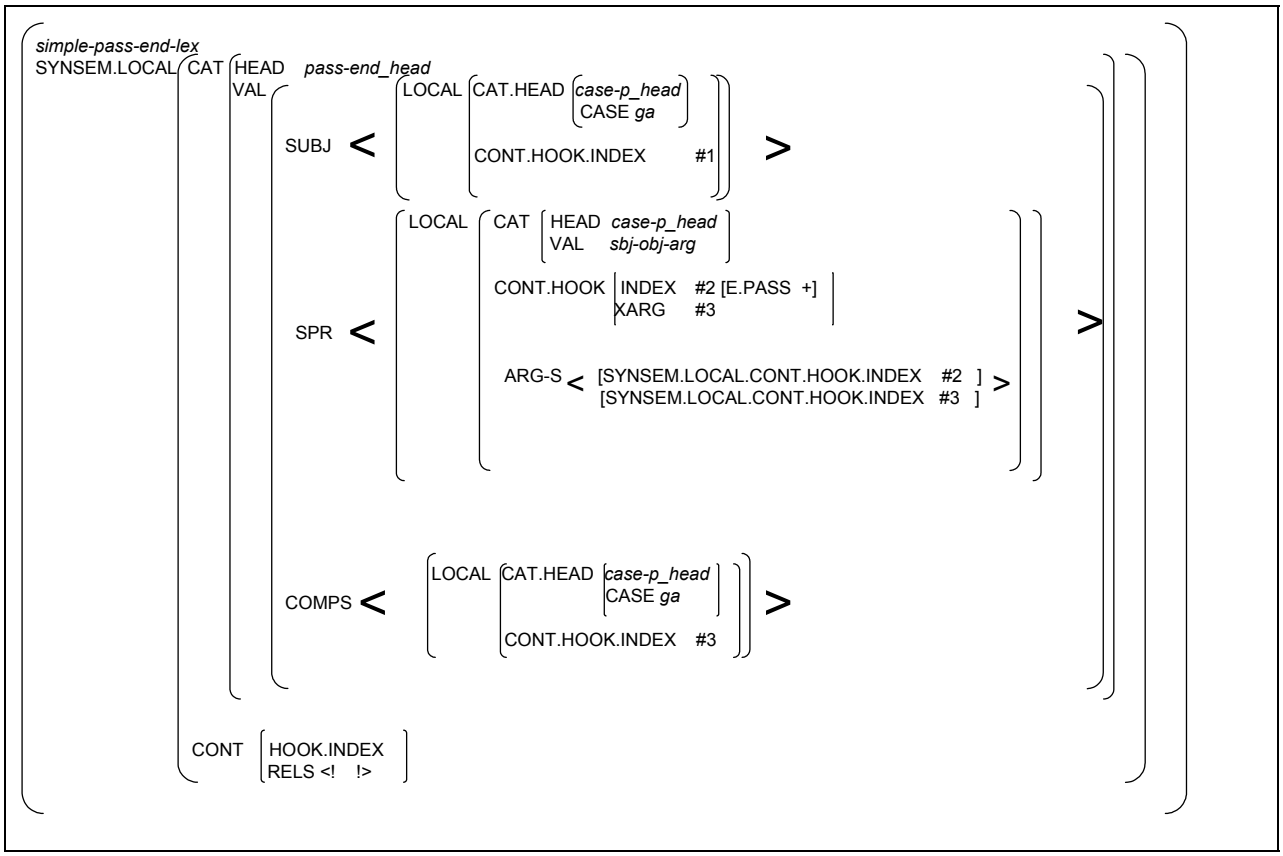


Figure 34: Feature structure of a passive ending

The verb stem and the verbal ending are combined with the *vstem-vend-rule*, which is an instance of the *head-specifier-rule* type.

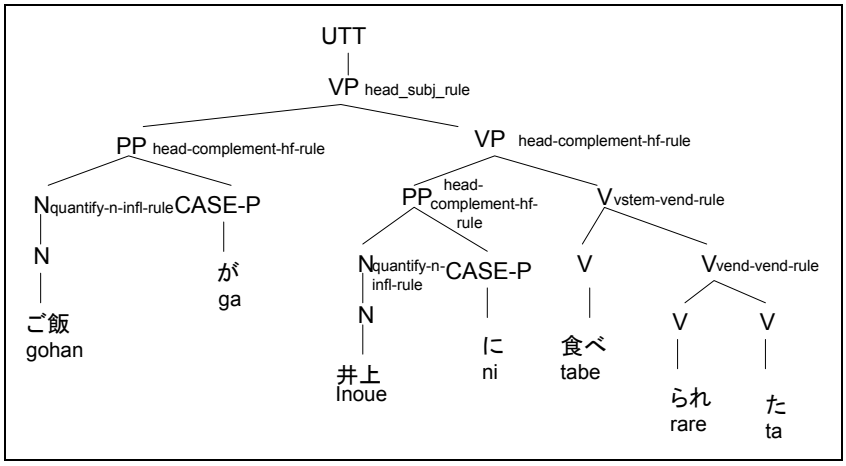


Figure 35: Passive treatment

If a verbal noun – light verb construction is involved, the argument access for the passive ending is slightly more complicated, as the passive must access the arguments of the verbal noun. But, with accessing ARG-S of the subcategorized verb, the problem can be solved.

Ditransitive verbs can be passivized in two ways, such that both complements can be subjects of the compound, as can be seen in Example 39 and Example 40[15].

[15] Thanks to T. Kuribayashi for the examples.

Example 38: active

フランシス	が	田中	に	ボール	を	当てた
Francis	ga	Tanaka	ni	booru	wo	ateta
Francis	NOM	Tanaka	DAT	ball	ACC	hit

(Francis threw the ball to Tanaka.)

Example 39: passive (acc-arg)

ボール	が	フランシス	によって	田中	に	当てられた
booru	ga	Francis	niyotte	Tanaka	ni	aterareta
ball	NOM	Francis	DAT	Tanaka	DAT	hit-passive

(The ball was thrown to Tanaka by Francis.)

Example 40: passive (dat-arg)

田中	が	フランシス	によって	ボール	を	当てられた
Tanaka	ga	Francis	niyotte	booru	wo	aterareta
Tanaka	NOM	Francis	DAT	ball	ACC	hit-passive

(Tanaka was hit by the ball by Francis.)

Passivization of a ditransitive verb changes the case marking of the arguments and keeps the argument structure. Emphasis (information in CONTEXT) is set to the surface subject and the semantics gets a mark in the event index.

Different to languages like English or German, the Japanese language contains a passive, which can be attached to intransitive as well as transitive verbs. This is used to express adversative relations. An example for adversative passive with an intransitive verb is given in Example 41.

Example 41

花子	が	弟	に	寝	られた
Hanako	ga	otouto	ni	ne	rareta
Hanako	NOM	brother	DAT	sleep	PASSIVE

(Hanako was affected by the fact that her brother slept.)

The adversative ending adds a *ga*-marked subject and changes the *ga* marking of the subject argument of the verb to *ni*. An adversative relation is added to the MRS, which contains as its ARG1 the extra *ga*-marked subject and as its ARG2 the handle of the event that is denoted by the verb. This is the MRS for the example:

```
h1,e2:PAST:INDICATIVE,
h1:proposition_m(e2, h3),
h4:named(x5, "hanako"),
h6:(x5, h7, h8),
h9:_otouto_n(x10:THREE),
h11:u(x10, h12, h13),
h14:_neru_v(e15, x10),
h16:adversative(e2, x5, h14),
h3 qeq h16,
h7 qeq h4,
h12 qeq h9
```

Figure 36: MRS for *Hanako ga otouto ni nerareta*

The adversative passive can be applied to transitive verbs as well. In this case, the verb complex requires a *ni*-marked argument, co-indexed with the *ga*-marked argument that was

required by the verb, a *wo*-marked argument, co-indexed with the *wo*-marked argument of the verb and an extra *ga*-marked argument. An adversative relation is added to the MRS, containing as its ARG1 the index of the extra *ga*-argument and as its ARG2 the handle of the event denoted by the main verb.

An example for transitive adversative is given in Example 42.

Example 42: transitive adversative

花子	が	弟	に	ケーキ	を	食べ	られた
Hanako	ga	otouto	ni	keeki	wo	tabe	rareta
Hanako	NOM	brother	NI	cake	ACC	eat	passive

(*Hanako was affected by the fact that her brother ate the cake*.)

This is the MRS of the transitive adversative in Example 42:

```
h1,e2:PAST:INDICATIVE,
h1:proposition_m(e2, h3),
h4:named(x5:PNG, "hanako"),
h6:(x5, h7, h8),
h9:_otouto_n(x10:THREE),
h11:u(x10, h12, h13),
h14:_keeki_n(x15:THREE),
h16:u(x15, h17, h18),
h19:_taberu_v(e20, x10, x15),
h21:adversative(e2, x5, h19),
h3 qeq h21,
h7 qeq h4,
h12 qeq h9,
h17 qeq h14
```

Figure 37: MRS for *Hanako ga otouto ni keeki wo taberareta.*

The same morphological process is used for a different purpose: Honorification. Adding reru/rareru to a verbal stem can as well have honorificational meaning, as can be seen in Example 43.

Example 43: honorific passive

先生	が	ご飯	を	食べ	られた。
sensei	ga	gohan	wo	tabe	rareta.
teacher	NOM	rice	ACC	eat	HON-PASSIVE

(*The teacher ate rice*.)

This verbal ending adds BACKGROUND information (a *subj-honor-rel*) to the feature structure, but does not affect the MRS. As it behaves morphologically just like passive, there is systematic ambiguity between the honorific and the passive reading.

4.5 Causative

The Japanese language contains the phenomenon of causative, as can be seen in Example 44 and Example 45.

Example 44: transitive causative

花子	が	妹	に	ピアノ	を	習わ	せる
Hanako	ga	imooto	ni	piano	wo	narawa	seru
Hanako	NOM	sister	DAT	piano	ACC	learn	causative

(Hanako makes her sister learn piano.)

Example 45: intransitive causative

花子	が	妹	に	寝	させる
Hanako	ga	imooto	ni	ne	saseru
Hanako	NOM	sister	DAT	sleep	causative

(Hanako puts her sister to sleep.)

Research literature has basically two positions for the treatment of the Japanese causative: The phrasal approach, as represented by the work of Gunji (1996a), and the lexical approach, as represented by the work of Manning et al. (1998). The position of the phrasal approach states that the combination of verbal stem and causative ending has to be done on the phrasal level. The main argumentation for this is based on the semantic behaviour of the causative construction and the possibilities for low and high attachment of modifiers. The position of the lexical approach takes the viewpoint that the phonological and morphological facts of the causative constructions trigger an approach where the combination has to be done on the lexical level.

The discussion shows that in any case there is a mismatch between morpho-phonological and syntactic or semantic constructions. As with other verbal stem – ending complexes, we make use of binary rules that are neither strictly lexical nor strictly phrasal, reflecting the fact that the status of these phenomena is unclear. The rule type **head-specifier-rule** is used for this kind of rules. It allows the combination of semi-lexical elements, such that the result allows different options for argument composition. Some verbal endings can add semantics (such as the causative ending), others do not (such as the past tense ending). Some endings change the valence structure and case marking of the verb complex, others do not.

In order to account for the different possibilities to access the verb or the causative relation for semantic modification (as described by Manning and Sag as well as by Gunji), we assume two types of causatives, which propagate the index of the causative relation or the verbal relation to the top. See, how the two MRSs for Example 46 in Figure 40 show the modification of the verbal and the cause event.

Example 46: From Manning et al. (1998)

紀子	が	勝	に	学校	で	走らせた
Noriko	ga	Masaru	ni	gakkou	de	hashiraseta
Noriko	NOM	Masaru	DAT	school	LOC	run (causative)

(Masaru made Noriko run at school)

h1,e2:PAST:INDICATIVE, h1:proposition_m(e2, h3), h4:named(x5, ″noriko″), h6:(x5, h7, h8), h9:named(x10, ″masaru″), h11:(x10, h12, h13), h14:_gakkou_n(x15:THREE), h16:u(x15, h17, h18), h19:_de_p(e20:NO_TENSE, x15, e2), h21:_hashiru_v(e2, x10, x5), h22:cause(e25, u24, x10, h23), h23:proposition_m(u27, h26), *h3 qeq h22,* *h7 qeq h4,* *h12 qeq h9,* *h17 qeq h14,* *h26 qeq h21*	h1,e2:PAST:INDICATIVE, h1:proposition_m(e2, h3), h4:named(x5, ″noriko″), h6:(x5, h7, h8), h9:named(x10, ″masaru″), h11:(x10, h12, h13), h14:_gakkou_n(x15:THREE), h16:u(x15, h17, h18), h19:_de_p(e20:NO_TENSE, x15, e2), h21:_hashiru_v(e22, x10, x5), h23:cause(e2, u25, x10, h24), h24:proposition_m(u27, h26), *h3 qeq h23,* *h7 qeq h4,* *h12 qeq h9,* *h17 qeq h14,* *h26 qeq h21*

Figure 38: Two MRSs for modified causative

Morphologically, the verb stem is changed by an inflectional rule (the very general **i-lexeme-v-stem-infl-rule** for *ne* and **pass-lexeme-stem-infl-rule** for *hashira*). Then, *seru* is attached to the verb stem in case of c-stem verbs and *saseru* in case of v-stem verbs[16].

In the case of intransitive causative, the causer (*Hanako* in the example) is marked by *ga* and the causee (*imooto* in the example) by *wo* or *ni*. In case of transitive causative, the causer is marked by *ga* and the causee by *ni*. Both types add the causer to the argument structure (such that a zero pronoun rule is needed that applies to the causative ending in case of a missing subject).

The intransitive causative ending links the external argument XARG of the specifier (i.e. the verb) to the causee, such that the arguments are correctly linked in the MRS. The transitive causative ending additionally takes the value of the complement of the verb.

[16] The irregular verbs *kuru* and *suru* become *kosaseru* and *saseru*, respectively.

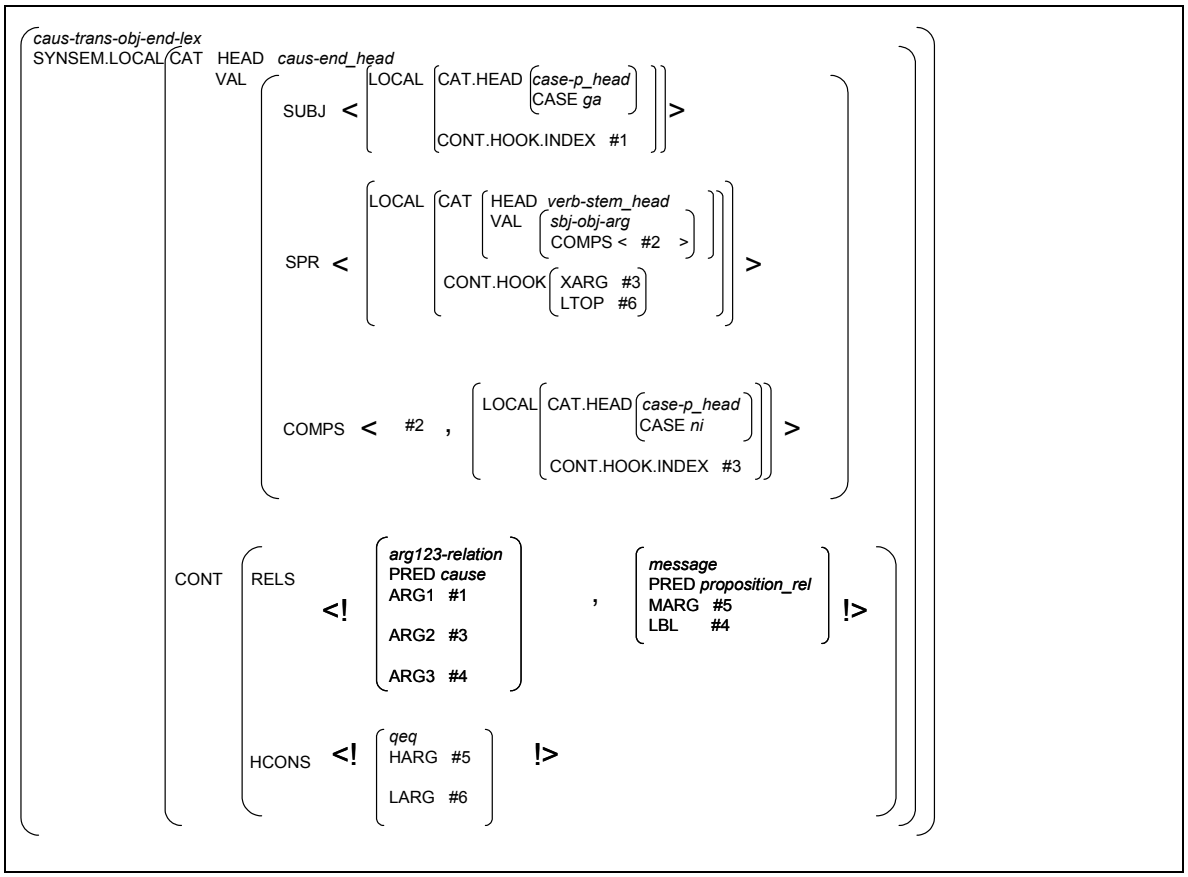


Figure 39: Causative-transitive verbal ending

The responsible rule is the **vstem-vend-rule**, a subtype of the **head-specifier-rule**. This is due to the peculiarities of argument composition here.

```
h1,e2:PRESENT,
h1:proposition_m(e2, h3),
h4:named(x5, "HANAKO"),
h6:def(x5, h7, h8),
h9:_imouto_n(x10:THREE),
h11:udef(x10, h12, h13),
h14:_neru_v(e15, x10),
h16:cause(e2, x5, x10, h17),
h17:proposition_m(e15, h18)},
h3 qeq h16,
h7 qeq h4,
h12 qeq h9,
h18 qeq h14
```

Figure 40: MRS for *Hanako ga imouto ni nesaseta*

5 Nominal Structures: Linking Syntax, Semantics and Pragmatics

Japanese nominal phrases show an interesting interaction of syntax, semantics and pragmatics. For processing natural language, it is necessary not to isolate the levels of linguistics, but to take their interactions into account. This is especially valid for Japanese, where much linguistic expression is dependent on semantics (such as it is the case with numeral classifiers) and context (the external context, such as in the case of honorification and the discourse context, as in the case of zero pronouns). The HPSG grammar framework is well suited for building representations and restrictions for this interaction, as it makes use of complex signs. We describe some phenomena of Japanese noun phrases and how the interaction can be used for processing and representation of the knowledge in Japanese language.

We first describe the basic structures of Japanese ordinary nouns in our grammar, where it can already be seen how syntax, semantics and pragmatics interact. The analysis of Japanese pronouns is based on the analysis of ordinary nouns (which is reflected by the fact that the pronouns type hierarchy is part of the noun hierarchy). Their analysis sets a stronger focus on the context. It will be shown that the reflexive is part of the pronoun type hierarchy. We show that the characteristics of the Japanese reflexive can be expressed in this system of interaction of linguistic levels. The analysis of named entities shows how information from external resources is included into the grammar. Nominalizations are in the noun type hierarchy as well, but have restricted semantic content and subcategorization features. Next, we show how the analysis of temporal expressions fits well into the general account and does not need special grammar structures. A description of noun modification by the genitive particle and noun modifiers follows. Numeral classifiers are a specific example of noun modification, and show interesting behaviour. Relative sentence constructions and pre-nominal adjectives show surprising similarities.

5.1 Ordinary nouns

An ordinary noun (belonging to the lexical type **ordinary-noun-lex**, as most of the nouns) does not specify any other category. It can have different honorific forms and it can occur with or without a particle in spoken language. For example, *hon* (book) is a non-honorific ordinary noun and *kyouju* (professor) is an honorific ordinary noun, which requires pragmatic agreement with the verb when being in the subject position of the phrase headed by this verb (see Chapter 9 for honorific agreement).

Japanese noun phrases usually do not contain determiners. Bond et al. (1994), Bond and Ogura (1998), Bond (2005), Siegel (1996a) and Heine (1998) describe this phenomenon and the problems and solutions for machine translation of Japanese into languages with determiners, such as German or English. Though, determiners like *kono, sono* or *ano* (this, that) are possible, if the determination cannot be inferred by the utterance context. Thus, the ordinary noun subcategorizes for an optional specifier, which is a determiner (*kono toki*, this time). The determiner, if expressed, adds quantificational information to the MRS of the noun phrase, as can be seen in Figure 41.

```
h4:_kono_det(x6, h5, h7),
h8:_toki_n(x6),
h5 qeq h8
```

Figure 41: MRS of *kono toki* (*this time*)

If, as in most cases for Japanese, a determiner is unavailable, there is still the need to express an underspecified quantification on the noun in the MRS, in order to make the semantics compatible with semantic output of other languages and to make scope restrictions work.

Therefore, we added a lexical rule (**quantify-n-inflectional-rule**) to the rule set that takes a noun as its argument and adds quantificational information. The lexical rule can be seen in Figure 42. It shows, how syntactic information, such as head type and valence information is linked with semantic information when building relations in RELS. The C-CONT contains information that is added to the MRS by the rule. It contains the (underspecified) quantification relation **udef_rel** with a reference to the noun index (ARG0 #i) and the scope information (RSTR #restr).

The resulting MRS for a noun phrase containing only the ordinary noun *hon* (*book*) can be seen in Figure 43. It is similar to the noun phrase containing a determiner and a noun, except for the underspecified quantification relation **udef**.

```
quantify-n-infl-rule :=
word2word-rule &
[SYNSEM [LOCAL [CAT [VAL saturated & [UNSAT -]]],
         LEX #lex],
 C-CONT [RELS <! [PRED #rel & udef_rel,
                  ARG0 #i,
                  RSTR #restr] !>,
         HCONS <! qeq &
                 [HARG #restr,
                  LARG #h] !>],
 ARGS.FIRST.SYNSEM [LOCAL [CAT [HEAD noun_head,
                                VAL [UNSAT +,
                                     SPR opt-1-arg &
                                       [FIRST [OPT + ,
                                               LOCAL.CAT.HEAD.KEYS.KEY #rel]],
                                     SUBJ olist,
                                     COMPS olist]],
                           CONT.HOOK [LTOP #h,
                                      INDEX #i]],
                LEX #lex]].
```

Figure 42: quantify-n-infl-rule

```
h4:_hon_n(x5:THREE),
h6:udef(x5, h7, h8),
h7 qeq h4
```

Figure 43: MRS of *hon*

5.2 Pronouns

Pronouns are a type of nouns and therefore build their type hierarchy under **n-lex**. This type hierarchy is shown in Figure 44.

Pronouns can refer to locations (**pron-loc-ref-lex**) or persons (**pers-pron-lex**) or can be demonstrative pronouns (**pron-demon-lex**). Table 7 shows the pronoun types and examples. Different to ordinary nouns, personal pronouns in Japanese often contain number information. This is reflected in the type hierarchy.

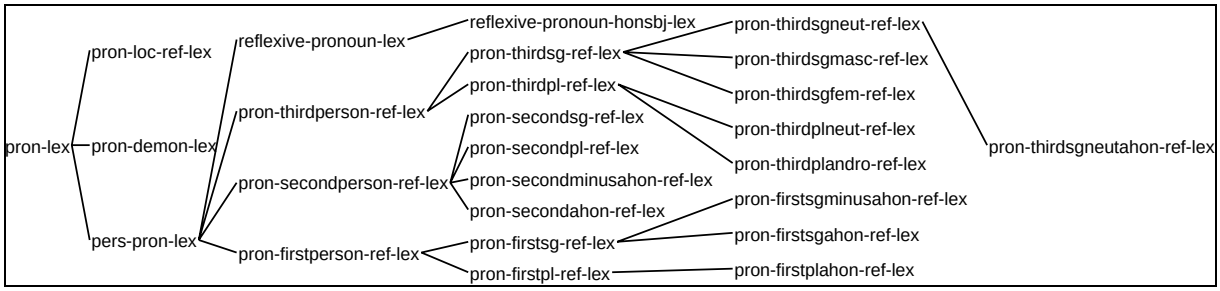


Figure 44: Type hierarchy of pronouns

Table 7: Pronouns

Type of Pronoun	Explanation	Example
pron-firstpl-ref-lex	pronoun with reference to first person plural	私達 (watashitachi, we)
pron-firstplahon-ref-lex	pronoun with reference to first person plural and addressee honorification with positive polarity	こちら (kochira, we)
pron-firstsgminusahon-ref-lex	pronoun with reference to the first person singular and addressee honorification with negative polarity	ぼく (boku, I)
pron-firstsg-ahon-ref-lex	pronoun with reference to the first person singular and addressee honorification with positive polarity	私(watashi, I)
pron-secondperson-ref-lex	pronoun with reference to second person	そっち (socchi, you)
pron-secondsg-ref-lex	pronoun with reference to second person singular	あなた (anata, you)
pron-secondpl-ref-lex	pronoun with reference to second person plural	君達 (kimitachi, you)
pron-secondminusahon-ref-lex	pronoun with reference to second person and addressee honorification with negative polarity	君 (kimi, you)
pron-secondahon-ref-lex	pronoun with reference to second person and addressee honorification with positive polarity	そちら (sochira, you)
pron-thirdsgneut-ref-lex	pronoun with reference to third person singular	それだけ (soredake, only that)
pron-thirdsgneutahon-ref-lex	pronoun with reference to third person neuter and addressee honorification	あちら (achira, it)
pron-thirdsgmasc-ref-lex	pronoun with reference to third person masculine	かれ (kare, he)
pron-thirdsgfem-ref-lex	pronoun with reference to third person feminine	彼女 (kanojo, she)
pron-thirdpl-ref-lex	pronoun with reference to third person plural	かれら (karera, they)
reflexive-pronoun-lex	reflexive pronoun	自分 (jibun, self)
reflexive-pronoun-honsbj-lex	reflexive pronoun with honorification	ご自分 (go-jibun, self)

pron-loc-ref-lex	pronoun with reference to a location	そこ (soko, there)
pron-demon-lex	demonstrative pronoun	それ (sore, that)

5.2.1 Personal pronouns

Pronouns that refer to persons encode person, number and gender – just as in English or German. They are thus less underspecified than ordinary nouns in Japanese. Pronouns are always definite and thus get a definite quantification in the MRS. Additionally; they can contain information about addressee honorification. This is reflected in the type hierarchy of personal pronouns, which links syntactic, semantic and pragmatic information in the types. Person, number and gender are added to the MRS (see Figure 45 for the MRS of a first person singular pronoun), while honorification is added as CONTEXT information. First person pronouns set a pragmatic perspective to the speaker of the utterance, namely empathy, and reference the pronoun with the speaker. So, they identify their HOOK.INDEX in CONT with the EMPEE (empathee) in CONTEXT.EMPATHY and the SPEAKER in C-INDS in CONTEXT. Additionally, they add an **entity-honor_rel** with negative polarity to CONTEXT.BACKGROUND, in order to reflect the fact that reference to oneself usually happens in humble or neutral form with respect to honorification. The HEAD.FORMAL gets [SHON -], such that agreement phenomena of honorification can be accounted for as well. The CONTEXT of a first person singular pronoun, as can be seen in Figure 46, thus co-indexes the speaker with the pronoun index, the empathy setting person as well as the person empathy is set to, and an **entity-honor_rel** in the BACKGROUND, which further identifies the honorer and the honored with a negative polarity.

Honorific second person pronouns on the other hand get [SHON +] in their HEAD information. They identify their INDEX with the ADDRESSEE and insert an **entity-honor_rel** with positive polarity to the CONTEXT.BACKGROUND.

```
h4: pron_rel (x5 [PNG.PN: ONESG ])
h6: def_rel (x5, h7)
h7 qeq h4
```

Figure 45: MRS of first person singular pronoun

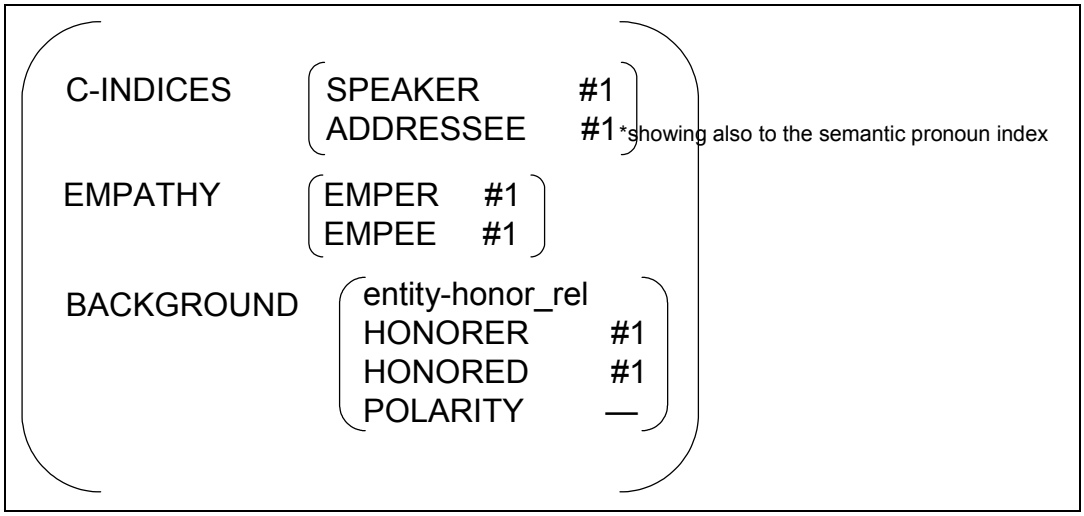


Figure 46: CONTEXT of a first person singular pronoun

5.2.2 Locative pronouns

Pronouns with reference to a location get an MRS, which is just like the MRS of the English and German location reference pronouns, reflecting the fact that their semantic behaviour shows no difference. The location reference pronoun inserts two relations to the MRS: A relation relative to the pronoun, such as **_soko_n_rel**, and a **place_relation**. Both identify their ARG0 and their LBL. The difference between the locative pronouns concerns the spatial position of the designated location relative to speaker and addressee (comparable to English "here" and "there"). This difference should in principle go into the pragmatic representation of the resulting sign and be linked to speaker, addressee, and locations in the utterance context. So far, we have no encoding for this difference in CONTEXT, but differ only in the relation names (e.g., **_soko_n_rel**, **_koko_n_rel**), representing "this place", or "that place".

```
h4: _soko_n_rel (x5)
h4: place_rel (x5)
h6: def_rel (x5, h7)
h7 qeq h4
```

Figure 47: MRS of the locative pronoun *soko*

5.2.3 Demonstrative pronouns

The MRS of demonstrative pronouns is as well just like the MRS for demonstrative pronouns in the English Resource Grammar. It contains a **dem_q_rel**, functioning as a quantifier and a **generic_nom_rel**. Their gender information is **neutr**. The meaning difference in the spatial relation of the pronoun to speaker, addressee and utterance location is not encoded.

```
h4: generic_nom_rel (x5:NEUT)
h6: dem_q_rel (x5, h7)
h7 qeq h4
```

Figure 48: MRS of demonstrative pronouns

5.2.4 The reflexive *jibun*

The Japanese reflexive does not appear very frequently, as we found no occurrence in the Verbmobil dialogue data and 0.009 % of the words in one year of Mainichi Shinbun newspaper data. Although this is the case, there is quite a lot of research literature on the binding properties of *jibun*. In order to set up a grammar, we need to explore the syntax, the semantics and the pragmatics of the reflexive and see how they interfere. Binding properties are relevant for this HPSG only, if they can be expressed as restrictions (as opposed to preferences) and if they are on the sentential level.

There is a fundamental difference of reflexive binding in Japanese and English as also being discussed in Sag/Wasow (1999, 166f): The Japanese reflexive can be bound to a subject in a higher sentence or an external entity that is not contained in the ARG-ST of a verb. Furthermore, the Japanese reflexive typically lacks number and gender information.

At first sight, the Japanese reflexive syntactically looks like an ordinary noun. It is followed either by a case particle (as in Example 48, Example 49 and Example 50) or by an adverbial particle (as in Example 47)) and can therefore function as a verbal argument as well as an adjunct. The difference is though the impossibility to take a determiner (see Example 51). This qualifies the reflexive to be a personal pronoun. Syntactically, all occurrences of *jibun* in the examples could be replaced by a pronoun like *kare* or *kanojo* (although the semantic changes). The difference of the reflexive to pronouns is in the restrictions to binding properties; it can be syntactically arranged in the personal pronoun type.

Example 47: From Makino/Tsutsui (1986)

Mearii(i)	wa	jibun(i)	de	nan demo		suru
Mary(i)	TOP	REFL	DE	everything		do

(Mary(i) does everything by herself(i).)

Example 48: From Makino/Tsutsui (1986)

Nakagawa(i)	wa	jibun(i)	ga	kyoudai	ni	haireru	to	omotte	inakatta
Nakagawa	TOP	REFL	NOM	Kyoto-Univ.	NI	enter	TO	think	AUX

(Nakagawa(i) didn't think that he(i) could enter Kyoto University.)

Example 49: From Makino/Tsutsui (1986)

Kazuo(i)	wa	jibun(i)	wo	hagemashita
Kazuo	TOP	REFL	ACC	brace

(Kazuo(i) braced himself(i))

Example 50: From Sag/Wasow (1999)

Hanako(i)	ga	jibun(i)	wo	tataita
Hanako(i)	NOM	REFL	ACC	hit

(Hanako(i) hit herself(i))

Example 51

*Mearii(i)	wa	sono	jibun(i)	de	nan demo	suru
Mary(i)	TOP	this	REFL	DE	everything	do

Mitsue (2001) states that the reflexive is not a lexical element, but a grammatical formative introduced to save derivation. This expresses the view that *jibun* has no semantics on its own. It is a right observation that the reflexive inherits its semantics from its antecedent. The reflexive itself has no explicit information about gender. Number information is though possible when using *jibuntachi*, the plural form of *jibun*, although *jibun* is underspecified for number, and the information must be coherent with the antecedent's information. The person information is – different to other personal pronouns – also not encoded in the reflexive itself, but dependent on the antecedent. It is though possible to encode honorification with the reflexive, when using *go-jibun*.

The reflexive therefore has some semantic and pragmatic content. It lacks person and gender information, but sometimes contains number and honorification. Unification seems a natural operation to combine the information on the reflexive and its antecedent. The unification for reflexives and their antecedents is not operating on the syntactic, but on the semantic and pragmatic levels.

I would therefore propose to view the reflexive as a personal pronoun with special implications to semantics and pragmatics. The semantics is fairly underspecified, but not empty, and can be enriched, when combined with the antecedent's content value.

Syntax, semantics and pragmatics constrain the binding of the reflexive pronoun. On the one hand, it is not obligatory for a reflexive to be bound in the clause or even in the sentence. It can also be bound by the discourse topic, as Example 52 shows, or it can function as a contrastive marker. Reflexive binding in the sentence is optional.

Example 52: From Gunji (1983)

jibun	wo	Naomi	ga	aisiteiru
self	ACC	Naomi	NOM	love

(Naomi loves s.o.)

On the other hand, there seem to be syntactic restrictions for reflexive binding. McCawley (1976) formulates the subject-antecedent conditions:

"...the reflexive refers back to the subject in the same simplex sentence or the subject in any higher sentence." (page 53)

"...the antecedent of the reflexive not only must be the subject but also must command the reflexive."(page 58)

Gunji (1983) adds:

"There is no object control." (page 133)

Here is an example for the subject-antecedent-condition given by McCawley (1976):

Example 53: From McCawley (1976)

Satoui	wa	$Tanaka_k$	ga	Harada	ni	$jibun_{i/k}$	ga	sukina	musume
Satou	TOP	Tanaka	NOM	Harada	DAT	REFL	NOM	like	daughter
wo	shoukaishita	koto	ni	odoroita					
ACC	introduced	NOM	DAT	surprise					

(Satou was surprised that Tanaka introduced to Harada the girl he loves.)

In this case, a topicalized subject and a nominative subject are possible antecedents of *jibun*.

The example above (Example 52) shows the command condition. While *Naomi* is the subject, (marked by *ga*), it cannot be the antecedent of *jibun*.

So-called backward reflexivization is based on the following kind of examples:

Example 54: From McCawley (1976)

jibuni	ga	gan	kamo	shirenai	koto	ga	Hiroshii	wo	nayamaseta
REFL	NOM	cancer	if	not know	NOM	NOM	Hiroshi	ACC	worried

(That he might have cancer worried Hiroshi)

This example contains a causative form of the verb. The antecedent of the reflexive is a surface complement, not a subject. It seems that the subject-antecedent-condition does not hold for surface subjects in case of causativization. The adversative passive shows similar effects:

Example 55

Keni	ga	Naomi	ni	$jibun_i$	wo	hihan	sareta[17]
Ken	NOM	Naomi	DAT	REFL	ACC	criticize	PASS

(Ken was adversely affected by Naomi's criticizing himself)

Another example of backwards reflexivization is the following that is quite similar to Example 52:

[17] Thanks to Akira Kusamoto for verifying this example.

Example 56

jibuni wo Naomii wa aisite-iru
REFL ACC Naomi TOP love

(Naomi loves herself)

It seems that topicalization and subject marking underlie different constraints in *jibun* binding. Manning and Sag (1998), using examples from a couple of languages including Japanese, show that "theories of grammar that define binding on surface phrase structure configurations or surface valence lists are unable to satisfactorily account for binding patterns". They propose to use the ARG-ST list as the locus of binding theory. As ARG-ST does not underlie changes in lexical rules for, e.g., passive, restrictions apply on the lexical level and account for the given effects. On the other hand, there seems to be a complex relation between semantics and pragmatics in Japanese reflexive binding. We conclude that the binding is not a question of syntactic functions, but of semantic indices and pragmatic restrictions.

Multiple occurrences of *jibun* in one clause must be bound to the same antecedent, as Example 57 shows:

Example 57: From Gunji (1983)

Ken wa Naomi ga jibun$_i$ ni jibun$_i$ no hon wo
Ken TOP Naomi NOM REFL DAT REFL GEN book ACC
okutta to omotteiru
send COMP think

(Ken thinks that Naomi has send herself her book)

Mitsue (2001) gives the following example to show that split antecedents are not allowed:

Example 58: From Mitsue (2001)

Takashii ga Mariko$_j$ ni Kenji$_k$ ga jibun$_{i/k/*i+j/*i+k}$ wo
Takashi NOM Mariko DAT Kenji NOM REFL ACC
suisenshita to tsugeta
recommended COMP reported

(Takashi reported Mariko that Kenji recommended self)

Katagiri (1991) explains jibun-binding from a semantic perspective:

"'jibun' ... could be explained solely in terms of coreference to semantic agent/experiencer of a clause containing 'jibun'."

This explanation is used to explain the correlation between perspective auxiliaries and *jibun* binding. He gives the following examples:

Example 59: From Katagiri (1991)

Hanako$_i$ wa Taro$_k$ ga jibun$_i$ ni hon wo yonde kureta
Hanako TOP Taro NOM REFL DAT book ACC read got
koto wo oboeteiru
NOM ACC remember

(Hanako remembered that Taro read a book for her)

Example 60: From Katagiri (1991)

*Hanako$_i$	wa	jibun$_i$	ga	Taro$_k$	ni	hon	wo	yonde	kureta
Hanako	TOP	REFL	NOM	Taro	DAT	book	ACC	read	got
koto	wo	oboeteiru							
NOM	ACC	remember							

(Hanako remembered that she read a book for Taro)

Reflexive binding is not restricted to the sentence. The reflexive *jibun* can as well be bound by the speaker or the discourse topic. It can also function as a contrastive marker. Pragmatic binding of *jibun* is due to the notions of old/new information and world view. Katagiri (1991) gives a coreference rule for *jibun* that is based on world view and semantics:

"The use of "jibun" is based on the judgement of identity of the referent of "jibun" to the semantic agent of an action or to the semantic experiencer of a mental state described in the sentence."

As sentential control of reflexives is optional, we decided to leave the control to a grammar-external module, just as the anaphoric binding. Though, we give the necessary information that can be accessed from the linguistic input, such that it is available for such a potential module.

Information for reflexive binding restrictions in JACY is stated on the semantic and pragmatic level, interconnecting the information available on these levels. The reflexive introduces REL to the NONLOCAL sign, reflecting the fact that antecedent and reflexive do not have to be bound locally. The value of REL is the semantic index of the reflexive pronoun. First of all, this locates the binding conditions to the semantic index and not to a syntactic function, due to the arguments given above. NONLOCAL is passed up in the trees, as they are built. Thus, the reflexive index is available for binding at any place in the tree. On the other hand, it is not possible to parse a second reflexive that inserts a different semantic index (see Example 60). Thus, the restriction that multiple occurrences of *jibun* must be bound to the same index is met. Furthermore, split antecedents are not possible and restricted by this. By passing up the NONLOCAL.REL, the MRS of a sentence with two occurrences with *jibun* shows the same index for both pronoun relations.

On the pragmatic level, the reflexive pronoun sets empathy to its own index. This accounts for the fact that binding to entities outside and inside the sentence is possible, if the speaker's empathy is focussed on the binding entity. If other entities in a sentence set empathy restrictions, such as for example the adversative passive, the reflexive binding gets a restriction as well. This is due to the fact that empathy can be set maximally to one index in a sentence. In Example 55, the adversative passive reading of the sentence allows only reference of *jibun* to *Ken*, because the reflexive sets empathy to *jibun* and the adversative passive to *Ken*. A similar effect can be seen with the perspective auxiliaries that as well set empathy to their arguments.

An external component for anaphoric binding will make use of the information given by the grammar:

- It will combine the index information with ontological information, such that the reflexive will be bound by animated entities only.
- It will reason about empathy in the discourse context, to find out where the speaker empathy is set to. If this is found, the reflexive can be bound. If more than one reflexive appears, all will be bound to the same antecedent, as they share their index in the MRS.

- It will take into account the situatedness in a physical and social environment, as described be Katagiri (1991). JACY gives a linking of speaker, addressee, perspective (empathy) and honorification, and therefore the basic information for situated binding.

5.3 Named entities

Any deep linguistic grammar has to decide on how much structure will be provided for the recognition of named entities. On the one hand, named entities with a clear structure can well be described by a deep grammar. On the other hand, the actual lexicalization of named entities is a potentially never ending set and constantly emerging, such that it might well be better covered by named entity recognition tools that can include regular expressions in the rule definitions. We decided to go for a combined approach. We provide a grammar structure for named entities and a restricted lexicon of known names. Additionally, we connected a named entity recognition tool (Sprout, see Drozdzynski et al. 2004) to the grammar processing, such that names not available in the JACY lexicon, but recognized by the tool, can be included into parsing (see Figure 50 for a JACY result that includes named-entity recognition results). In Section 10.3., we describe this integration, while here we will give an overview over the structures and types provided by the grammar.

Names are nouns that are neutral concerning honorification. We have first names, surnames, names of institutions, names of locations and product names[18]. They have different HEADs that are sorted in a type hierarchy of name-heads (which is itself a subtype of *noun_head*) as can be seen in Figure 49.

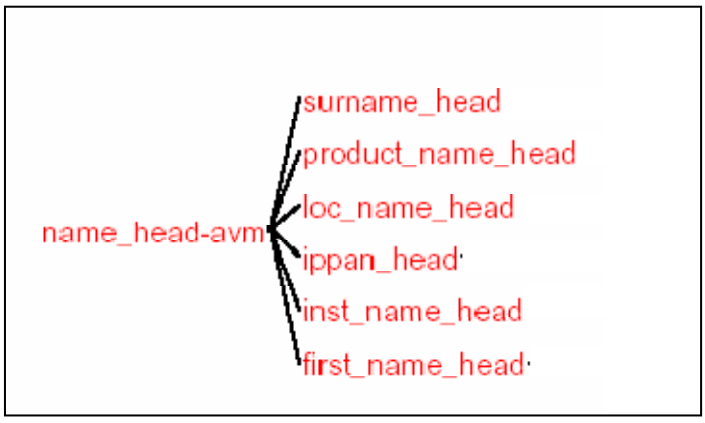


Figure 49: Type hierarchy of name heads

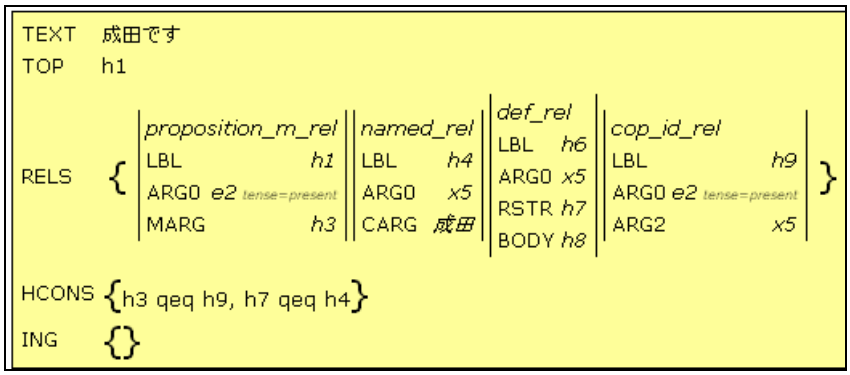


Figure 50: Including a result from NER to JACY

[18] *ippan-name* is used for those cases, where the named-entity recognition detects a name, but doesn't give the information about the name type.

Person names can occur with a determiner (as in Example 61), but much more often don't. Therefore, a person name undergoes a unary rule that inserts a quantifying relation, just as ordinary nouns. The difference, though, is that person names restrict their determiners to be definite in their lexical type.

Example 61: Person name with determiner, from the internet

この 田中 先生 から
kono tanaka sensei kara
this Tanaka Prof from

First and surnames combine in Japanese usually in the order surname – first name, but the other order is possible. The **compound-name-rule** takes a name and adds the possibility to modify another name. This rule inserts a relation named **compound** to the MRS, which combines the information on the names. The **compounds-rule**, an instance of head-final head-adjunct rules, combines the names.

```
h4:named(x5, "HIRATSUKA"),
h6:def(x5, h7),
h9:named(x10, "HANAKO"),
h11:def(x10, h12),
h9:compound(e14, x5, x10),
h7 qeq h4,
h12 qeq h9
```

Figure 51: MRS for "Hiratsuka Hanako"

First names modify a surname; surnames and institutions do not modify, and location names modify institutions (*aoyama daigaku*, Aoyama University).

All names specify a title. Titles are words like *kyouju, saN, kuN, sama, seNsei* that subcategorize for a specifier that can be a human name, but also institutional titles like *keNkyuushitsu* in *Fujita kenkyuushitsu (Fujita research institute)*, or *daigaku (University)*. Titles that attach to person names can add information on subject honorification with positive or negative polarity. They introduce two relations to the MRS: a *title* relation and a **title-id_rel** that combines the information on the title and the name. Institutional titles like *kenkyuushitsu* add their specific relation (such as **_kenkyuushitsu_rel**) and the **title-id_rel**.

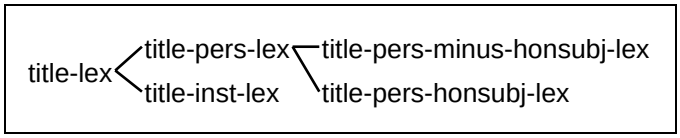


Figure 52: Type hierarchy of titles

5.4 Nominalizations

Some Japanese nouns, such as *koto*, *tame* or *mono*, have restricted semantic content. Therefore, they need something to specify the phrase content. This could be a determiner, a genitive, an adjective or a sentence. The external syntactic function of the phrase headed by these nouns is that of a noun. Uda (2001) describes the semantic function of a sentence containing such an "internally-headed relative clause" as being "descriptive; it does not restrictively modify the target".

These nouns cannot occur by themselves and need an obligatory argument (see Example 62, Example 63, Example 64 and Example 65). We call them nominalizations. Typically, nominalizations take a verb phrase and nominalize it (see Example 62). Examples for those nominalizing nouns are: *hou, koto, tame, katachi* and *no*. Some of the nominalizations, such as *, koto, tame* and *katachi,* can as well take a determiner (*sono hou*) or a PP with the particle

no (*watashi no hou*). The nominalizer *no* takes only a verb phrase. The argument is an obligatory specifier in any case.

Example 62

こちら	の	ほう	で	四時	に	終わる	こと	は	できます	けども
kochira	no	hou	de	yoji	ni	owaru	koto	ha	dekimasu	kedomo
we	GEN	side	DE	4 o'clock	NI	end	NOM	TOP	can	SP

(We could end at 4 o'clock.)[19]

Example 63

*こと	は	いい	です
*koto	ha	ii	desu
NOM	TOP	good	COP

Example 64

その	ため	に	ちょっと	スケジュール	の	ほう
sono	tame	ni	chotto	sukejuuru	no	hou
that	purpose	NI	somehow	schedule	NO	side

を	調整	させて	いただきたい	と	思いまして
wo	chousei	sasete	itadakitai	to	omoimashite
ACC	order	do	want	TO	think

(For that purpose, I think I want to order my schedule somehow.[20]*)*

Example 65

*ため	です
*tame	desu
purpose	COP

The structures these nominalizations occur in with verb phrases resemble relative sentence, but there are principal differences:

- The phrase left to the nominalization is obligatory, and therefore an argument of the nominalization, as can be seen in Example 65.
- The nominalization does not fill in an argument position in the subcat frame of the verb, as can be seen in Example 62.
- There are semantic differences of descriptiveness and truth condition differences between these and relative clauses, as described by Uda (2001) for *no*.

Nominalizations that occur with determiners, such as *kono*, *sono*, or *ano* (Example 64), resemble common nouns with these determiners, with just the difference that the determiners are obligatory.

Nominalizations with a modification by a *no* phrase (*kochira no hou* in Example 62) resemble modified nouns, but here as well the *no* phrase is obligatory.

Nominalizations are organized in a type hierarchy of nominalizers (see Figure 54). All nominalizations inherit from the type *nom-lex*. This is a sub type of nouns, with its HEAD

[19] From Verbmobil data.

[20] From Verbmobil data.

being a sub type of **noun_head**, which encodes the similarity of nominalizations to common nouns. The type **nom-lex** determines that the nominalization subcategorizes for a specifier. Nominalizations that take a VP as their argument determine that the specifier's head is a **verb_head** and that the argument is obligatory, using cross-classification with the subcategorization hierarchy to **nom_sc**. The nominalizations have sub types for special classes that have a noun head that can be subcategorized for by *na* to modify a noun. An example of these is the noun *you*, as in Example 66, with its tree structure in Figure 53.

Example 66

すぐ 帰る よう な 人
sugu kaeru you na hito
soon go home YOU NA human

(Someone who seems to go home soon.)

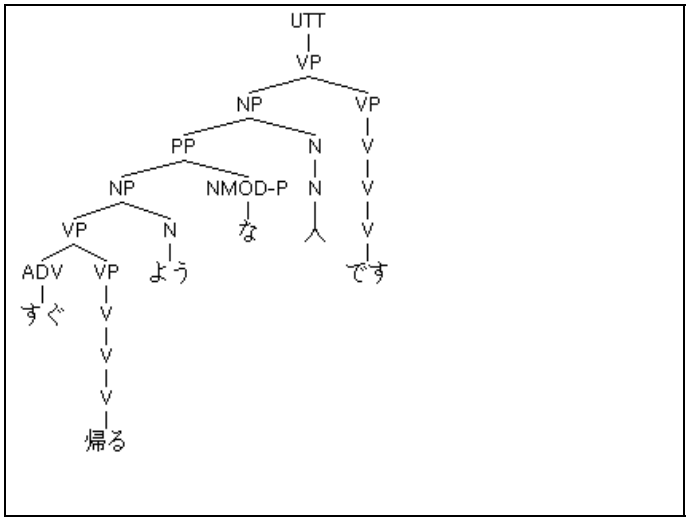


Figure 53: Tree structure of a phrase containing *you*

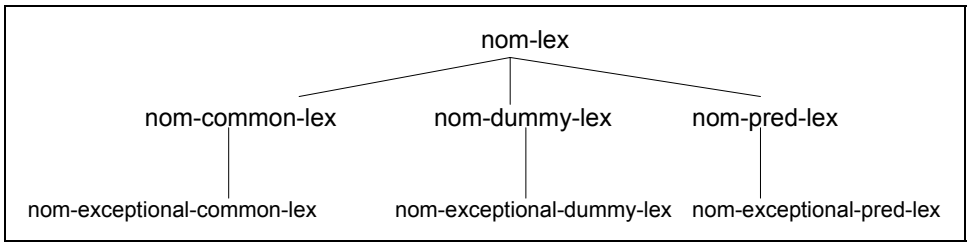


Figure 54: Type hierarchy of nominalizations

The MRS of a nominalization of a verb phrase contains the relation for the nominalization (e.g., **_koto_n_rel**), an undefined determination relation (**udef_rel**), a proposition on top of the event (**proposition_m_rel**), the verb relation and outscoping relations between **udef_rel** and nominalization as well as between proposition and verb. A simplified MRS can be seen in Figure 55.

```
h12:_koto_n_rel(x13, h14)
h15:udef_rel(x13, h16)
h14:proposition_m_rel(h18)
h9:taberu_rel (e11:INDICATIVE:PRESENT, x1,x2)
qeq (h16,h12)
qeq (h18,h9)
```

Figure 55: MRS for *taberu koto*

MRSs for nominalizations with determiners look just the same as those for noun phrases containing determiners (see Figure 56).

```
h4: _sono_det(x5, h6)
h8:_koto_n_rel(x5)
qeq (h6,h8)
```

Figure 56: MRS for *sono koto*

5.4.1 Data analysis of nominalizations

In order to explore the behaviour of nominalizations in newspaper data, we investigated the first 6 months of the Mainichi Shinbun newspaper corpus of 2002. We used ChaSen to tokenize and POS tag the corpus and Perl scripts to count the nominalizations. Table 8 shows the dissemination of nominalizations in the corpus.

Table 8: Dissemination of nominalizations in Mainichi Shinbun corpus

Nominalization	**Occurrences in Mainichi Shinbun corpus (First half year of 2002)**
no	53447
koto	45192
tame	22236
you	19303
mono	9593
n	6772
toki	6314
koro	5424
tokoro	3905
uchi	3766
wake	1949
hazu	1852

tsumori	795
hou	439
tabi	350
mon	170
hodo	115
goto	98
moto	6

We had a closer look at the most frequent nominalizations *koto*, *tame* and *mono* (*no* is too ambiguous to allow a proper quantitative investigation) to inspect their behaviour internally to the nominalization phrase in the data. The data was tagged with POS by ChaSen, and we applied Perl scripts to find categories that precede the nominalizations. Table 9 shows the contexts of occurring *koto*, *mono* and *tame* in the corpus. Far most words that occur before *koto*, *mono* or *tame* are of a verbal category. If we give an account for verb classes (incl. adjectives), *no*-particle and adnouns as occurring before *koto*, we cover 98.9% (44704) of *koto* occurrences. If we give an account for verb classes (incl. adjectives), *no*-particle and adnouns as occurring before *mono*, we cover 91.38% (8953) of *mono* occurrences. And if we give an account for verb classes (incl. adjectives), *no*-particle and adnouns as occurring before *tame*, 99.2% (17648) of *tame* occurrences are covered. We account for these. *Mono* can be preceded by nouns as well. This is due to the fact that *mono* occurs quite often as part of a nominal compound, being separated in some cases and not in others by ChaSen. This is covered by the rules for compound nouns.

Table 9: Occurring contexts of koto, mono and tame: What occurs before the nominalization

こと **koto**	もの **mono**	ため **tame**
45,192 occurrences of こと	9,798 occurrences of もの	17,792 occurrences of ため
verb classes: 19,045 occurrences of category POS 動詞-自立 with こと : independent verb 16,321 occurrences of category POS 助動詞 with こと : auxiliary 3,427 occurrences of category POS 動詞-非自立 with こと : dependent verb 1,823 occurrences of category POS 動詞-接尾 with こと : verb suffix	verb classes: 3,336 occurrences of category POS 助動詞 : auxiliary, べき , な , た 3,083 occurrences of category POS 動詞-自立 with もの : independent verb 418 occurrences of category POS 動詞-接尾 with もの : verb suffix 288 occurrences of category POS 動詞-非自立 : dependent verb	verb classes: 5,812 occurrences of category POS 動詞-自立 : independent verb 3,705 occurrences of category POS 助動詞 : auxiliary 551 occurrences of category POS 動詞-非自立 : dependent verb 463 occurrences of category POS 動詞-接尾 : verb suffix 247 occurrences of category POS 形容詞-自立 : adjective

adjectives: 887 occurrences of category POS 形容詞-自立 with こと : adjective	adjectives: 613 occurrences of category POS 形容詞-自立 : adjective 70 occurrences of category POS 形容詞-非自立 : dependent adjective 1 occurrence of category POS 形容詞-接尾 : adjective suffix, っぽい	
particles: 2,222 occurrences of category POS 助詞-連体化 with こと : の particle 1 occurrence of category POS 助詞-係助詞 with こと : particle も	particles: 890 occurrences of category POS 助詞-連体化 : particle, の 315 occurrences of category POS 助詞-格助詞-連語 : という, による, に関する, に対する 124 occurrences of category POS 助詞-格助詞-一般 : case particle 13 occurrences of category POS 助詞-係助詞 : particle 9 occurrences of category POS 助詞-接続助詞 : particle,て , で 6 occurrences of category POS 助詞-副詞化 : particle, に , と 4 occurrences of category POS 助詞-副助詞 : particle, なんて , など 4 occurrences of category POS 助詞-並立助詞 : particle,と, や 1 occurrence of category POS 助詞-格助詞-引用 : particle, と 1 occurrence of category POS 助詞-終助詞 : particle, っけ	particles: 6465 occurrences of 助詞-連体化 : の 10 occurrences of 助詞-接続助詞 : particle, ga 3 occurrences of 助詞-係助詞 : particle, は 2 occurrences of 助詞-終助詞: particle, よ、な 2 occurrences of 助詞-副助詞 : particle, made

adnoun: 979 occurrences of category POS 連体詞 with こと : adnoun (この, そんな, どんな, 小さな, たいした, 同じ)	adnoun: 254 occurrences of category POS 連体詞 : adnoun	adnoun: 405 occurrences of 連体詞 : adnoun
sentence beginning: 376 occurrences of category POS 記号-括弧閉 with こと : closing brackets 66 occurrences of category POS 記号-一般 with こと : symbol (――, ◇, ◆, ＝) 10 occurrences of category POS 記号-括弧開 with こと : opening brackets (（, 『, 「, “) 9 occurrences of こと in the beginning of a sentence (8 times こと は, 1 time こと も)	sentence beginning: 48 記号-括弧開 : symbol, opening brackets 31 記号-読点 : symbol, comma 29 記号-括弧閉 : symbol, closing brackets 19 mono as beginning of sentence 11 記号-空白 : symbol, space 4 記号-一般 : ◇ , －－	
	nouns: 112 occurrences of category POS 名詞-一般 : noun 13 occurrences of category POS 名詞-固有名詞-人名-名 : person name 8 occurrences of category POS 名詞-数 : number 7 名詞-代名詞-一般 : pronoun, なに 5 occurrences of category POS 名詞-形容動詞語幹 : adjectival stem noun 3 occurrences of category POS 名詞-接尾-一般 : noun suffix, ら, め, たて 3 occurrences of category POS 名詞-接尾-助数詞 : numeral classifier, 本, 年 1 occurrence of category	

	POS 名詞-固有名詞-一般 : name 1 occurrence of category POS 名詞-固有名詞-地域-国 : location name, 中国	
minor: 12 occurrences of category POS 名詞-一般 with こと : ゲタ, す, 妻, 部長, ナイ, カイ, サイ occurrences of category POS 助詞-接続助詞 with こと : て and ば occurrences of category POS 名詞-サ変接続 with こと : ぐ	minor: 38 名詞-サ変接続 : verbal noun 14 副詞-一般 : adverb 11 接頭詞-名詞接続 : nominal prefix, す、生、大 8 副詞-助詞類接続 : adverb - particle connection, はっきり, まだ 感動詞 : emotional expression, う、な	minor: 91 occurrences of 記号-括弧閉 : symbol, closing bracket 11 occurrences of 記号-読点: symbol, comma 9 occurrences of 名詞-一般 : 天国, す、一つ、作品、いが、良識、国旗 8 occurrences of 記号-一般 : symbol 6 occurrences of 記号-括弧開 : symbol, opening bracket 1 occurrence of 記号-空白 : symbol, blank 1 occurrence of 副詞-一般 : 本当に

The external function of the nominalizations in the sentence can be approximated by the words that follow. In Mainichi Shinbun, nominalizations take particles in 96.73%. These are 46.6% case particles *ga* and *wo*, 19.27% adverbial particles *de* and *ni* and 18.23% topic particles *wa*. Therefore, nominalizations behave like ordinary nouns in their external functions.

Chung and Kim (2002) state for Korean Internally Headed Relative Clauses that one of the arguments of the main predicate is associated to either an argument in the verb phrase inside of the nominalization phrase or to the event of the verb phrase inside of the nominalization phrase. If the matrix verb is an action verb in Korean, we obtain a reading where the verb in the nom-construction has the same arguments as the matrix verb, as in the following example from Chung and Kim:

Example 67

```
John-un   Mary-ka   talli-nun  kes-ul   capassta
John-TOP  Mary-NOM  run-PNE    KES-ACC  caught
```

(John caught Mary who was running)

But if the matrix verb is a type of a recognition verb (such as see, remember etc.), we have event readings:

Example 68

John-un Mary-ka talli-nun kes-ul mollassta
John-TOP Mary-NOM run-PNE KES-ACC not.know

(John didn't know that Mary was running)

The action verb possibility (Example 67) does not exist in Japanese in the same way as in Korean. This seems to be one of the differences of Korean and Japanese grammatical structures. We therefore have to connect the argument structure of the main verb in the sentence with the event of the complement sentence. As because this can as well be a negated event or a question in Japanese, the connection is done on the proposition on the event, as shown in Figure 55.

5.5 Temporal expressions

The Verbmobil domain of appointment scheduling requires precise analysis of various types of temporal expressions. Example 69 to Example 73 show some typical temporal expressions in Japanese, in the Verbmobil corpus.

Nouns used in temporal expressions can syntactically behave like ordinary nouns. An example is the word 日 (*hi*, day) in a construction like Example 69.[21] A special class are the nouns that denote days, such as 月曜日 (*getsuyoubi*, Monday) or 火曜日 (*kayoubi*, Tuesday). As these get a special semantic description, they get their own type in the lexical type hierarchy, **day-lex,** which has the sub types **dofw-n-lex** (weekdays) **dofm-n-lex** (month days) and **mofy-n-lex** (month names) . A third class are the nouns that occur frequently without particles, such as 午後 (*gogo*, afternoon), 朝 (*asa*, morning), 2時 (*niji*, two o'clock), 一月 (*ichigatsu*, January), or 三日 (*mikka*, the third) and that belong to a type **temp_numeral-lex**, which is also a sub type to nouns.

Temp-numerals and day nouns occur in constructions as in Example 71, where the semantic relation between the two words requires a head-complement structure, while in combinations of ordinary nouns and, e.g. day-of-week nouns there is a head-adjunct relation (see Example 72).

Basically, no special rules are required for the treatment of Japanese temporal expressions, as restrictions are encoded lexically in the subcategorizational and modificational behavior of the lexical types or items. Consider Example 73 with its chart in Figure 57, where several phrasal types are applied that are used for other constructions as well.

Example 69

その 日 は いい です
sono hi wa ii desu
that day TOP good COP

(That day is good.)

Example 70

六月 十三日 の 火曜日 から は いかが でしょう か
rokugatsu juusannichi no kayoubi kara wa ikaga deshou ka
June 13th NO Tuesday from TOPIC good COP QUE

(Would Tuesday the 13th of June, in the afternoon, suit you?)

[21] Others are 週 (shuu, week) and 時間 (jikan, time)

Example 71

十七日　　の　月曜日
juunananichi　no　getsuyoubi
17th　　GEN　Monday

(Monday the 17th)

Example 72

来週　　の　火曜日
raishuu　no　kayoubi
next week　GEN　Tuesday

(Tuesday of next week)

Example 73

来週　　の　水曜日　　十七日　　は　どう　です　か
raishuu　no　suiyoubi　juunananichi　ha　dou　desu　ka
next week　GE　Wednesday　17th　　TOP　how　COP　QUE

(How would next week Wednesday the 17th be?)

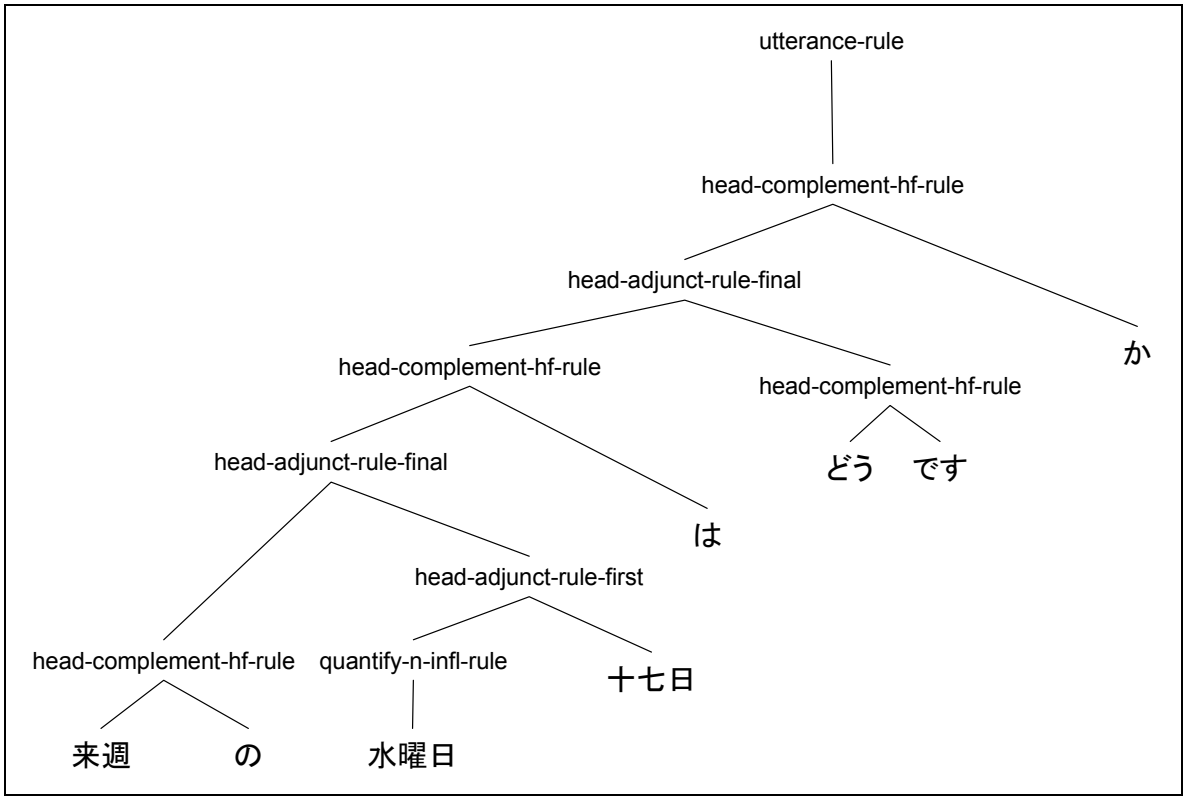


Figure 57: Use of phrasal rules for temporal constructions

```
h1,x2
h3:_raishuu_n(x4)
h5:u(x4, h6)
h8:_no_p(e10:NO_TENSE, x4, x9)
h8:named(x9, "_wed")
h11:(x9, h12)
h14:card(u16, x15, "10")
h17:plus(u19, x15, h14, h18)
h18:card(u20, x15, "7")
h21:u(x15, h22)
h17:_nichi_numcl(x15)
h17:degree(u25, e24)
h8:unspec_adj(e24, x9)
h26:_wa_p(e27:NO_TENSE, x9, x2)
h28:prep-mod(e29:NO_TENSE, x2, u30)
h31:_way_a(x2)
h32:whq(x2, h33, h34)
h1:question_m(x2, h35)
h6 qeq h3,
h12 qeq h8,
h22 qeq h17,
h33 qeq h31,
h35 qeq h28
```

Figure 58: MRS for *raishuu no suiyoubi juunananichi wa dou desu ka*.

5.6 Numeral classifiers[22]

Much attention has been paid to the semantic aspects of Japanese numeral classifiers, and in particular, the semantic constraints which govern which classifiers co-occur with which nouns (Matsumoto 1993; Bond and Paik 2000). A more neglected aspect of this linguistic phenomenon is the syntax of numeral classifiers: How they combine with number names to create numeral classifier phrases, how they modify head nouns, and how they can occur as stand-alone NPs.

Paik and Bond (2002) divide Japanese numeral classifiers into five major classes: *sortal*, *event*, *mensural*, *group* and *taxanomic*, and several subclasses. The classes and subclasses can be differentiated according to the semantic relationship between the classifiers and the nouns they modify, on two levels: First, what properties of the modified noun motivate the choice of the classifier, and second what properties the classifiers predicate of the nouns. As we are concerned here with the syntax and compositional semantics of numeral classifiers, we will focus only on the latter. Sortal classifiers, (*kind*, *shape*, and *complement* classifiers) serve to individuate the nouns they modify. Event classifiers quantify events, characteristically modifying verbs rather than nouns. Mensural classifiers measure some property of the entity denoted by the noun they modify (e.g., its length). NPs containing group classifiers denote a group or set of individuals belonging to the type denoted by the noun. Finally, taxonomic classifiers force a kind or species reading on an NP.

Internally, Japanese numeral classifier expressions consist of a number name followed by a numeral classifier (Example 74, Example 75, and Example 76). In this, they resemble some date expressions:[23]

[22] This chapter is an extended version of joint research with Emily Bender, published in Bender and Siegel (2004).

[23] Note that many of the time units are ambiguous with date expressions, although some, like the one for months shown in (1), are distinguished.

Example 74

十　枚
juu　mai
10　NumCL

Example 75

十　円
juu　en
10　yen

Example 76

十　ヶ月
juu　kagetsu
10　months

Example 77

十　月
juu　gatsu
10　month

In fact, both numeral classifiers and date expressions are tagged as numeral classifiers by the morphological analyzer ChaSen (Asahara and Matsumoto, 2000). However, date expressions do not have the same combinatoric potential (syntactic or semantic) as numeral classifiers. We thus give date expressions a distinct analysis.

Externally, numeral classifier phrases (NumClPs) appear in at least four different contexts: alone, as anaphoric NPs (Example 78); preceding a head noun, linked by the particle *no* (Example 79); immediately following a head noun (Example 80); and 'floated', right after the associated noun's case particle or right before the verb (Example 81). These constructions are distinguished pragmatically (Downing, 1996).[24]

Example 78

二　匹　　を　飼う
ni　hiki　wo　kau
2　NumCl　ACC　raise

((I) am raising two (small animals).)

Example 79

二　匹　　の　猫　　を　飼う
ni　hiki　no　neko　wo　kau
2　NumCl　GEN　cat　ACC　raise

((I) am raising two cats.)

[24] Downing also notes NumClPs following the head noun with an intervening no. As this rare construction did not appear in our data, we have not incorporated it into our account.

Example 80

猫　二　匹　を　飼う
neko　ni　hiki　wo　kau
cat　2　NumCl　ACC　raise

((I) am raising two cats.)

Example 81

猫　を　(二　匹)　家　で　(二　匹)　を　飼う
neko　wo　(ni　hiki)　ie　de　(ni　hiki)　wo　kau
cat　ACC　2　NumCl　house　LOC　2　NumCl　ACC　raise

((I) am raising two cats in my house.)

NumClPs can be modified by elements such as *yaku* 'approximately' (before the number name) or *mo*'even' (after the floated numeral classifiers).

The above examples illustrate the contexts with a sortal numeral classifier, but mensural numeral classifiers can also appear both as modifiers (Example 82) and as NPs in their own right (Example 83):

Example 82

二　キロ　の　りんご　を　買った
ni　kiro　no　ringo　wo　katta
two　NumCl (kg)　GEN　apple　ACC　bought

((I) bought two kilograms of apples.)

Example 83

二　キロ　を　買った
ni　kiro　wo　katta
two　NumCl (kg)　ACC　bought

((I) bought two kilograms.)

NumClPs serving as NPs can also appear as modifiers of other nouns:

Example 84

三　人　の　出会い　は　８０　年　春
san　nin　no　deai　wa　80　nen　haru
3　NumCl　GEN　meeting　TOP　80　year　spring

(The three's meeting was in the spring of 80.)

Example 85

一　キロ　の　値段　は　百　円　です
ichi　kiro　no　nedan　wa　hyaku　en　desu
1　kg　GEN　price　TOP　100　yen　COP

(The price of/for 1 kg is 100 yen.)

As a result, tokens following the syntactic pattern of (Example 79) and (Example 82) are systematically ambiguous, although the non-anaphoric reading tends to be preferred.

Certain mensural classifiers can be followed by the word *han* 'half':

Example 86

二　キロ　半
ni　kiro　han
two　kg　half

(two and a half kilograms)

In order to build their semantic representations compositionally, we make the numeral classifier (here, *kiro*) the head of the whole expression, and *ni* and *han* its dependents. *Kiro* can then orchestrate the semantic composition of the two dependents as well as the composition of the whole expression with the noun it modifies.

Although they aren't tagged as numeral classifiers by ChaSen, we extended our analysis of mensural classifiers to certain elements that appear before numbers, namely currency symbols (such as $), and prefixes like *No.* 'number' in Example 87.

Example 87

講座　No.　１２３４　号
kouza　No.　1234　gou
account　number　1234　number

(account number 1234)

Finally, we found that number names can sometimes occur without numeral classifiers, either as modifiers of nouns or as anaphora:

Example 88

講座　１２３４　を　閉じたい
(kouza)　1234　wo　tojitai
(account)　1234　ACC　close.volitional

((I) want to close (account) 1234.)

5.6.1 Data: Distribution

We used ChaSen to segment and tag 10,000 paragraphs of the Mainichi Shinbun 2002 corpus. Of the resulting 490,202 words, 11,515 (2.35%) were tagged as numeral classifiers. 4,543 of those were potentially time/date expressions, leaving 6,972 numeral classifiers, or 1.42% of the words. 203 orthographically distinct numeral classifiers occur in the corpus. The most frequent is *nin* (the numeral classifier for people) which occurs 1,675 times.

We sampled 100 sentences tagged as containing numeral classifiers to examine the distribution of the constructions outlined. These sentences contained a total of 159 numeral classifier phrases and the vast majority (128) were stand-alone NPs. This contrasts with Downing's (1996) study of 500 examples from modern works of fiction and spoken texts, where most of the occurrences are not anaphoric.

Furthermore, while our sample contains no examples of the floated variety, Downing's contains 96. The discrepancy probably arises because Downing only included sortal numeral classifiers, and not any other type. Another possible contributing factor is the effect of genre.

5.6.2 Semantic representations

One of our main goals in implementing an analysis of numeral classifiers is to compositionally construct semantic representations, and in particular, Minimal Recursion

Semantics (MRS) representations. The representation we build for Example 79[25] and Example 80[26] is as in Figure 59.

```
h1,e2:INDICATIVE:PRESENT
h1:proposition_m(e2, h3),
h4:card(x5:THREE, "2"),
h4:_neko_n(x5),
h7:udef(x5, h8,
h10:_kau_v_2(e2, u11, x5)},
h3 qeq h10,
h8 qeq h4
```

Figure 59

This can be read as follows: A relation of raising holds between *z* (the unexpressed subject), and *x*. *x* denotes a cat entity, and is bound by an underspecified quantifier (as there is no explicit determiner). *x* is also an argument of a **card_rel** (short for 'cardinal relation'), whose other argument is the constant value 2, meaning that there are in fact two cats being referred to.

For anaphoric numeral classifiers, the representation contains an underspecified **noun_relation**, to be resolved in further processing to a specific relation:

```
h1,e2:INDICATIVE:PRESENT
h1:proposition_m(e2, h3),
h4:noun_relation(x5),
h4:card(x5, "2"),
h7:udef(x5, h8),
h10:_kau_v_2(e2, u11, x5)},
h3 qeq h10,
h8 qeq h4
```

Figure 60: MRS for *ni kiro wo katta*.

Mensural classifiers have somewhat more elaborated semantic representations, which we treat as similar to English measure NPs (Flickinger and Bond, 2003). On this analysis, the NumClP denotes the extent of some dimension or property of the modified N. This dimension or property is represented with an underspecified relation (**unspec_adj_rel**), and a **degree_rel** relates the measured amount to the underspecified adjective relation.

The underspecified adjective relation modifies the N in the usual way. This is illustrated in Figure 61, which is the semantic representation assigned to Example 82.[27]

25

二	匹	の	猫	を	飼う
ni	hiki	no	neko	wo	kau
2	NumCl	GEN	cat	ACC	raise

26

猫	二	匹	を	飼う
neko	ni	hiki	wo	kau
cat	2	NumCl	ACC	raise

27 The relationship between the **degree_rel** and the **unspec_adj**_*rel* is not entirely apparent in this abbreviated notation. The first argument of the **degree_rel** is in fact the predicate name of the **unspec_adj_rel**, and not the whole relation.

```
h1,e2:PAST:INDICATIVE,
h1:proposition_m(e2, h3),
h4:card(x5:PNG, "2"),
h7:u(x5, h8),
h4:_kiro_numcl(x5),
h4:degree(u11, e10),
h12:unspec_adj(e10, x13:THREE),
h12:_ringo_n(x13),
h14:u(x13, h15),
h17:_kau_v(e2, u18, x13),
h3 qeq h17,
h8 qeq h4,
h15 qeq h12
```

Figure 61: MRS for *ni kiro no ringo wo katta*

When mensural NumClPs are used anaphorically (Example 83), the element modified by the unspec_adj rel is an underspecified noun relation, analogously to the case of sortal NumClPs used anaphorically:

```
h1,e2:PAST:INDICATIVE,
h1:proposition_m(e2, h3),
h4:noun_relation(x5),
h6:card(x7, "2"),
h9:u(x7, h10),
h6:_kiro_numcl(x7),
h6:degree(u13, e12),
h4:unspec_adj(e12, x5),
h14:u(x5, h15),
h17:_kau_v(e2, u18, x5),
h3 qeq h17,
h15 qeq h4
```

Figure 62: MRS for *ni kiro wo katta*

5.6.3 The analysis

Our analysis consists of:

A lexical type hierarchy cross-classifying numeral classifiers along three dimensions (Figure 63).

A special lexical entry for *no* for linking NumClPs with nouns.

A unary-branching phrase structure rules for promoting NumClPs to nominal constituents.

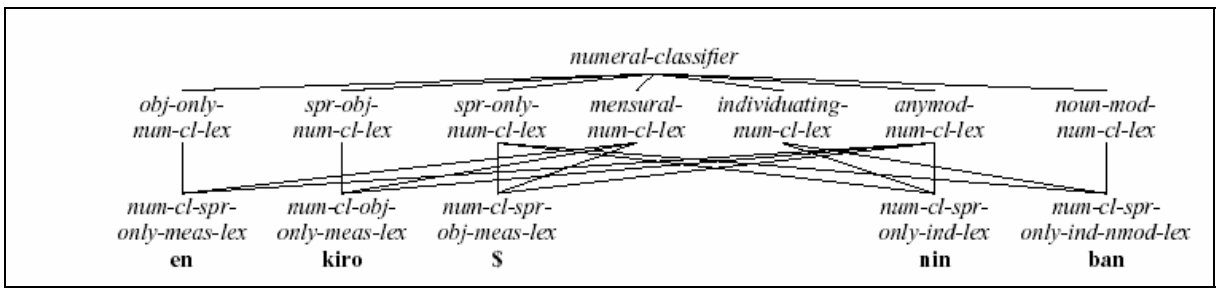


Figure 63: Type hierarchy under *numeral classifier*

5.6.4 Lexical types

Figure 63 shows the lexical types for numeral classifiers, which are cross-classified along three dimensions:

1. semantic relationship to the modified noun(**individuating** or **mensural**)
2. modificational possibilities (NPs or PPs: **anymod**/NPs: **noun-mod**)
3. relationship to the number name (number name precedes: **spr-only**, number name precedes but may take *han*: **spr-obj**, number name follows: **obj-only**).

Not all the possibilities in this space are instantiated (e.g., we have found no sortal classifiers which can take *han*), but we leave open the possibility that we may find in future work examples that fill in the range of possibilities.

The constraint in Figure 64 ensures that all numeral classifiers have the head type **num-cl head**, as required by the unary phrase structure rule. Furthermore, it identifies two key pieces of semantic information made available for further composition, the INDEX and LTOP (local top handle) of the modified element with the numeral classifier's own INDEX and LTOP, as these are intersective modifiers (Bender et al., 2002). The constraints on the type **num-cl head** (not shown here) ensure that numeral classifiers can modify only saturated NPs or PPs (i.e., NPs marked with a case postposition *wo* or *ga*), and that they only combine via intersective head-modifier rules.

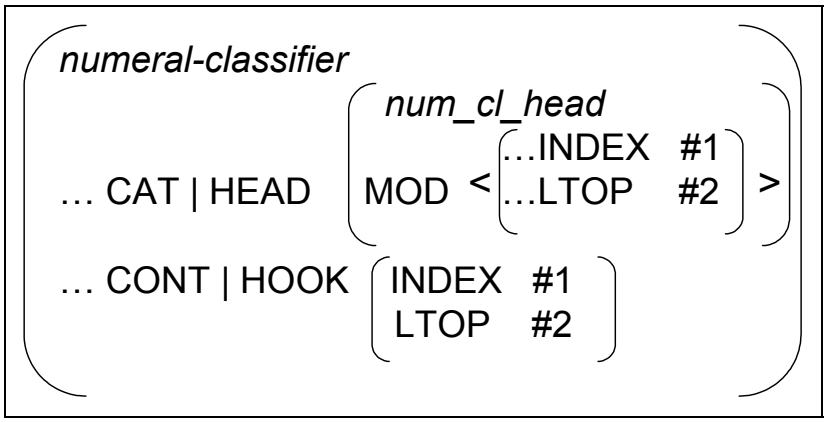


Figure 64

The constraints on the types **spr-only-num-cl-lex**, **obj-only-num-cl-lex** and **spr-obj-num-cl-lex** account for the position of the numeral classifier with respect to the number name and for the potential presence of *han*. Both the number name (a phrase of head type **int_head**) and *han* (given the distinguished head value **han_head**) are treated as dependents of the numeral classifier expression, but variously as specifiers or complements according to the type. In the JACY grammar, specifiers immediately precede their heads, while complements are not required to do so and can even follow their heads (in rare cases). Given all this, in the ordinary case (**spr-only-numcl-lex**), we treat the number name as the specifier of the numeral classifier. The other two cases involve numeral classifiers taking complements: with no specifier, in the case of pre-number unit expressions like the symbol $ (**obj-only-num-cl-lex**) and both a number-name specifier and the complement *han* in the case of unit expressions appearing with *han* (**spr-obj-num-cl-lex**).[28] Finally, the type **spr-obj-num-cl-lex** does some semantic work as well, providing the **plus_rel** which relates the value of the number name to the "half" contributed by *han*, and identifying the ARG1 of the **plus_rel** with the XARG the SPR and COMPS so that they will all share an index argument (eventually the index of the modified noun for sortal classifiers and of the measure noun relation for mensural classifiers).

[28] Because numeral classifiers are analyzed as taking posthead complements in these two cases, the head type **numcl_head** is a subtype of **init-head**, which contrasts with **final_head**. These types are used by the head-complement rules to determine the order of the head and complements.

The constraints which implement these aspects of our analysis are sketched in Figure 65–Figure 67.

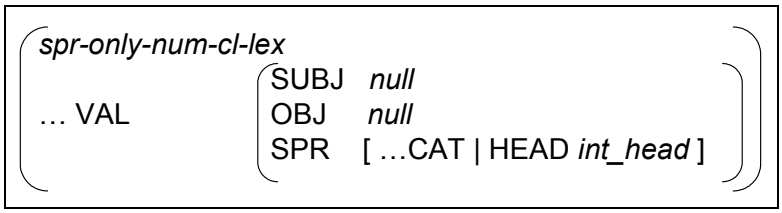


Figure 65

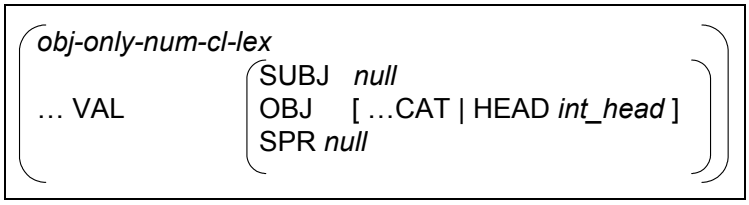


Figure 66

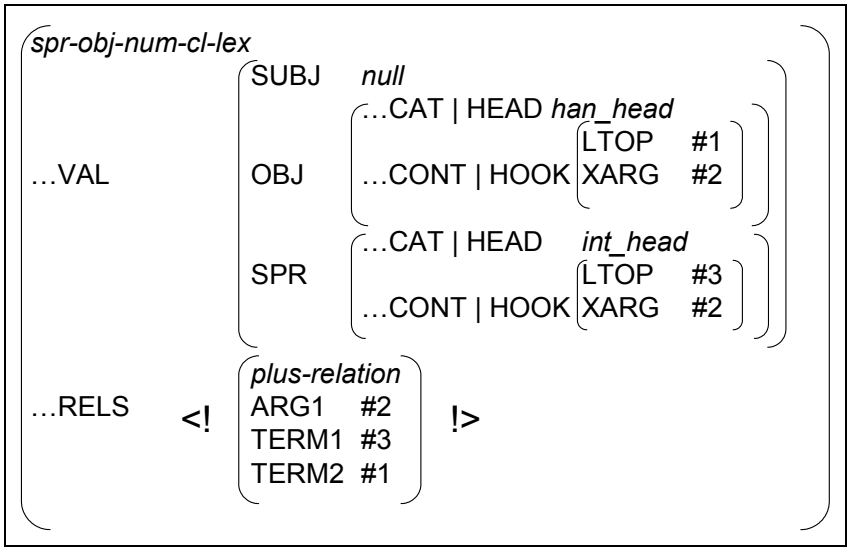


Figure 67

In the second dimension of the cross-classification, **anymod-num-cl-lex** and **noun-modnum-cl-lex** constrain what the numeral classifier may modify, via the MOD value. When numeral classifiers appear before the head noun, they are linked to it with *no*, which mediates the modifier-modifiee relationship. However, numeral classifiers can appear after the noun (Example 80), modifying it directly. Some numeral classifiers can also 'float' outside the NP, either immediately after the case postposition or to the position before the verb (Example 81).[29] We handle the former type by allowing most numeral classifiers to appear as post-head modifiers of PPs. Thus **noun-mod-num-cl-lex** further constrains the HEAD value of the

[29]Those that can't include expressions like *gou* in (i), cf. (ii):

(i) kouza 1234 gou wo tojitai
account 1234 number ACC close.volitional

((I) want to close account number 1234.)

(ii) *kouza wo 1234 gou tojitai
account ACC 1234 number close.volitional

element on the MOD list to be **noun_head**, but **anymod-num-cl-lex** leaves it as inherited (**noun-or-case-p head**). This type does, however, constrain the modifier to show up after the head ([POSTHEAD **right**]), and further constrains the modified head to be [NUCL **nucl_plus**], in order to rule out vacuous attachment ambiguities between numeral classifiers attaching to the right and other modifiers appearing to the left of the NP.

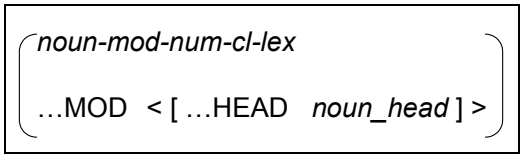


Figure 68

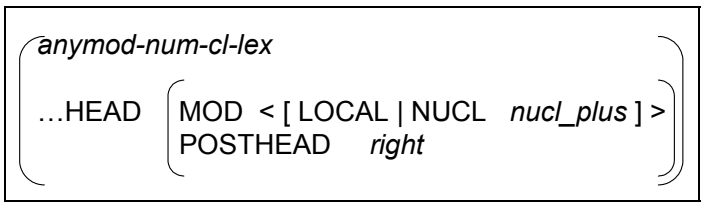


Figure 69

There are two types of floating numeral classifiers: One that counts the verbal subject and one that counts the verbal complement.

Example 89: Floated numeral classifier counting the verbal object

友達	を	3	人	待って	います。
tomodachi	wo	san	nin	matte	imasu
friends	ACC	three	persons	wait	progressive

(I am waiting for three friends.)

Example 90: Floated numeral classifiers counting the verbal subject

友達	が	3	人	待って	います。
tomodachi	ga	san	nin	matte	imasu
friends	NOM	three	persons	wait	progressive

(Three friends are waiting for me.)

Two lexical rules take a numeral classifier and turn it into a type that modifies the verb, but counts the verbal external argument or the first complement:

`numeral-classifier-sbj-float`
`numeral-classifier-obj-float`

The rules are restricted to arguments that are not subject to zero pronominalization.

The lexical rules turn the numeral classifier into a lexical type for floated numeral classifiers.

This is its head:

```
num-cl-float_head := num-cl-mod_head &
                     [ MOD < [LOCAL scopal_mod & [CAT.HEAD verb_head]] >,
                               EMPTY -].
```

Figure 70

The floated classifier modifies a verb, which shouldn't be empty.

There are two types of floated numeral classifiers, one that counts the verbal subject and one that counts the verbal first complement.

The type for floated numeral classifiers identifies the XARG of its specifier (the number) with its own XARG. This type is supertype to subject counting and object counting types.

The type counting the verbal subject identifies the XARG of the modified verb with its specifier's XARG. The type counting the verbal first complement identifies the INDEX of the first argument of the modified verb with its specifier's XARG. In all cases, the counting is restricted to open (non-zero pronoun) arguments.

```
floated-num-cl-lex := numeral-classifier &
[ SYNSEM.LOCAL [CAT [HEAD num-cl-float_head &
                     [MOD < [ LOCAL.CONT.HOOK [LTOP #top,
                                               INDEX #ind]] >],
                 VAL [SPR.FIRST.LOCAL.CONT.HOOK [XARG #xarg],
                      COMPS < >,
                      SUBJ < >]],
               CONT [HOOK [LTOP #top,
                           INDEX #ind,
                           XARG #xarg],
                     HCONS <! !>]]].
```

Figure 71

```
floated-ind-sbj-num-cl-lex := floated-num-cl-lex &
[ SYNSEM.LOCAL [CAT [HEAD [MOD < [ LOCAL.CONT.HOOK.XARG #xarg & full_ref-
ind] >],
                     VAL [SPR.FIRST.LOCAL.CONT.HOOK [XARG #xarg]]]]].


floated-ind-obj-num-cl-lex := floated-num-cl-lex &
[ SYNSEM.LOCAL [CAT [HEAD [MOD <[ LOCAL.CAT.VAL.COMPS.FIRST.LOCAL.CONT.HOOK
                                  [INDEX #xarg & full_ref-ind]] >],
                     VAL [SPR.FIRST.LOCAL.CONT.HOOK [XARG #xarg]]]]].
```

Figure 72

Floated numeral classifiers should not go into the nominal-numcl-rule, as this would cause spurious ambiguity. Therefore, we use the feature EMTPY: The nominal-numcl-rule requires its argument to be EMPTY +, while the num-cl-float_head is EMPTY -. The type for the special "no" that is used in the case of numeral classifiers modifying the counted noun requires its complement to be EMPTY +.

The MRS for sentences with floated numeral classifiers reflect the fact that these modify verbs and count their arguments:

```
h1,e2:INDICATIVE:PRESENT:PROGRESSIVE,
h1:proposition_m(e2, h3),
h4:_tomodachi_n(x5:THREE:GENDER),
h6:udef(x5, h7),
h9:card(x5, "3"),
h11:_matsu_v(e2, u12, x5)},
h3 qeq h11,
h7 qeq h4
```

Figure 73

The final dimension of the classification captures the semantic differences between sortal and mensural numeral classifiers. The sortal numeral classifiers contribute no semantic content of their own.[30] They are therefore constrained to have empty RELS and HCONS lists:

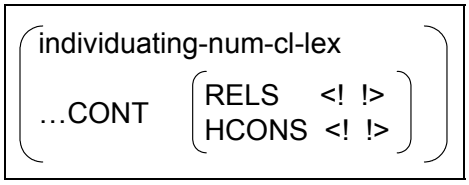
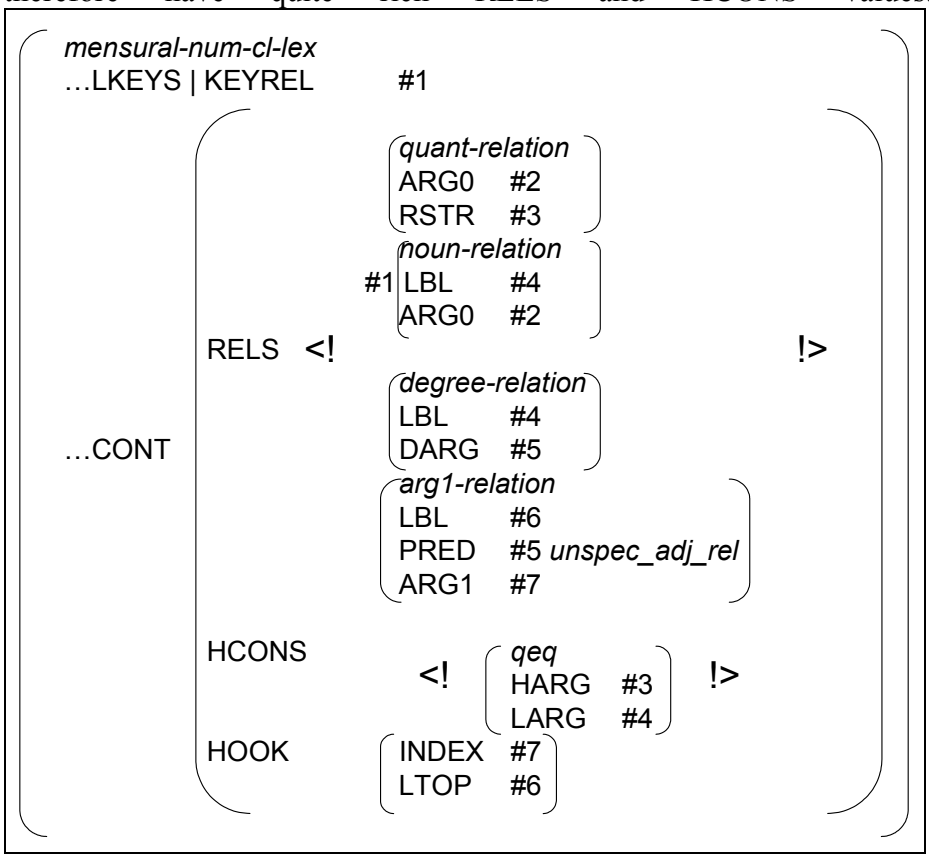


Figure 74

In contrast, mensural numeral classifiers contribute quite a bit of semantic information, and therefore have quite rich RELS and HCONS values. As shown in



[30] The individuating function they serve we take to be implicit in the linkage they provide between the **card_rel** and the noun relation.

Figure 75, the **noun-relation** is identified with the lexical key relation value (LKEYS.KEYREL) so that specific lexical entries of this type can easily further specify it (e.g., *kiro* constraints its PRED to be **kilogram_n_rel**).

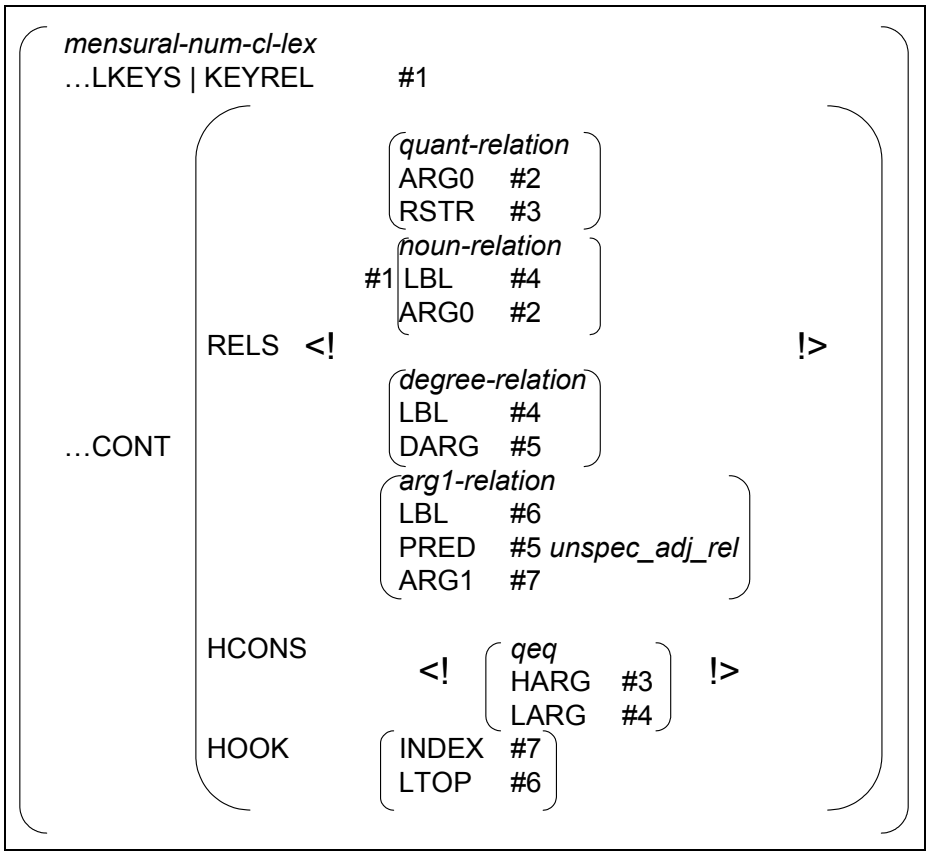


Figure 75

The type also makes reference to the HOOK value so that the INDEX and LTOP (also the INDEX and LTOP of the modified noun, see Figure 64) can be identified with the appropriate values inside the RELS list. The length of the RELS list is left unbounded, because some mensural classifiers also inherit from **spr-obj-num-cl-lex**, and therefore must be able to add the **plus_rel** to the list.

The types in the bottom part of the hierarchy in Figure 63 join the dimensions of classification. They also do a little semantic work, making the INDEX and LTOP of the modified noun available to their number name argument, and, in the case of subtypes of **mensural-num-cl-lex**, they constrain the final length of the RELS list, as appropriate.

5.6.5 The linker no

We posit a special lexical entry for *no* which mediates the relationship between NumClPs and the nouns they modify. In addition to the constraints that it shares with other entries for *no* and other modifier heading postpositions (see Section 6.4.3.1), this special *no* is subject to the constraints shown in Figure 76. These specify that *no* makes no semantic contribution, that it takes a NumClP as a complement, and that the element on the MOD list of *no* shares its local top handle and index with the element on the MOD list of the NumClP (i.e., that *no* effectively inherits its complement's MOD possibility). Even though (most) numeral classifiers can either modify NPs or PPs, all entries for *no* are independently constrained to only modify NPs, and only as pre-head modifiers.

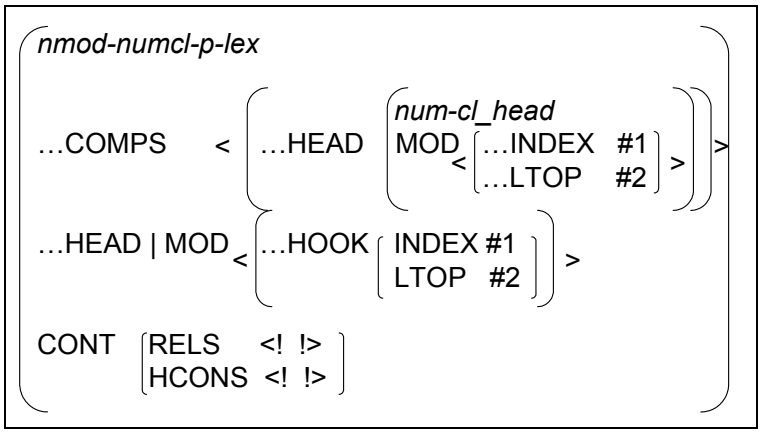


Figure 76: nmod-numcl-p-lex

5.6.6 Examples: NumCLPs as modifiers

We illustrate our analysis with sample derivations, displayed as trees with (abbreviated) rule names and lexical types on the nodes. Figure 77 corresponds to Example 79, Figure 78 to Example 80, and Figure 79 to a shortened Example 81.

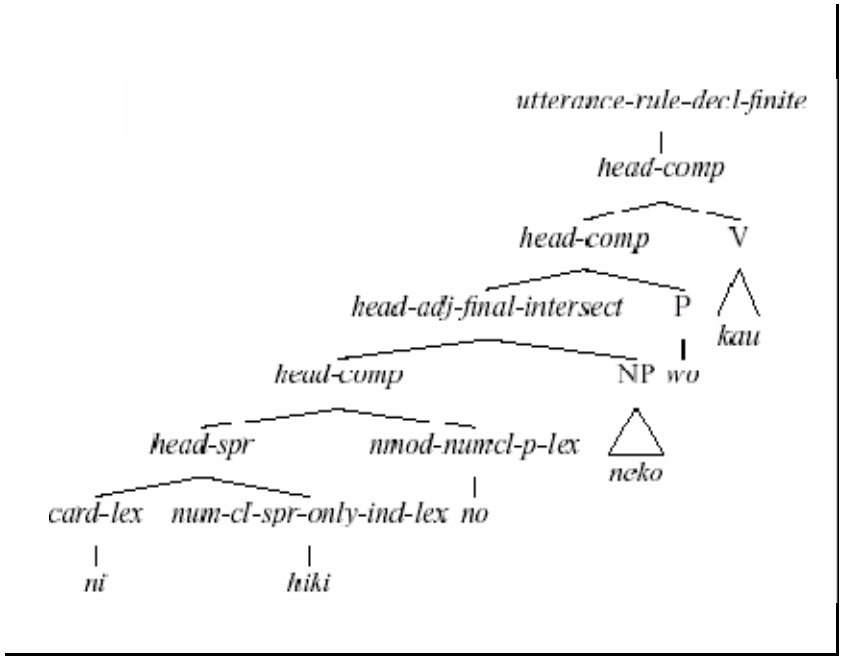


Figure 77: Tree for *ni hiki no neko wo kau*

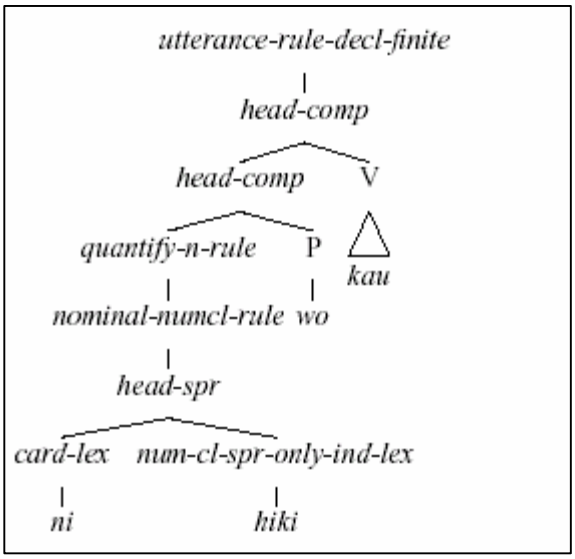


Figure 78: Tree for *neko ni hiki wo kau*

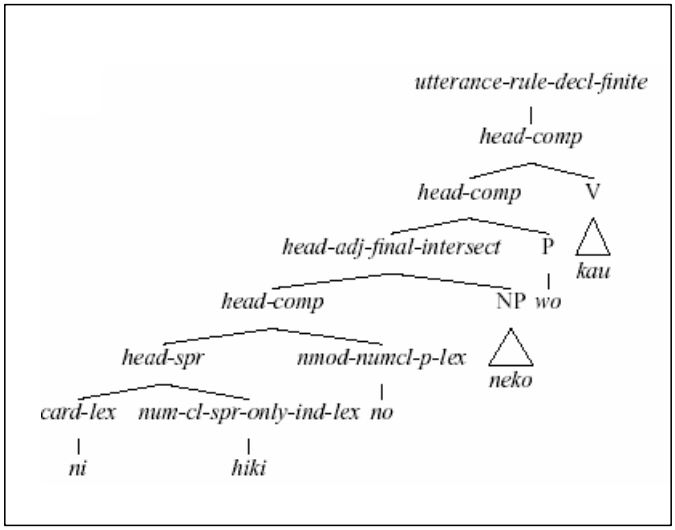


Figure 79: Tree for *neko wo ni hiki wo kau*

5.6.7 Unary-branching phrase structure rule

We treat NumClPs serving as nominal constituents by means of an exocentric unary-branching rule.[31]

This rule specifies that the mother is a noun subcategorized for a determiner specifier (these constraints are expressed on **noun_sc**), while the daughter is a numeral classifier phrase whose valence is saturated.

Furthermore, it contributes (via its C-CONT, or constructional content feature) an underspecified **noun-relation** which serves as the thing (semantically) modified by the numeral classifier phrase. The reentrancies required to represent this modification are implemented via the LTOP and INDEX features.

[31] In the analysis of number names used as NumClPs, we posit a second unary-branching rule. The mother of that rule (a NumClP) can then serve as the daughter of the rule discussed here.

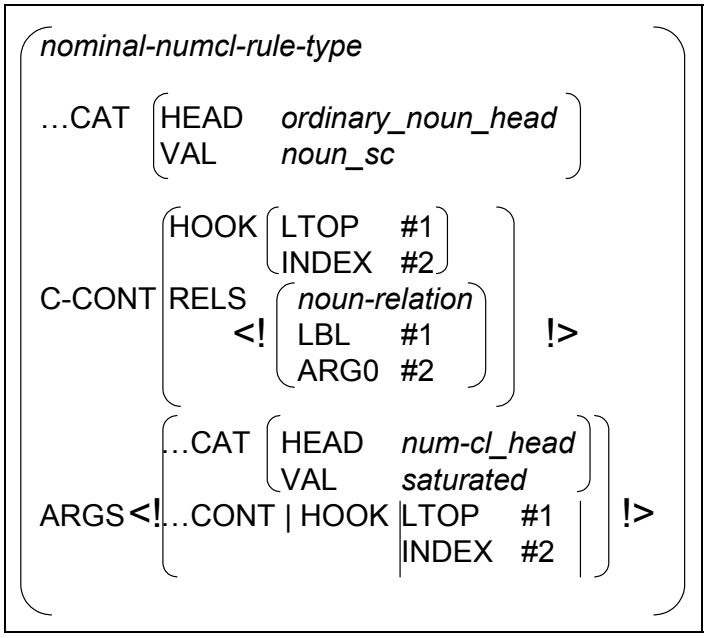


Figure 80: Nominal-numcl-rule-type

This rule works for both sortal and mensural NumClPs, as both are expecting to modify a noun.

5.6.8 Examples: NumClPs as nouns

Again, we illustrate the interaction of these various constraints with an example derivation (Figure 81) for Example 78.

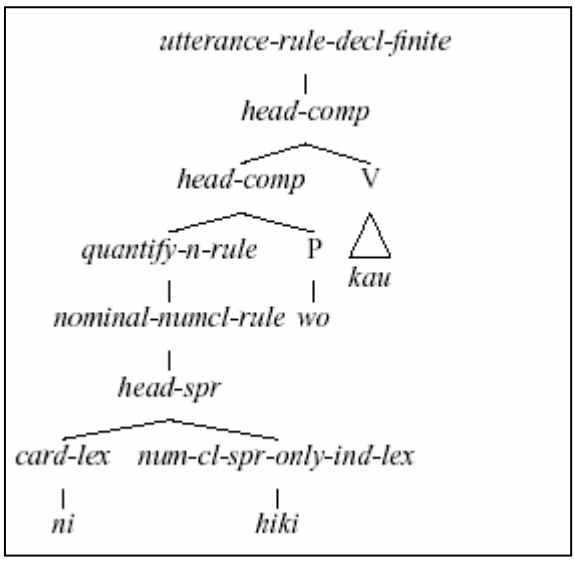


Figure 81: Tree for *ni hiki wo kau*

5.7 Noun modification

Typically, Japanese nouns are modified by other noun phrases via the particle *no*, as in Example 91. The particle *no* inserts a relation to the MRS, which combines the indices of the nouns via its argument structure, as can be seen in Figure 82.

Example 91

私　　の　本
watashi　no　hon
I　　NO　book

(my book)

```
h4:pron_rel(x5)
h6:def_rel(x5,h7)
h9:_no_p_rel(e11, x5, x10)
h9:_hon_n_rel(x10)
h12:udef_rel(x10,h13)
qeq (h7,h4)
qeq (h13,h9)
```

Figure 82: MRS of *watashi no hon*

With relational nouns like *ue* (above), *shita* (under), *kita* (north) etc. there is the possibility of noun modification using *kara* or *yori*, as in Example 92.

Example 92

大阪　は　東京　から　南　　だ
Osaka　wa　Tokyo　kara　minami　da
Osaka　TOP　Tokyo　from　south　COP

(Osaka is south of Tokyo)

The entry for this particle is of the same lexical type as the *no* entry, restricted to relational nouns (their head type being a subtype of **noun_head**).

There are some noun modifiers that occur before the noun. Examples of these are *purasu* (plus), *onaji* (the same), *ironna* (various), *ichiban* (most), as in Example 93.

Example 93: *onaji* modifiying a noun

あの　人　　は　同じ　本　　を　読んで　います
ano　hito　　wa　onaji　hon　wo　yonde　imasu
that　person　TOP　same　book　ACC　read　progr.

(That person is reading the same book.)

There are, though, modifiers that occur between noun and particle, such that we have examples for head-initial modification. Examples of these are *nado* (and so on) and *kurai* (about). The modifier *dake* 'only' occurs between nouns and case particles, as for example in:

Example 94

野村さん　　だけ　が　　来た
Nomura-san　dake　ga　　kita
Ms. Nomura　only　CASE　came

(Only Ms. Nomura came)

The head of the construction *Nomura-san dake ga* is the case particle *ga*, because the verb *kita* selects for a subject marked by *ga* and therefore *ga* contributes the information for syntactic selection. The head of *Nomura-san dake* must be *Nomura-san,* because *ga* selects for a noun. Leaving *dake* out in this construction leads to a grammatical sentence *Nomura-san*

ga kita, while leaving *Nomura-san* or the case particle out, leads to ungrammatical sentences[32]. Therefore we conclude that *dake* in this construction is a modifier to *Nomura-san*; and we have a good example of head-initial noun modification.

The lexical type **noun_mod-lex** contains the possibility for noun modification in:

[SYNSEM|LOCAL|CAT|HEAD|MOD|LOCAL|CAT|HEAD noun_head]

The lexical entries of the noun modifiers get the information about right (head-initial) or left (head-final) modification in their HEAD:

[POSTHEAD left]; or

[POSTHEAD right].

The grammar rules for head-final and head-initial modification then refer to this POSTHEAD information (see Chapter 8 for the treatment of head-initial and head-final modification).

5.8 Relative sentence constructions

Noun phrases can be modified by predicates. This builds the Japanese relative sentence construction. The noun modification with a relative sentence is possible for verbs without addressee honorification, as in Example 95.

Example 95: From the Verbmobil corpus

nakanaka	aite	iru	jikaN	ga	arimaseN	node
more and more	free	be	time	NOM	not exist	SAP

(There is less and less free time.)

Research literature (such as Uda 2001) gives these constructions the name "externally-headed relative clauses", as opposed to "internally-headed relative clauses" which we described in Chapter 5.4 as nominalization structures. Uda (2001) explains the semantic difference between these structures in the restrictiveness of modification, such that "*only the EHRC appropriately restricts the target*" (page 205).

The problem of argument binding in relative clauses is quite similar to the problem of topics in Japanese: The noun that is modified by the relative clause can fill a subject position in the argument structure of the verb (see Example 97), a complement position (see Example 98), or no position at all (see Example 96). Sirai and Gunji (1998) designate these as "internal relationship" and "external relationship" (page 17). Syntax on its own cannot solve this ambiguity, such that we have an example of systematic ambiguity[33]. The decision can in many cases be made by selectional restrictions and world knowledge, which we view to be external to the HPSG grammar. A solution can be not to try to connect the head noun to the argument structure of the relative clause verb, thus leaving the decision to further NLP components that have access to world knowledge and selectional restrictions. Relative clauses are then seen as adjuncts to the head noun in all cases. This treatment is seen in the tradition of underspecification, as for example scope representation in MRS is.

Another possibility is to give ambiguous readings and select the best one based on treebanking with relevant examples in a relevant domain, using the result for stochastic disambiguation, as is described by Oepen et al. (2002b). Treebanking using JACY is already being done at NTT Japan (see Bond et al. 2004b).

---

[32] Although particle omissions can occur in spoken language.

[33] Baldwin (2004) states after a data evaluation of 5143 relative clause instances from the EDR corpus that in 64.0% of cases the subject gap interpretation is the correct one.

Example 96: Adjunctive relative clause

魚	を	焼く	如意
sakana	wo	yaku	niyoi
Fish	ACC	grill	smell

(The smell of grilling fish)

Example 97: Relative clause with a subject head noun

本	を	読んだ	人
hon	wo	yonda	hito
book	ACC	read	person

(The person that read the book.)

Example 98: Relative clause with a complement head noun

人	が	読んだ	本
hito	ga	yonda	hon
person	NOM	read	book

(The book that a person read.)

Sirai and Gunji (1998) give an approach for the relative clauses with internal relationship to the head noun which includes a lexical rule that builds up a SLASH list with subcategorized arguments of the head verb, such that elements on the SLASH list can be bound by the head noun. We opt for a more direct approach, circumventing the building of SLASH lists, as we did for the treatment of zero pronouns (see Section 3.3).

In order to allow for disambiguation by treebanking, we added four possibilities for relative sentence constructions (and set a switch for the root node, which allows for a non-ambiguous parsing of these). All relative clause rules are subtypes of head-final intersective modification rules. They add a proposition on top of the relative clause verb to the MRS, using C-CONT. The *relative-clause-rule* views the relative clause as a pure modification of the head noun and does not give the head noun any role in the verb's argument structure. Using C-CONT, it adds a topic relation to the MRS that takes the verbal event and the head noun as its arguments and thus links the structures of relative sentence and head noun. The first argument (the relative sentence) of the rule is restricted to have the feature [POSTHEAD *rels*] in its HEAD. The POSTHEAD feature is determined by the verbal ending, such that plain endings like *ru* can undergo relative sentence constructions, while the *te* ending, for example, leads the sentence into a coordinated sentence construction. Figure 83 shows the MRS for Example 97 with the **relative-clause-rule** applied. It can be seen that the verbal relation **_yomu_v_rel** contains a zero pronoun u10 as its first semantic argument.

The **rel-cl-sbj-gap-rule** takes the index of the head noun and identifies it with the index of the subject of the relative clause verb. Figure 84 shows the MRS for Example 83 with the subject gap reading. It can be seen that the verbal relation **_yomu_v_rel** contains the index of **_hito_n_rel** as its first argument.

The **rel-cl-obj1-gap-rule** and the **rel-cl-obj2-gap-rule** do the same thing with the complement indices.

```
h4:_hon_n_rel (x5:THREE)
h6:udef_rel (x5, h7)
h9:_yomu_v_rel(e11::PAST:INDICATIVE, u10, x5)
h12:_hito_n_rel (x13:THREE)
h14:udef_rel (x13, h15)
h12:proposition_m_rel (h17)
h9: topic_rel (e18:NO_TENSE, e11, x13)
h7 qeq h4
h15 qeq h12
h17 qeq h9
```

Figure 83: MRS for *hon wo yonda hito*, adjunct reading

```
h4:_hon_n_rel (x5:SEMSORT:THREE:GENDER)
h6:udef_rel (x5, h7)
h9:_yomu_v_rel (e11:PAST:INDICATIVE,x10:THREE,x5)
h12:_hito_n_rel (x10)
h13:udef_rel (x10, h14)
h12:proposition_m_rel (h16)
h7 qeq h4
h14 qeq h12
h16 qeq h9
```

Figure 84: MRS for hon wo yonda hito, subject-gap reading

The nominative case inside of relative clauses can be changed to genitive, as in Example 99. In this case, a lexical rule is applied to the verbal stem that changes the case of the subcategorized subject noun.

Example 99: Ga-no conversion in relative clauses

田中	の	食べた	ご飯
Tanaka	no	tabeta	gohan
Tanaka	GEN	eat (past)	rice

(The rice that Tanaka ate)

5.9 Pre-nominal adjectives

There are two types of adjectives. The one type directly modifies nouns. It can also be used as a sentence predicative, with predicative inflections. This is analyzed as a subtype of verbs, as shown in Chapter 4. An example for this kind of noun modification can be seen in Example 100. The other type of adjectives needs the copula form *na* for noun modification.[34] They are subcategorized by *na* and cannot be used in a predicative way (see Example 102.

Example 100: Verbmobil example

いい	じかん	だ	と	思います
ii	jikan	da	to	omoimasu
good	time	COP	COMP	think

(I think this is a good time)

[34] See Nightingale (1996) for *na* as a copula construction.

Example 101

時間　が　いい
jikan　ga　ii
time　NOM　good

(The time is good.)

Example 102: Verbmobil example

きれい　な　ホテル　に　止まって　みたい
kirei　na　hoteru　ni　tomatte　mitai
beautiful　COP　hotel　LOC　stay　want to try

(I want to try to stay in a beautiful hotel.)

Example 103: Ungrammatical

*ホテル　が　きれい
hoteru　ga　kirei
hotel　NOM　beautiful

The predicative adjective modification of nouns as in Example 100 is treated as a relative sentence construction, just as described in Chapter 5.8. The na adjective is subcategorized for by the copula *na*. This modifies the head noun, such that a relative sentence modification takes place as well. In both cases there is ambiguity between the adjunctive relative clause and the subject-gap relative clause constructions. So the MRS for the *na* adjective example looks just the same as the MRS for the predicative adjective example, reflecting the semantic parallelism of the constructions.

```
h4: _kirei_a / _ii_a (e6,x5)
h7: _hoteru_n(x5)
h8: udef(x5,h9)
h7: proposition_m(h11)
h11 qeq h4
h9 qeq h7
```

Figure 85: MRS for kirei na hoteru / ii hoteru

6 Particles

The treatment of particles is essential for the processing of Japanese language for two reasons. The first reason is that these are the words that occur most frequently. In 800 Japanese dialogues on appointment scheduling, the particle *wa* occurs 5765 times, *ga* occurs 5909 times, *ni* occurs 4358 times, *kara* occurs 2802 times and *made* occurs 1158 times, while the noun *kaigi* (which means *meeting* and is therefore essential for appointment scheduling dialogues) occurs only 792 times. The second reason is that particles have various central functions in the Japanese syntax:

- Case particles mark subcategorized verbal arguments.
- Postpositions mark adjuncts and have semantic attributes.
- Topic particles mark topic adjuncts or topicalized verbal arguments.
- *no* marks an attributive nominal adjunct.

Their treatment is difficult for three reasons:

- Despite their central position in Japanese syntax, the omitting of particles occurs quite often in spoken language.
- One particle can fulfil more than one function.
- Particles can co-occur, but not in an arbitrary way.

In order to set up a grammar that accounts for a large amount of spoken language, a comprehensive investigation of Japanese particles is thus necessary. Such a comprehensive investigation of Japanese particles (and its implementation in an HPSG grammar) was missing up to now.[35] Two kinds of solutions have previously been proposed:

1. Particles are divided into case particles and postpositions. The latter build the heads of their phrases, while the former do not (cf. Miyagawa 1986, Tsujimura 1996).
2. All kinds of particles build the head of their phrases and have the same lexical structure (cf. Gunji 1987)

Both kinds of analyses lead to problems: If postpositions are heads, while case particles are non-heads, a sufficient treatment of those cases where two or three particles occur sequentially is not possible, as we will show. If on the other hand there is no distinction of particles, it is not possible to encode their different behaviour in subcategorization and modification.

We carried out an empirical investigation of co-occurrences of particles in Japanese spoken language. We show that the problem is essentially based on the lexical level. Instead of assuming different phrase structure rules for sentences with different types of particles we state a type hierarchy of Japanese particles. This allows a uniform treatment of phrase structure as well as a differentiation of subcategorization patterns. We therefore adopt the `all-head' analysis, but extend it with a type hierarchy in order to be able to differentiate between the particles.

[35] Pollard and Sag (1994) mention a manuscript that was written by Tomabechi in 1989 that seems not to be available any more.

6.1 Co-occurrence of particles

Japanese noun phrases can be modified by more than one particle at a time. There are many examples in our data where two or three particles occur sequentially. On the one hand, this phenomenon must be accounted for in order to attain a correct processing of the data. On the other hand, the discrimination of particles is motivated by their modificational and subcategorizational behaviour. The analysis that we describe in this section is based on a large amount of dialogue data: The 800 Japanese dialogues concerning appointment scheduling that were collected and transcribed in the Verbmobil project, which dealt with English, German and Japanese machine translation (see Wahlster 2000 for further information.). We carried out an empirical analysis, based on this dialogue data. Table 10 shows the frequency of co-occurrence of two particles in the dialogue data. Table 11 shows the frequency of co-occurrence of three particles.

The co-occurrence of *wa* and *de mo* in two cases of the dialogue data are cases of *wa demo*, where *demo* functions as an adverb, rather than as a particle, e.g.:

Example 104

火曜日	は	でも	一日	開いて	います	ね
kayoubi	wa	demo	ichiNchi	aite	imasu	ne
Tuesday	WA	also	whole day	free	AUX	tag

(Also on Tuesday, the whole day is free.)

The same applies to *wa nanka*, where *nanka* occurs in its function as an adverb:

Example 105

来週	は	なんか	うまちゃって	いる	ｎです	けども
raishuu	wa	nanka	umachatte	iru	N desu	kedomo
next week	WA	somehow	occupied	AUX	COP	SAP

(Next week is somehow occupied.)

Table 10: Co-occurrence of two particles in the 800 Verbmobil dialogues[36]

left ↓ right →	ga	wo	ni	de	e	kara	made	wa	mo	nanka	to
ga	0	0	0	0	0	0	0	0	0	0	0
wo	0	0	0	0	0	0	0	0	0	0	3
ni	0	0	0	19	0	0	0	137	49	0	15
de	2	0	0	0	0	0	0	158	241	0	30
e	0	0	0	1	0	0	0	1	0	0	0
kara	23	0	30	81	0	0	0	69	12	0	123
made	17	1	66	32	0	0	0	63	1	0	79
mo	0	0	0	0	0	0	0	0	0	0	0
nanka	3	0	0	1	0	0	0	30	0	0	0
to	0	0	0	1	0	0	0	17	58	0	0

[36] We have not taken *no* into account here, because *no* is ambiguous between nominalization and particle and occurs very often in both functions.

toshite	0	0	0	0	0	0	0	36	15	0	0
toshimashite	0	0	0	0	0	0	0	15	0	0	0
wa	0	0	0	0	0	0	0	0	0	1	1

Table 11: Co-occurrence of 3 particles in the 800 Verbmobil dialogues

left ↓ right →	de mo	de wa	ni wa
ga	0	0	0
wo	0	0	0
ni	15	4	0
de	2	0	0
kara	12	5	0
made	2	1	16
wa	2	0	0
mo	0	0	0
nanka	0	1	0

ga can follow the particles *de*, *kara*, *made* and *nanka*. *kara ga* and *made ga* occur quite often in the dialogue data, but there are no examples of the other particles.

The dialogue data shows that combinations with *wo* occur quite seldom, we found only one example of *wo* following *made*.

The dialogue data shows that there are several occurrences of *kara ni* and *made ni*, but no examples of other co-occurrences. Here is an example for *kara ni*:

Example 106

何時	ぐらい	から	に	します	か
nanji	gurai	kara	ni	shimasu	ka
what time	about	KARA	NI	do	QUE

(At about what time shall we start?)

de can follow verb-modifying particles in its case marking function. *ni de*, *kara de,* and *made de* occur quite often in the dialogue data. Here is an example:

Example 107

三時	ぐらい	から	で	よろしい	でしょう	か
saNji	gurai	kara	de	yoroshii	deshou	ka
3 o'clock	about	KARA	DE	good	COP	QUE

(Would about 3 o'clock suit you?)

In their modifying function, *de* and *ni* can follow particles like *kara*, *made*, *nanka* and *toshite*; in their case marking function they can follow different kinds of particles. It is in principle even possible to have the co-occurrence of the case particle de (respectively *ni*) with its modifying counterpart:

Example 108

東京　で　で　いかが　でしょう　か
Tokyo　de　de　ikaga　deshou　ka
Tokyo　DE　DE　good　COP　QUE

(Would it suit you (to meet in) Tokyo?)

There is a tendency to avoid the co-occurrence of particles with the same phonology, even if it is possible on principal in some cases. The reason is obvious: Such sentences are difficult to understand. *kara* as well as *made, nanka* and *e* cannot follow any other particles. *wa* does not follow *ga, wo, e* or *mo*, but all other kinds of (analyzed) particles. *mo* behaves like *wa*, except that it did not follow *nanka*.

In some case three particles occur in a row, as for example:

Example 109

五時　ごろ　まで　に　は　お電話　さしあげます　ので
goji　goro　made　ni　wa　odeNwa　sashiagemasu　node
5 o'clock　about　MADE　NI　WA　telephone　do　SAP

(I will phone you before about 5 o'clock.)

The reason is that *wa* can follow *ni*. This again can follow *made*. Another linearization like e.g. *made-wa-ni* or *ni-made-wa* would not be possible. Table 11 shows the frequency of co-occurrence of three particles in the dialogues.

A first classification based on these co-occurrence results can be seen in Table 12.

Table 12: A first classification based on co-occurrence

left ↓ right →	case particle	postposition	adverbial particle	topic particle
case particle	-	-	-	-
postposition	+	-	+	+
adverbial particle	-	-	-	+
topic particle	-	-	-	-

6.2 The type hierarchy of Japanese particles

Kuno (1973) treats *wa, ga, wo, ni, de, to, made, kara* and *ya* as `particles'. They are divided into those that are in the deep structure and those that are introduced through transformations. An example for the former is *kara*, examples for the latter are *ga* (SBJ), *wo* (OBJ), *ga* (OBJ) and *ni* (OBJ2).

Gunji (1987) assigns all particles the part-of-speech P. Examples are *ga, wo, ni, no, de, e, kara* and *made*. All particles are heads of their phrases. Verbal arguments get a grammatical relation [GR OBJ/SBJ]. In Gunji (1991) though, the part-of-speech class P contains only *ga*, *wo* and *ni*:

"For example, the class of postpositionals only include particles that indicate grammatical relations `subject' and `object'. Thus, only ga, wo and ni are in this class. Other particles traditionally called postpositions (`zyosi') are classified as either an adnominal (e.g., the possessive no), or an adverbial (e.g., the locative de)."

Tsujimura (1996) defines postpositions and case particles:

"Postpositions are the Japanese counterpart of prepositions in English, and as the term indicates, postpositions are placed after nouns while prepositions occur before nouns. ... Postpositions cannot stand independently.

Case particles include Nominative (Nom)-ga, Accusative (Acc)-wo, Dative (Dat)-ni, and Genitive (Gen)-no, and to these we add the Topic(Top) marker -wa. ... Case particles can follow postpositions although particles following nouns comprise a far more general pattern."

Nightingale (1996) divides case markers (*ga, wo, ni* and *wa*) from copula forms (*ni, de, na* and *no*). He argues that *ni, de, na* and *no* are the infinitive, gerund and adnominal forms of the copula.

We assume a common type of particles, which gives us the possibility to state general restrictions on particles as well as restrictions on sub-types to them (see Figure 86 for the type hierarchy under **p-lex**). This general particle type is divided into case particles (**case-p-lex**) and other particles (**p-lex-c**). Case particles assign a case to the argument they take, others do not. We assume not only a differentiation between case particles and postpositions, but a finer graded distinction that includes different kinds of particles not mentioned by the other authors. *de* is assumed to be a particle, and not a copula, as Nightingale proposes. It belongs to the class of adverbial particles. One major motivation for the type hierarchy is the observation we made of the co-occurrence of particles.

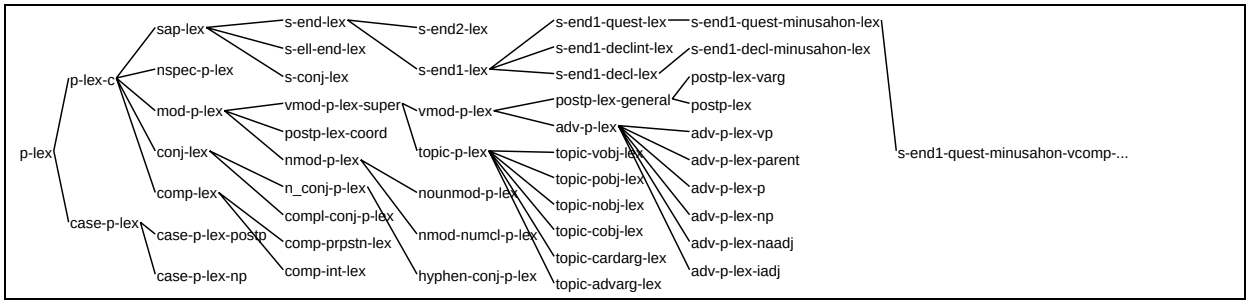
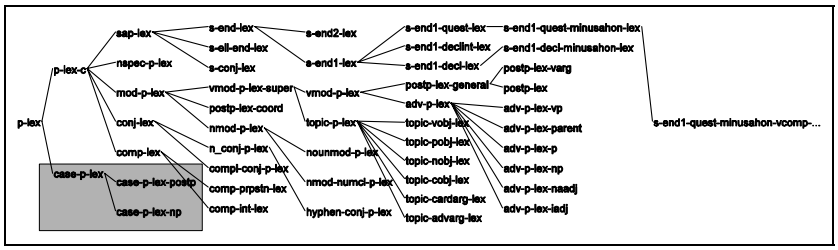


Figure 86: Type hierarchy of Japanese particles

Case particles are those that attach to verbal arguments and assign case, which can be subcategorized for. A complementizer (**comp-lex**) marks complement sentences. Modifying particles (**mod-p-lex**) attach to adjuncts. They are further divided into noun-modifying particles (**nmod-p-lex**), verb-modifying particles (**vmod-p-lex-super**) and others. Verb modifying particles can be topic particles (**topic-vmod-p-lex**), adverbial particles (**adv-p-lex**), or postpositions (**postp-lex-general**). Some particles can have more than one function; as for example *ni* has the function of a case particle and an adverbial particle.

The next sections feature the individual types of particles.

6.3 Case particles



Case particles show the case of the phrase they head, which is selected for by subcategorization. They add no further semantics. The type of case particles contains two sub-types: **case-p-lex-np** and **case-p-lex-postp**. This division is motivated by the arguments they

can take: **case-p-lex-np** type particles take noun phrases and **case-p-lex-postp** attach to other particles of the postposition type. Table 13 shows the division of particles to the case particle types.

case-p-lex-np	case-p-lex-postp
が (ga)	が(ga)
を (wo)	を(wo)
に (ni)	に(ni)
と (to)	
だけ (dake)	
は (wa)	
の (no)	

Table 13: Case particles

There is neither number nor gender agreement between the subcategorized noun phrase and the verb. The verbs subcategorize for case marked entities. Case is assigned by the case particles. Therefore these have a syntactic function, but not a semantic one. Different from English, the grammatical functions cannot be assigned through positions in the sentence or c-command-relations, since Japanese knows no fixed word position for verbal arguments. Hence, the following variations are possible, for example:

Example 110

花子	が	本	を	買います
Hanako	ga	hon	wo	kaimasu
Hanako	NOM	book	ACC	buy

(Hanako buys a book.)

Example 111

本	を	花子	が	買います
hon	wo	Hanako	ga	kaimasu
book	ACC	Hanako	NOM	buy

(Hanako buys a book.)

The assignment of the grammatical function is not achieved by the case particle alone but only in connection with the verbal valence. There are verbs that require *ga*-marked objects, while in most cases the *ga*-marked argument is the subject:

Example 112

なんとか	予定	が	取れる	んです	が
nantoka	yotei	ga	toreru	N desu	ga
somehow	time	NOM	can take	COP	SAP

(Somehow (I) can find some time.)

Japanese is described as a head-final language. Gunji (1987) therefore assumes only one phrase structure rule:

$$Mother \rightarrow Daughter \quad Head$$

However, research literature questions whether this also applies to nominal phrases and their case particles. Pollard and Sag (1994:45) assume Japanese case particles to be markers. Miyagawa (1986) makes a phrase-structural distinction between case particles and

`postpositions': While ‘postpositions’ are assumed to be heads, case particles are not. He gives two arguments for this assumption. The first is that a distinction between case particles and `postpositions' is semantically necessary, because the case particles assign no theta-role to the marked NPs. We follow the argument that a distinction is necessary and distinguish case particles and other particles in the type hierarchy, although we do not follow the idea to distinguish them on the phrase structural level.

The second argument concerns the numeral classifiers. They can occur within or outside the NP+case particle (called `NP' by Miyagawa) which they classify. But they cannot occur outside of an NP+`postposition' (called `PP' by Miyagawa):

Example 113[37]

a.	学生	三人	が	本	を	読んだ
	gakusei	sannin	ga	hon	wo	yonda
	students	3-NK	NOM	book	ACC	read(PAST)

(Three students read a book.)

b.	*人	が	小さい	村	から	二つ	来た
	hito	ga	chiisai	mura	kara	futatsu	kita
	people	NOM	small	village	from	2	come(PAST)

c.	人	が	二つ	の	小さい	村	から	来た
	hito	ga	futatsu	no	chiisai	mura	kara	kita
	people	NOM	2	GEN	small	village	from	come(PAST)

(People came from two small villages.)

d.	先生	が	三人	来た
	sensei	ga	sannin	kita
	teacher	NOM	3-CL	come(PAST)

(Three teachers came.)

The restriction that Miyagawa (1986:162) sets up is based on phrase structure:

> *Definition: X is bijacent to NP, iff:*
>
> *X is a sister to NP, or*
>
> *X is immediately dominated by a sister of NP.*

His restriction for numeral classifiers says that the classifier must be bijacent to the antecedent. Thus, every structure in which the antecedent of the numeral classifier is embedded in a PP is excluded.

Bijacency is however not a sufficient restriction for numeral classifiers, as the following example from Gunji and Hasida (1998b) shows, where the numeral classifier refers to the subject, while a complement is between the two entities:

[37] b) and c) from Miyagawa(1986:162), d) from Miyagawa (1986:157).

Example 114

去年	は	アメリカ人	が	日本	を	３０００人	訪れた
kyonen	wa	Amerikajin	ga	Nihon	wo	3,0000-nin	otozureta
last year	TOP	Americans	NOM	Japan	ACC	30.000 persons	visit(PAST)

(Last year, 30.000 Americans visited Japan.)

It is not possible to set up adequate restrictions on an (exclusively) syntactic base. The phrase-structural distinction between case-marked nominal phrases and nominal phrases marked with modifying particles does not further help here. Different restrictions for numeral classification with case particles and postpositions support our claim that they must be distinguished, but not necessarily on the phrase structural level. Gunji and Hasida (1998b) show that instead of syntactic restrictions for numeral classifiers, semantic ones should be used. They use the notions of measurability, coercion, contrastivity and incremental theme in order to explain the phenomena of connection of numeral classifiers and discover two conditions (Gunji and Hasida (1998b:71)):

Coercion *Coerced quantification caused by an adverbial measurement.*

Intervention *Intervention of an adverbially measurable NP in an NP-MP pair.*

When both conditions are fulfilled, the sentence is assigned as not acceptable.

On the one hand, there are several reasons to distinguish case particles and modifying particles, as has been shown. On the other hand, I doubt whether it is necessary to assume different *phrase structures* for NP+case particle and NP+modifying particle.

Yoshimoto (1997:35) argues that Japanese case particles cannot function as heads, because they can be omitted in spoken language. Ellipsis would be universally seen as a criterion to divide heads and non-heads. However, the ellipsis of heads also often occurs in other languages, as for example in German:

Example 115

Wen	hat	Peter	geküsst?
whom	did	Peter	kissed
Maria			
Maria			

(Whom did Peter Kiss? - Maria.)

The phrase-structural distinction of case particles and postpositions leads to problems, when more than one particle occurs, as in our data anlysis. The following example comes from the Verbmobil corpus:

Example 116

何時	から	が	よろしい	です	か
nanji	kara	ga	yoroshii	desu	ka
what time	from	NOM	good	COP	QUE

(At what time would you like to start?)

If one now assumes that the modifying particle *kara* is head of *nanji* as well as of the case particle *ga*, the following results for *nanji kara ga* with the head-marker structure described in Pollard and Sag (1994)[38]:

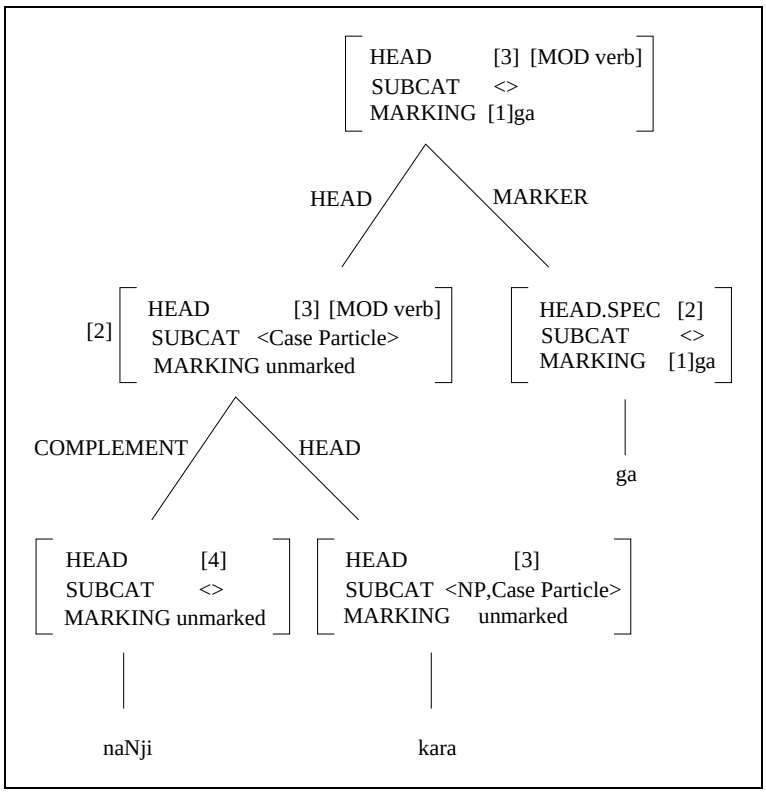


Figure 87

The case particle *ga* would have to allow nouns and modifying particles in SPEC. The latter are however usually adjuncts that modify verbal projections, as the following example shows:

Example 117: From the Verbmobil corpus

ことら	から	先生	の	ほう	の	研究室	に	お伺い
kochira	kara	sensei	no	hou	no	kenkyuushitsu	ni	o-ukagai
we	from	professor	GEN	side	GEN	institute	NI	HON-visit

する	という	形	で	よろしい	でしょう	か
suru	to iu	katachi	de	yoroshii	deshou	ka
do	COMPL	way	DE	good	COP	QUE

(Would it suit you if we come to your institute?)

Therefore the head of *kara* entails the information that it can modify a verb[39]. This information is inherited to the head of the whole phrase by the Head-Feature Principle as is to be seen in the tree above[40]. As a result, this is also admitted as an adjunct to a verb, which leads to wrong analyses for the following sentences:

---

[38] The Marking Principle says: In a headed phrase, the MARKING value is token-identical with that of the MARKER-DAUGHTER if any, and with that of the HEAD-DAUGHTER otherwise.

[39] [SYNSEM.LOCAL.CAT.HEAD.MOD *verb*]

[40] The Head Feature Principle says*: The HEAD value of any headed phrase is structure-shared with the HEAD value of the head daughter* (Pollard and Sag 1994).

Example 118

a. *何時 から が そとら が 時間 が 取れます か
nanji kara ga sochira ga jikan ga toremasu ka
what time from NOM you NOM time NOM can take QUE

b. *セミナー が 何時 から が 入って
seminaa ga nanji kara ga haitte
seminar NOM what time from NOM inserted

いらっしゃいます か
irasshaimasu ka
AUX-HON QUE

If, on the other hand, case particles as well as topic markers are heads, one receives a consistent and correct processing of this kind of example too. This is because the head information [MOD *none*] is given from the particle *ga* to the head of the phrase *nanji kara ga*. Thus, this phrase is not admitted as an adjunct:

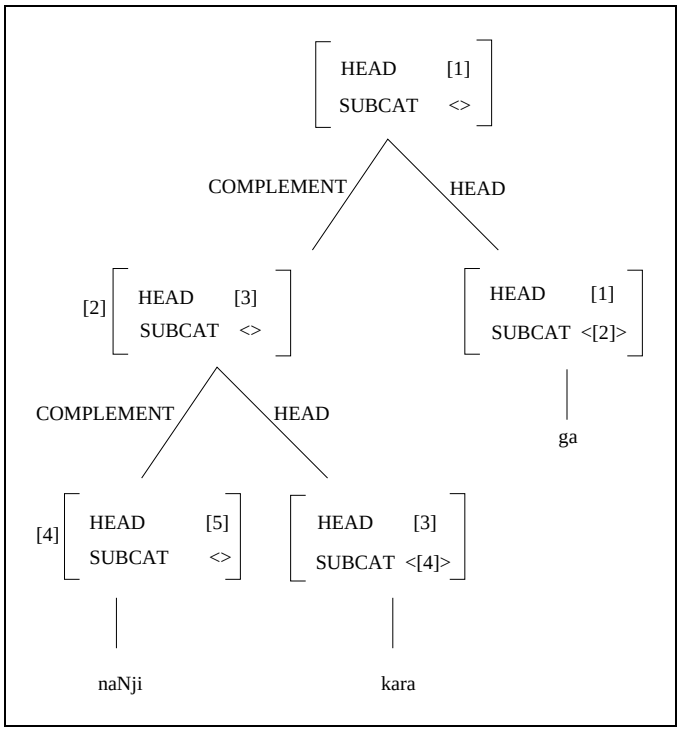


Figure 88

Similar problems occur during the analysis of the following (Verbmobil) sentences:

Example 119

明日 の 一日 から 四日 まで の ほう です
ashita no tsutachi kara yokka made no hou desu
tomorrow GEN first from fourth till GEN side COP

(It is from tomorrow the first to the fourth.)

Example 120

一時	から	三時	ぐらい	まで	を
ichiji	kara	sanji	gurai	made	wo
one o'clock	from	three o'clock	about	till	ACC

開けていただけます	か
akete itadakemasu	ka
hold free (HON)	QUE

(Could you keep the time between one o'clock and three o'clock free?)

In the case of Example 119, the phrase *made no* directly modifies the noun *hou*. *no* has the possibility to modify a noun, while *made* does not. Therefore, *no* must be the head. In Example 120, the phrase *saNji gurai made wo* is object of the sentence (marked by *wo*). Therefore, *wo* must be the head.

Pollard and Sag (1994) describe English complementizers as markers. However, a problem results, if the Japanese complementizer *to* is described as a marker. Let's have a look at the following sentence:

Example 121 (Verbmobil)

そう	なります	と	大分	先	に	なってしまう	んです	が
sou	narimasu	to	daibu	saki	ni	natte shimau	ndesu	ga
so	become	TO	a lot	earlier	NI	become	COP	SAP

(If it is like this, it will be a lot earlier.)

The complement sentence *sou narimasu* cannot be adjunct to a sentence or a VP. Therefore its head contains the entry [MOD *none*]:

Example 122

*そう	なります	大分	先	に	なってしまう	んです	が
sou	narimasu	daibu	saki	ni	natte shimau	ndesu	ga
so	become	a lot	earlier	NI	become	COP	SAP

The complement sentence with *to* – on the other hand – can modify a sentence, as Example 121 shows. It must therefore have the information [MOD *utterance*] in his head. The modification could not be realized, if *to* would be marker and *sou narimasu* would be head. Thus, we view *to* as the head of its phrase[41].

Instead of assuming different phrase structure rules, a distinction of the kinds of particles can be based on lexical types. HPSG offers the possibility to define a common type and to set up specifications for the different types of particles.

We assume Japanese to be head-final in this aspect, as the general pattern for Japanese is. All kinds of particles are analysed as heads of their phrases.

The relation between case particle and nominal phrase is a `Complement-Head' relation. The complement is obligatory and adjacent, as the following examples show:

[41] See Müller (1997) and Kiss (1995) for an argumentation against analyzing German complementizers as markers.

Example 123

a. *が
ga
NOM

b. 家 が
ie ga
house NOM

(the house)

c. 大きい 家 が
ookii ie ga
large house NOM

(the large house)

d. その 大きい 家 が
sono ookii ie ga
that large house NOM

(that large house)

e. *家 大きい が
ie ookii ga
house large NOM

f. *家 その 大きい が
ie sono ookii ga
house that large NOM

Normally, the case particle *ga* marks the subject, the case particle *wo* the direct object and the case particle *ni* the indirect object. There are however exceptions. We therefore use predicate-argument-structures instead of a direct assignment of grammatical functions by the particles (and possibly transformations). The valence information of the Japanese verbs does not only contain the syntactic category and the semantic restrictions of the subcategorized arguments, but also the case particles they must be annotated with[42].

6.3.1 The case particle *ga*

In most cases the *ga*-marked noun phrase is the subject of the sentence:

Example 124

何 日 が よろしい でしょう か
nan nichi ga yoroshii deshou ka
which day NOM good COP QUE

(Which day would suit you?)

However, this is not always the case. Notably stative verbs subcategorize for *ga*-marked objects. An example is the stative verb *dekimasu*[43]:

---

[42] Ono (1996) investigates the particles *ni*, *ga* and *wo* and also states that grammatical functions must be clearly distinguished from surface cases.

[43] See Kuno (1973) for a semantic classification of verbs that take *ga*-objects.

Example 125

彼女	が	泳ぎ	が	できます
kanojo	ga	oyogi	ga	dekimasu
she	NOM	swimming	NOM	can

(She can swim.)

These and other cases are sometimes called `double-subject constructions' in the literature. But these *ga*-marked noun phrases do not behave like subjects. They are not subject to restrictions on subject honorification or reflexive binding by the subject. This can be shown by the following example:

Example 126 (Verbmobil)

午後	の	ほう	が	ゆっくり	話	が	できます	ね
gogo	no	hou	ga	yukkuri	hanashi	ga	dekimasu	ne
afternoon	GEN	side	NOM	at ease	talking	NOM	can	SAP

(We can talk at ease in the afternoon.)

hanashi does not meet the semantic restriction [+animate] stated by the verb *dekimasu* for its subject. Nor is it constrained by subject honorification or subject binding of *jibun* in the following variants:

Example 127

私	が	ゆっくり	話	が	できております
watashi	ga	yukkuri	hanashi	ga	dekite-orimasu
I	NOM	at ease	talking	NOM	can-HON

(I can talk at ease.)

The honorification of *dekite-orimasu* does not refer to *hanashi*, but to *watashi*.

Example 128

自分	が	ゆっくり	話	が	できます
jibun	ga	yukkuri	hanashi	ga	dekimasu
self	NOM	at ease	talking	NOM	can

(? can talk at ease.)

The antecedent of *jibun* in Example 128 is outside of the sentence.

There are even *ga*-marked adjuncts, as in Example 126 and the following:

Example 129

いつ	が	ご都合	が	よろしい	でしょう	か
itsu	ga	go-tsugou	ga	yoroshii	deshou	ka
when	NOM	HON-circumstances	NOM	good	COP	QUE

(When does it suit you?)

The first NP-*ga* in these Example 126 and Example 129 is not a subject. It is not subcategorized for by the verb. It is the interrogative word in Example 129 that is marked by *ga*. *dekimasu* in Example 126 subcategorizes for two *ga*-marked NPs, but *gogo no hou ga* can neither be the subject nor the object, as it does not fulfil the semantic restrictions for these. Kuroda (1992) assumes these `double-subject constructions' to be derived from genitive

relations. This means that the meaning of the following sentence from Farmer (1984) is derived from one with a *no*-marked NP:

Example 130

a.	山	が	木	が	きれい	です
	yama	ga	ki	ga	kirei	desu
	mountain	NOM	tree	NOM	pretty	COP

(The mountains: Their trees are pretty.)

b.	山	の	木	が	きれい	です
	yama	no	ki	ga	kirei	desu
	mountain	GEN	tree	NOM	pretty	COP

(The mountain's trees are pretty.)

But this analysis seems not to be true for Example 126, because the following sentence is wrong:

Example 131

*午後	の	ほう	の	ゆっくり	話	が	できます	ね
gogo	no	hou	no	yukkuri	hanashi	ga	dekimasu	ne
afternoon	GEN	side	GEN	at ease	talking	NOM	can	SAP

ga marks a true verbal adjunct in this example.

To summarize, *ga* is a case particle that usually attaches to the sentence subject and adds case information to the entity it attaches to. Sometimes the object is marked by *ga*. This means that the grammatical function is not allocated by the case particle, but by the verbal valence. In some cases *ga* can even mark an adjunct. The case particle *ga* subcategorizes for noun phrases (as in most of the examples) and postpositions, as in Example 116. Therefore, we have two entries for the case particle *ga*: a *case-p-lex-np* and a *case-p-lex-postp* and a topic entry (see section 6.4.3.3.4: Ga-adjuncts).

6.3.2 The case particle *wo*

The case particle *wo* usually attaches to the direct object of the sentence:

Example 132

澤田	の	ほう	が	雑誌	の	インタビュー	を	受けます
Sawada	no	hou	ga	zasshi	no	intabyuu	wo	ukemasu
Sawada	GEN	side	NOM	journal	GEN	interview	ACC	give

(Sawada gives an interview to a journal.)

In contrast to *ga*, no two phrases in one clause may be marked by *wo*. This restriction is called `double-wo constraint' in research literature (see, for example, Tsujimura 1996:249ff.). Consider the following examples from the Verbmobil corpus:

Example 133

混同	に	も	再度	確認	を	して	みます	けれども
Kondou	ni	mo	saido	kakunin	wo	shite	mimasu	keredomo
Kondou	NI	too	again	confirmation	ACC	do	try	SAP

(I will confirm (it) with Mrs. Kondou again.)

Example 134

混同	の	スケジュール	を	確認	いたします
Kondou	no	sukejuuru	wo	kakunin	itashimasu
Kondou	GEN	plan	ACC	confirmation	HON-do

(I confirm Mrs Kondou's schedule.)

suru can occur with an *wo* marked argument or in a light verb construction. *kakunin* is an argument in Example 133 and the verbal noun in a light verb construction in Example 134. *kakunin* as an argument would not be possible in Example 134, according to the `double-wo constraint', because there is already an *wo*-marked argument in the sentence:

Example 135

*混同	の	スケジュール	を	確認	を	いたします	
Kondou	no	sukejuuru	wo	kakunin	wo	itashimasu	
Kondou	GEN	plan	ACC	confirmation	ACC	HON-do	

The restriction is not valid for embedded sentences:

Example 136

混同	研究室	の	ほう	で	実演	を	する	という
Kondou	kenkyuushitsu	no	hou	de	jitsueN	wo	suru	to iu
Kondou	institute	GEN	side	DE	presentation	ACC	do	COMPL

予定	を	立てている	んです	けれども
yotei	wo	tatete iru	N desu	keredomo
plan	ACC	build	COP	SAP

(There is a plan to perform the presentation at Mr Kondou's institute.)

Actually there are some violations of the restriction in the Verbmobil data corpus. Examples are:

Example 137

今日	お電話	した	の	は	(P)(h)ええと	本	を	出版
kyou	o-denwa	shita	no	wa	(P)(h)/eto/	hon	wo	shuppaN
today	telephone	did	GEN	TOP	(pause)	book	ACC	publication

する	ために	(P)	その	原稿	を	いつ	(P)	こう	一緒	に
suru	tame ni	(P)	sono	genkou	wo	itsu	(P)	kou	issho	ni
do	because	(P)	that	manuscript	ACC	when	(P)	so	joint	NI

打ち合わせ	を	したら	よろしい	か	という	こと	で	(P)	お電話
uchiawase	wo	shitara	yoroshii	ka	to iu	koto	de	(P)	o-deNwa
appointment	ACC	do-cond	good	QUE	COMPL	NOM	DE	(P)	telephone

さして	いただいた	んです	けれども	
sashite	itadaita	N desu	keredomo	
do	HON	COP	SAP	

(This is the reason, why I am calling today: When would it suit you to have a joint discussion of that manuscript?)

Example 138

うち	の	佐藤	が	(P) あの	学会誌	の	特修
uchi	no	Satou	ga	(P)/ano/	gakkaishi	no	tokushuu
we	GEN	Satou	NOM	(pause)	academic journal	GEN	special edition

の	出費津	警官	を	(P) (h)	混同	先生	と
no	shuppitsu	keikaku	wo	(P) (h)	Kondou	sensei	to
GEN	article	timetable	ACC	(P) (h)	Kondou	Prof.	with

打ち合わせ	を	したい	と	申して	おりました	けれども
uchiawase	wo	shitai	to	moushite	orimashita	keredomo
appointment	ACC	want to do	COMPL	say	AUX-Past	SAP

(Our Mr. Satou said that he would like to agree upon an appointment to discuss the timetable for the article in the special edition of the academic journal.)

But these examples were described as ungrammatical by Japanese native speakers. They are very complex. In both cases there are pauses between the *wo*-marked entities. The *wo*-marked nominal phrases *sono geNkou* and *gakkaishi no tokushuu no shuppitsu keikaku* are not subcategorized by *uchiawase*. The examples become acceptable if one replaces *wo* with *nitsuite* and thus marks the NPs as adjuncts. These exceptions of the `double-wo constraint' are therefore rare effects of spoken language and shall not be introduced into the grammar.

Object positions with *wo*-marking as well as subject positions with *ga*-marking can be saturated only once. There are neither double subjects nor double objects. It will be shown that this restriction is also valid for indirect objects. Found arguments must be assigned a saturated status in the subcategorization frame, so that they cannot be saturated again (as it is in German and English). The verbs subcategorize for at most one subject, object and indirect object. Only one of these arguments may be marked by *wo*, while a subject and an object may both be marked by *ga*. These attributes are determined by the verbal valence. The effects of the so-called 'Double-Wo-Constraint' come from the fact that *wo* has only the function of marking direct objects, while *ga* and *ni* can have different functions.

The *wo*-marked argument is not adjacent to the verb. It is possible to reverse NP-*ga* and NP-*wo* as well as to insert adjuncts between the arguments and the verb:

Example 139 (Verbmobil)

意見	交換	を	島図	の	ほう	が	さして	いただきたい
iken	koukaN	wo	Shimazu	no	hou	ga	sashite	itadakitai
opinion	exchange	ACC	Shimazu	GEN	side	NOM	do	HON-want

という	こと	で	お電話	させていただきました
to iu	koto	de	o-deNwa	sasete itadakimashita
COMPL	NOM	DE	telephone	do-HON-Past

(I've called today because Mr Shimazu would like to exchange opinions (with you))

Example 140

パネルヂスカッション	を	今度	行う	んです	けども
paneru disukasshon	wo	kondo	okonau	N desu	kedomo
panel discussion	ACC	next time	perform	COP	SAP

(Next time we will perform a panel discussion)

6.3.3 The case particle *ni*

The particle *ni* can have the function of a case particle as well as that of an adjunct particle modifying the predicate and is therefore the one that causes most problems in interpretation and processing. The task to distinguish *ni*-marked adjuncts from *ni*-marked arguments is not a trivial one. Sadakane and Koizumi (1995) also identify homophoneous *ni* that can mark adjuncts or complements. They use the notion of `affectedness' to distinguish them. This is however not useful in our domains. Ono (1996) suggests testing the possibility of passivization. This is helpful in many cases.

Some verbs subcategorize for a *ni*-marked object, as for example *naru*:

Example 141

来月	に	なる	んです	が
raigetsu	ni	naru	N desu	ga
next month	NI	become	COP	SAP

(It will become next month.)

ni-marked objects cannot occur twice in the same clause, just as *ga*-marked subjects and *wo*-marked objects. The `double-wo constraint' is neither a specific Japanese restriction nor a specific peculiarity of the Japanese direct object. It is based on the wrong assumption that grammatical functions are assigned by case particles.

There are a lot of examples with double NP-*ni*. But these are adjuncts, as in the following one:

Example 142

十時	に	研究室	の	ほう	に	お伺い	いたします
juuji	ni	keNkyuushitsu	no	hou	ni	o-ukagai	itashimasu
10 o'clock	NI	institute	GEN	side	NI	come	AUX-HON

(I'll come to your institute at 10 o'clock.)

In order not to cause massive spurious ambiguity in the interpretation of *ni*-marked entities, we follow a conservative approach to subcategorization. In the decision on adding entities with *ni* case to the subcategorization frame of a verb, we perform the tests sketched in Chapter 3, considering as arguments those *ni*-marked entities that are obligatory, can be passivized and/or get a semantic restriction from the head verb. Others are adjuncts.

6.3.4 Other case particles

The case change rules that can apply to verbs with *kata* (and in other cases) replace the case of the subject or the object argument with *no*-case. Therefore, we have an entry for a case particle *no* that exactly assigns this case.

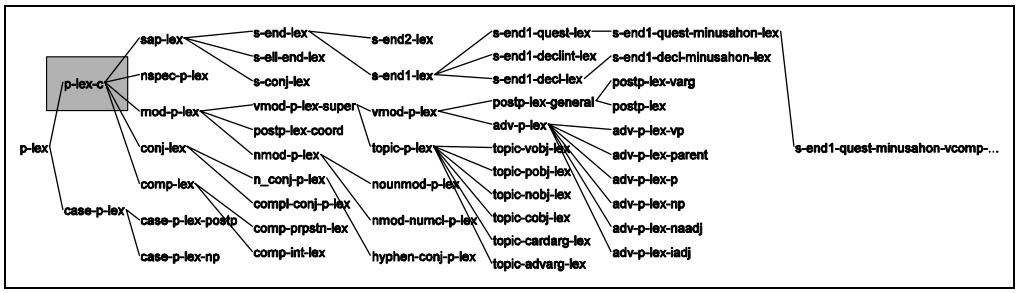
Subcategorization for a case particle *to* is also possible:

Example 143

花子	と	喧嘩	した。
Hanako	to	kenka	shita
Hanako	TO-CASE	fight	light verb

((I) fought with Hanako.)

6.4 Particles with semantic content



Semantic particles add semantics to the parse, comparable to prepositions in English. An essential problem is to find criteria for the classification and distinction of case particles and modifying particles. On the semantic level they can be distinguished in particles that introduce their own semantics and those that have a functional meaning. This is the main distinction between *case-p-lex* and *p-lex-c* in our type hierarchy. According to this distinctive feature the particle *no* is a non-case particle, because it introduces attributive meaning, as opposed to Tsujimura (1996:134), who classifies it as a case particle. Another distinctive criterion that is introduced by Tsujimura (1996:135) says that "postpositions" (as she calls them) are obligatory in spoken language, while case particles can be omitted. Case particles are indeed suppressed much more often, but there are also cases of suppressed modifying particles. These occur mainly in temporal expressions in our dialogue data:

Example 144

それでは	十四日	の	午後	2時	を	ロッビー	の
soredewa	juuyokka	no	gogo	niji	wo	robii	no
then	14th	GEN	afternoon	2 o'clock	ACC	lobby	GEN

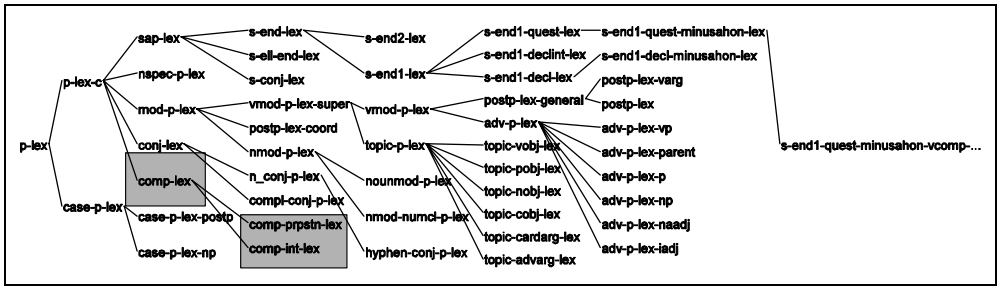
ほう	で	お待ち	して	おります
hou	de	o-machi	shite	orimasu
side	DE	HON-wait	do	AUX-HON

(I will then wait in the lobby at 2 o'clock on the 14th.)

Finally Tsujimura gives the criterion that case particles can follow modifying particles while "postpositions" cannot follow particles. This criterion in particular implies that a finer distinction is necessary, as we have shown that it is not that easy. This finer distinction can be realized with HPSG types. Further, topic particles are not taken into account in Tsujimura's classification.

We first distinguish on the top level of the type hierarchy of particles case particles from others, using the criterion of semantic contribution. Case particles have a syntactic (i.e. functional) function and do not add to the MRS, while p-lex-c particles do.

6.4.1 Complementizers



Complementizers take complement sentences and assign case, such that these can be subcategorized for by verbs. This raises the question, why they are not sorted under case particles in the type hierarchy. The reason is that case particles have no semantic contribution, while complementizers do: They add a message relation to RELS in the MRS. The type hierarchy contains complementizers that take a sentence and add a proposition to the MRS (comp-prpstn-lex) and those that take a sentence and add a question to the MRS (comp-int-lex). An example for the proposition adding complementizer can be seen in Example 145, Example 146 shows the question adding complementizer.

Example 145

そちら	に	伺いたい	と	思います
sochira	ni	ukagaitai	to	omoimasu
you	NI	visit	COMPL	think

(I would like to visit you.)

Example 146

花子	が	ご飯	を	食べた	か	分からない
Hanako	ga	gohan	wo	tabeta	ka	wakaranai
Hanako	NOM	rice	ACC	ate	COMPL	not know

((I) don't know if Hanako ate the rice.)

The *to* complementizer attaches to complement sentences that are subcategorized for by verbs like *omou*, *iu* or *kaku*. These complement sentences are adjacent (see Example 145 above). The complement sentence is an utterance that can have sentence end particles:

Example 147

そろそろ	打ち合わせ	を	しよう	か	と	思う	のです	が
sorosoro	uchiawase	wo	shiyou	ka	to	omou	no desu	ga
soon	meeting	ACC	let's do	QUE	COMPL	think	COP	SAP

(I think we should soon arrange a meeting.)

The MRS for Example 145 contains the extra proposition that is added by the complementizer (see Figure 89).

```
h1: proposition_m_rel(h3)
h4: _sochira_n_1_rel(x5)
h6: def_rel(x5,h7)
h4: place_rel(x5)
h9: _ni_p_rel(e11,x5,e10)
h12: _ukagau_visit_rel(e14,u13)
h15: tai_rel(e10,u13,h16)
h16: proposition_m_rel(h19)
h20: _omou_v_rel(e2,u21,h18)
h3 qeq h20
h7 qeq h4
h19 qeq h12
h18 qeq h15
```

Figure 89

The complementizer gets a case entry, because its head is a subtype of **case-particle-head**. It can therefore be subcategorized for by verbs such as *omou*. Lexical entries are of the subtypes **comp-prpstn-lex** (*to, kamo*) or **comp-int-lex** (*ka, kadouka, noka*), introducing a proposition or an interrogative to the MRS. They subcategorize for verbal or sentence particle heads (*to, ka, kadouka, noka*), or for heads of the type **sentence-valid**, which contains verbal heads, quotations or idioms (*kamo*). The complement is obligatory.

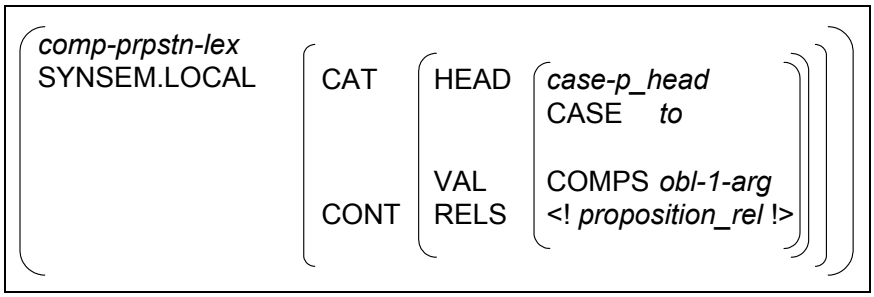


Figure 90: The complementizer *to*

6.4.2 Genitive specifying *no*

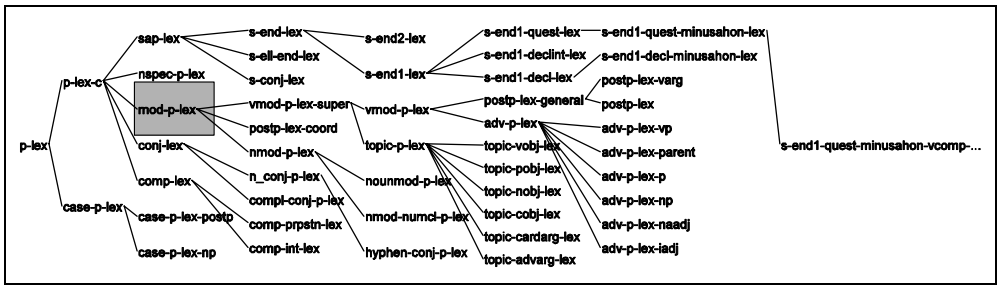
As described in Section 5.4, some nominalizations subcategorize for a complement sentence, a determiner or a genitive. In order to provide the genitive structure (see Example 148), we need an entry of *no* that contains SPEC information in its head, can be subcategorized for by the nominalization and contributes some semantic relation. This is accounted for by the particle type **nspec-p-lex**. The actual relation is quite underspecified and the interpretation must be left to discourse interpretation. It is an **unspec_compound_rel**, which has as its arguments the indices of the subcategorized noun and the nominalization.

Example 148

そちら　の　方
sochira　no　hou
you　GEN　side

(your side)

6.4.3 Modifying particles



Modifying particles (**mod-p-lex**) get the information in MOD that they can become adjuncts to verbs (verb modifying particles) or nouns (the noun modifying particle *no*) and their specific semantic information. They subcategorize for a noun, as all particles do.

The modifying particles share in their lexical entries the information about a case that is never subcategorized for by verbs (mod), some content in the MOD feature, an obligatory complement and some content in the MRS RELS (see Figure 91).

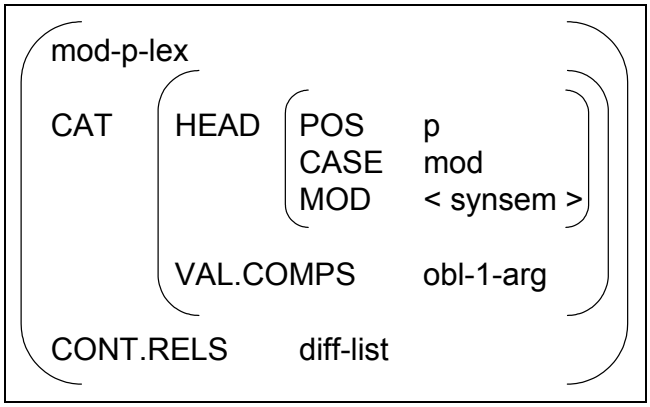
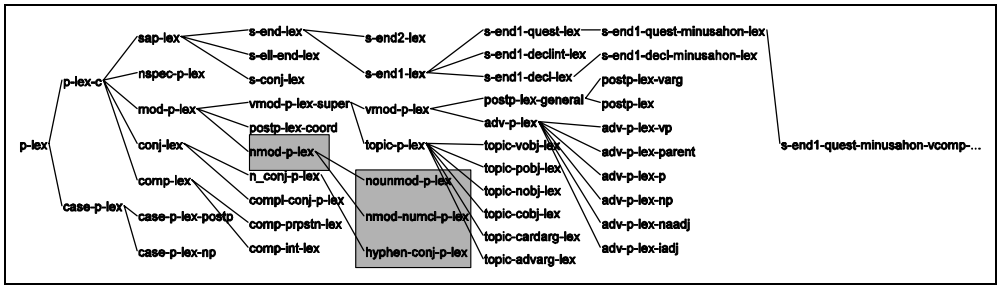


Figure 91: Modifying particles

6.4.3.1 The noun modifying particle *no*



no is a particle that modifies nominal phrases. This is an attributive modification and has a wide range of meanings, as the following examples indicate:[44]

Example 149

a. ほか　の　日
hoka　no　hi
another　GEN　day

(another day)

[44] See also Tsuda and Harada (1996).

b. 次　　の　日
tsugi　no　hi
next　GEN　day

(next day)

c. 私　　の　研究室
watakushi　no　kenkyuushitsu
I　　GEN　institute

(my institute)

d. 二十九日　　の　午前中
nijuukunichi　no　gozenchuu
29th.　　GEN　afternoon

(the afternoon of the 29th.)

e. 京都　大学　　の　川村
Kyouto　daigaku　no　Kawamura
Kyoto　University　GEN　Kawamura

(Kawamura of Kyoto University)

Tsujimura (1996:134ff.) assigns *no* to the class of case particles. The following criteria support this classification:

- Tsujimura's postpositions are obligatory in spoken language, case particles are optional.
- Case particles can - as Tsujimura states - follow postpositions, but postpositions cannot follow case particles.

However, the criterion on semantic contribution supports the idea to classify *no* as something else than a case particle:

- Tsujimura's postpositions have their own semantic meaning. Case particles have a functional meaning. *no* however has a semantic, namely attributive meaning.

As described above, our first distinction is based on semantic contribution. Therefore *no* is classified into the type of contributing particles, **p-lex-c**. Further, the particle contains information in MOD, such that it can be sorted into modifying particles **mod-p-lex**. We further distinguish **noun-mod-p-lex** for noun modifying particles and **nmod-numcl-p-lex** for *no* used in numeral classifier constructions. The latter ones are explained in Section 5.6.

The particle *no* subcategorizes for a noun, as the other particles do. It also modifies a noun. This separates it from the other modifying particles in **vmod-p-lex**, which modify verbal heads. NP-*no* is an adjunct to a nominal phrase. As a result, the analysis of multiple NP-*no* is possible:

Example 150

先生　の　ほう　の　大学　　の　研究室　　に　伺えば
sensei　no　hou　no　daigaku　no　kenkyuushitsu　ni　ukagaeba
Prof.　GEN　side　GEN　University　GEN　institute　NI　go (COND.)

いい　んです　ね
ii　　ndesu　ne
good　COP　　SAP

(It would be good to come to your institute, wouldn't it?)

Besides the function of a particle, the word *no* can also have the function of a nominalizer[45]. In this case, it subcategorizes for a verbal head and builds an NP (and can thus be followed by any particle).

The particle *no* modifies a noun phrase and occurs after a noun (as in Example 150) or a verb modifying particle, as in:

Example 151

四日　から　の　週　です　ね
yokka　kara　no　shuu　desu　ne
4th　from　GEN　week　COP　SAP

(It's the week beginning from the fourth, isn't it.)

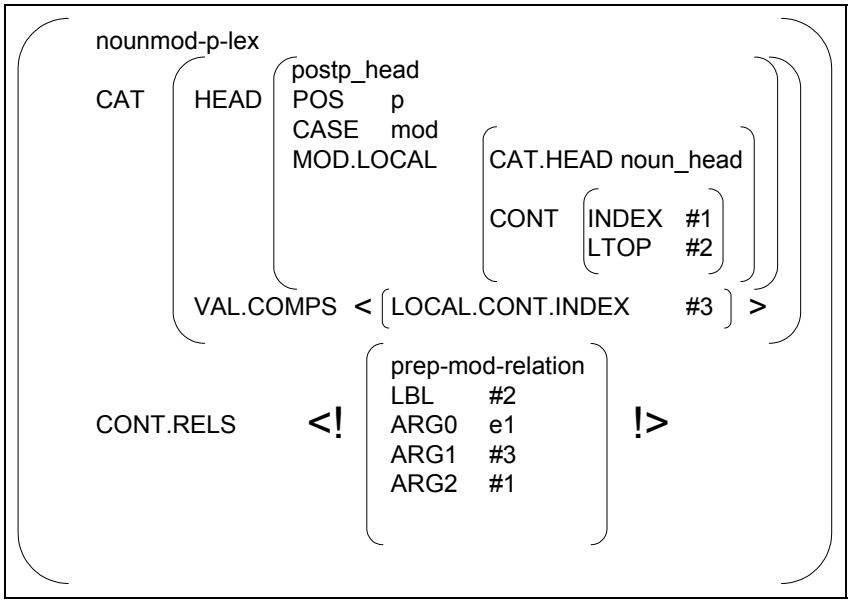
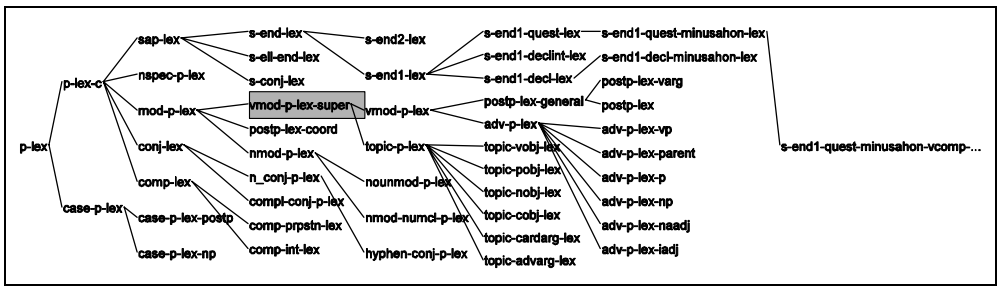


Figure 92

The semantic contribution of *no* gets the underspecified value of **_no_p_rel**, as it can be highly ambiguous and resolvable only in the linguistic context, as shown above. The arguments of the prepositional relation are the subcategorized and the modified nominal entities.

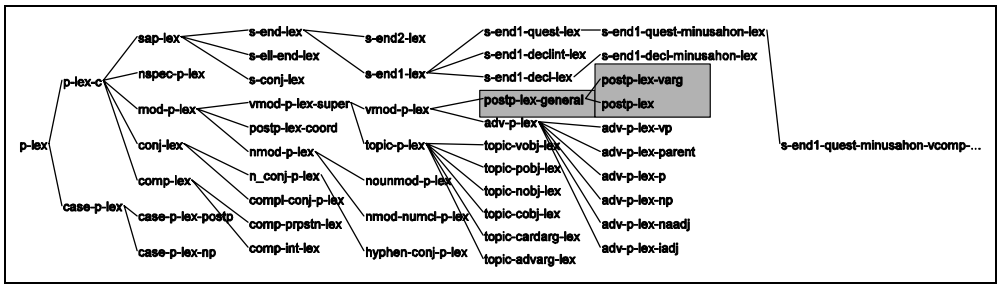
[45] See Nightingale (1996) and chapter 5.4 for a detailed description of this function of *no*.

6.4.3.2 Verb modifying particles



The verb modifying particles specify the modification of the verb in SYNSEM.LOCAL.CAT.HEAD.MOD. They link their semantic ARG2 with its INDEX and their local LTOP with its LTOP. Topic particles and other verb modifying particles are distinguished. This is due to their different behaviour in co-occurrence (topic particles can follow other particles) and the information they contribute. While modifying particles add to the semantics, topic particles add to the semantics and the context of the sentence.

6.4.3.2.1 Postpositions



The postpositions modify a verb as an adjunct and subcategorize for a nominal object (**postp-lex**), a verb (**postp-lex-varg**) or a conjunction (**postp-lex-coord**).

There are quite a lot of postpositions in Japanese and we will explain some of them in further detail here.

e is a non-ambiguous particle. It is verb modifying and has a directional function co-occuring with verbs of movement. *e* shares this function with *ni*:

Example 152

九時	に/*へ	そちら	に/へ	伺います
kuji	ni/*e	sochira	ni/e	ukagaimasu
9 o'clock	NI/*E	you	NI/E	go

(I'll come to you at 9 o'clock.)

The postpositions *kara* and *made* attach to verb modifying adjuncts. These are - as far as the Verbmobil data is concerned - mainly temporal and locative expressions:

Example 153

先生	の	おうち	から	遠い	ので	お昼	から
sensei	no	o-uchi	kara	tooi	node	o-hiru	kara
Prof.	GEN	hon_home	from	far	because	HON-noon	from

に　しましょう　か
ni　shimashou　ka
NI　shall do　QUE

(Shall we start from noon, because it's far from your home?)

Time periods are realized with *kara ... made*:

Example 154

会議　が　朝　の　十一時　から　昼　の　一時
kaigi　ga　asa　no　juuichiji　kara　hiru　no　ichiji
meeting　NOM　a.m.　GEN　11 o'clock　from　p.m.　GEN　1 o'clock

まで　あります　けれども
made　arimasu　keredomo
till　exist　SAP

(There is a meeting from 11 a.m. to 1 p.m.)

kara as well as *made* can be complements of *desu*:

Example 155

三時　から/まで　です　か
sanji　kara/made　desu　ka
3 o'clock　from/to　COP　QUE

(Does it start/end at three?)

kara and *made* are non-ambiguous modifying particles, as *e* is. I will call them `postpositions'. They subcategorize for nominal phrases and may not follow any other particles. Another particle in this category is *nanka*. This word has one function as an adverb and one as a postposition. The postposition *nanka* marks a verb modifying adjunct. An example is:

Example 156

午後　なんか　お時間　よろしい　でしょう　か
gogo　nanka　o-jikan　yoroshii　deshou　ka
afternoon　NANKA　HON-time　good　COP　QUE

(Would the time in the afternoon be good for you?)

Further postpositions are *to-shite* and *to-shimashite*, as in:

Example 157

こちら　として　は　都合　が　いい　んです　けれども
kochira　to-shite　wa　tsugou　ga　ii　ndesu　keredomo
we　To-SHITE　TOP　circumstances　NOM　good　COP　SAP

(This is good for us.)

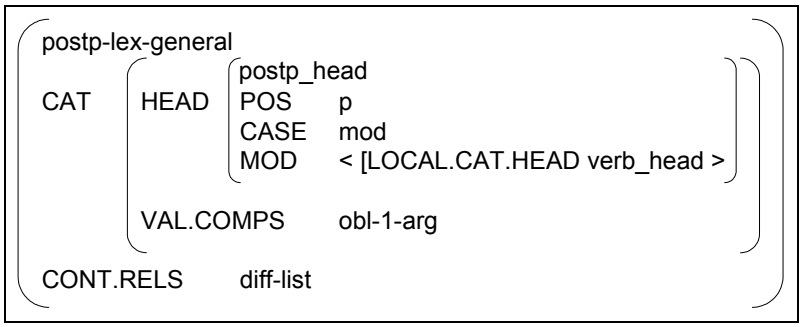
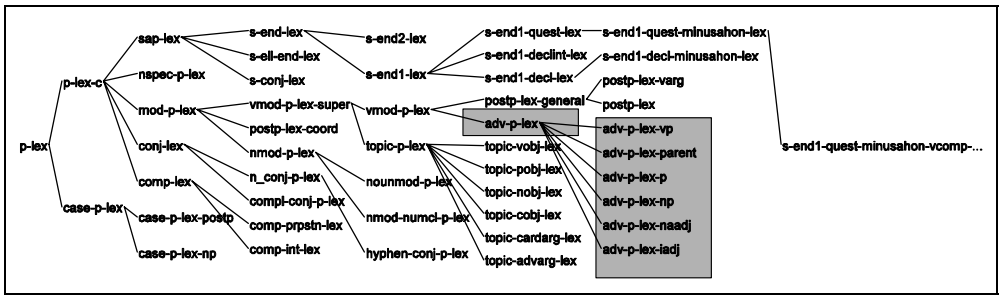


Figure 93: Postpositional particles

6.4.3.2.2 Adverbial particles



Nightingale (1996) treats *ni* and *de* as the infinitive and the gerund form of the copula. To account for this, it has to be clarified what the qualities of an infinitive and a gerundive form, a copula and a verb modifying particle are in our type system. Let us first consider the infinitive form. In our syntax, it has the following peculiarities:

- not honorific concerning addressee
- present tense
- indicative
- possible to use with *n desu* (*jikan ga toreru n desu ka*)
- possible as a relative sentence (V-*ru koto*/N)
- possible as a complement sentence (V-*ru to omou/iu*)
- can be modified by an adverb

ni is similar to the infinitive form in respect to the fact that it can take an adverb as its argument (*gogo wa furii ni natte imasu* -- afternoon - TOPIC - free - become). But the infinitive is clearly distinct from the characteristics of *ni*, that cannot be used with *N desu*, cannot mark a relative sentence *(*John ga furii ni koto*) and cannot be marked with the complementizer *to* (**John ga furii ni to omou*).

We define the gerundive form, a copula and a verb modifying particle as follows:

- A gerundive form is not finite, it can modify a verbal phrase and be specifier of an auxiliary.
- A copula is a nonauxiliary verb. It subcategorizes for an oblique object, which is an unmarked noun, a postpositional phrase or an adjective. It further subcategorizes for an optional subject, which is marked with *ga*.
- A verb modifying particle is a particle that modifies a nonauxiliar verbal phrase and subcategorizes for an oblique object. Adverbial particles in our type hierarchy subcategorize for a noun or a postposition.

The adjunctive form *de* has both qualities of a gerundive copula and of a particle:

- Subcategorizing for an unmarked noun or a postposition
- Being adjunctive to a verbal head
- Its semantic behaviour (see Nightingale 1996)

There are arguments for treating it as a copula:

- Historical derivation (see Nightingale 1996).
- *de arimasu* behaves like *desu.*
- The form *deshite* exists.

But there is some data that shows different behaviour of *de* and other gerundives. Firstly, it concerns the co-occurrence possibilities of *de* and other particles, compared to gerundive forms and particles:

- *de wa - V-te wa*
- *de mo - V-te mo*
- *de no - V-te no*
- *de ga - *V-te ga*
- *de wo - *V-te wo*
- *de ni - *V-te ni*
- *de de - *V-te de*

Secondly, a gerund may be subcategorized for by auxiliaries, e.g. *shite kudasai, shite orimasu*, but *de* may not. Additionally there is something which distinguishes *de* of a copula: It may not subcategorize for a subject.

A word that is an adjunct to verbs, subcategorizes for an unmarked noun or a phrase with a postposition and is subcategorized for by several particles (see above) fits well into our description of a verb modifying particle.

The adverbial particles *ni*, *de* and *to* subcategorize for a noun or a postposition, as can be seen in Example 158 and Example 159. The possibility to subcategorize for (i.e., occur after) a postposition is the main criterion to make the distinction between postpositions and adverbial particles in the type hierarchy.

Example 158

二十四日	から	に	迫って	います
nijuuyokka	kara	ni	sematte	imasu
24th	from	NI	be close	AUX

(The 24th is already close.)

Example 159

一時	から	で	お昼ご飯	の	ほう	は
ichiji	kara	de	ohirugohan	no	hou	wa
1 o’clock	from	DE	lunch	GEN	side	TOP

だいじょうぶ　です　ね
daijoubu　　desu　ne
good　　COP　SAP

(Would the lunch be fine from one o'clock?)

ni as a modifying particle can be found very often in temporal or locative expressions in the Verbmobil data.

Example 160

三時　　に　会議　　が　終わります
sanji　　ni　kaigi　　ga　owarimasu
3 o'clock　NI　meeting　NOM　end

(The meeting ends at three o'clock.)

Example 161

私　　が　そちら　の　研究室　　に　伺わない
watakushi　ga　sochira　no　kenkyuushitsu　ni　ukagawanai
I　　NOM　you　　GEN　institute　　NI　not visit

と　　いけない　　と　　思います　が
to　　ikenai　　to　　omoimasu　ga
COMPL　must not do　COMPL　think　　SAP

(I think I'll have to come to your institute.)

ni can subcategorize for predicates (*adv-p-lex-vp*), the verb being in infinitive form. Therefore, the verb gets an adverbial meaning:

Example 162

花　　を　見　　に　行く
hana　　wo　mi　　ni　iku
flowers　ACC　watch　NI　go

(go to watch flowers)

de can be a verb modifying particle. It has a temporal, locative or instrumental meaning. The temporal meaning of *de* is restricted to stative verbs:

Example 163

朝　　十時　　ぐらい　から　十二時　　まで　の　間　　で
asa　　juuji　　gurai　kara　juuniji　　made　no　aida　　de
morning　10 o'clock　ca.　　from　12 o'clock　till　GEN　interval　DE

やりたい　　と　　思う　んです　けれども　いかが　でしょう　か
yaritai　　to　　omou　n desu　keredomo　ikaga　deshou　ka
want to do　COMPL　think　COP　　SAP　　good　COP　　QUE

(I would like to do it between 10 and 12 o'clock in the morning. Would that suit you?)

The locative usage of *de* is non-directional:

Example 164

研究室	で	実験	の	実演	を	したい
kenkyuushitsu	de	jikken	no	jitsuen	wo	shitai
institute	DE	experiment	GEN	performance	ACC	want to do

んです	けれども
ndesu	keredomo
COP	SAP

(I would like to perform the experiment in the institute.)

An example for the instrumental usage is:

Example 165

バス	で	来ます
basu	de	kimasu
bus	DE	come

(I'll come by bus.)

The particle *to* can mark an adjunct to a predicate, which qualifies *to* as an adverbial particle[46]:

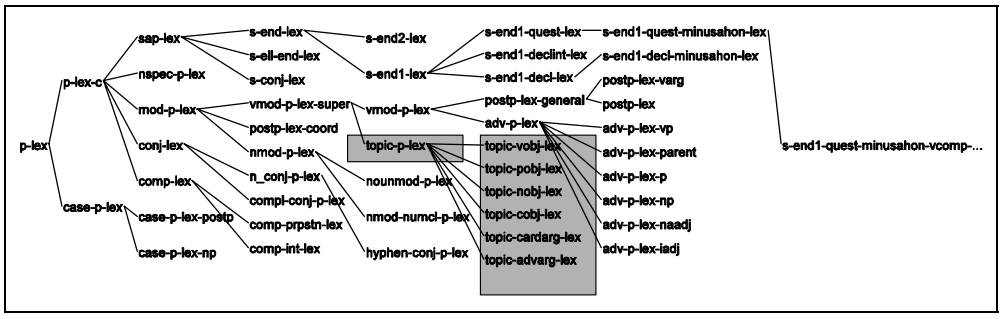
Example 166

清水	先生	と	展示会	を	ご一緒	させて	いただく
Shimizu	sensei	to	tenjikai	wo	go-issho	sasete	itadaku
Shimizu	Prof.	COMPL	exhibition	ACC	together	do	HON

(I would like to organize an exhibition with Prof. Shimizu.)

Further sub-types of *adv-p-lex* are due to the arguments the particles take: noun phrases, adjectives, verb phrases and parentheses.

6.4.3.3 Particles of topicalization



6.4.3.3.1 Wa

The topic particle *wa* can mark arguments as well as adjuncts. In the case of argument marking it replaces the case particle (see Example 167, where it replaces *ga*). In the case of adjunct marking it can replace the verb modifying particle (see Example 168, where it replaces *ni*) or it can occur after it (see Example 169):

[46] It is not categorized as a postposition, because it cannot be followed by case particles.

Example 167

午後	は	開いて	おります	ので
gogo	wa	aite	orimasu	node
afternoon	TOP	be free	HON-AUX	SAP

(The afternoon is free.)

Example 168

二十八日	の	月曜日	は	会議	が	午後	に
nijuuhachinichi	no	getsuyoubi	wa	kaigi	ga	gogo	ni
28th	GEN	Monday	TOP	meeting	NOM	afternoon	NI

入って	おります
haitte	orimasu
be inserted	HON-AUX

(On Monday the 28th there is a meeting in the afternoon.)

Example 169

今月中	に	は	ぜひ	お会いしたい	と
kongetsuchuu	ni	wa	zehi	o-ai shitai	to
this month	NI	TOP	certainly	want to meet	COMPL

思う	んです	が
omou	n desu	ga
think	COP	SAP

(I would certainly like to meet you within the month.)

Semantic interpretation has to divide anaphoric, generic and contrastive readings of *wa* (see Kuno 1973). On the syntactic level, it has to be decided, whether the topic particle marks an argument or an adjunct, when it occurs without a verb modifying particle. This is difficult because of the optionality of verbal arguments in Japanese. If it marks an argument, it has to be decided which grammatical function this argument has. This problem can often not be solved on the purely syntactic level. Lexical semantic restrictions for verbal arguments are necessary:

Example 170

場所	の	ほう	は	どう	しましょう	か
basho	no	hou	wa	dou	shimashou	ka
place	GEN	side	TOP	how	shall do	QUE

(How shall we resolve the problem of the place?)

Subject and object of the verb *shimashou* are suppressed in this example. The sentence can be interpreted as having a topic adjunct, but no surface subject and object, when using semantic restrictions for the subject (agentive) and the object (situation).

Gunji (1991) analyses Japanese topicalization with a trace that introduces a value in SLASH and the `Binding Feature Principle' that unifies the value of SLASH with a *wa*-marked

element[47]. This treatment is similar to the one introduced by Pollard and Sag (1994) for the treatment of English topicalization. There as well a trace introduces a SLASH value which is bound by the topicalized element. However, Japanese topicalization is fundamentally different from English one. Firstly, it occurs more frequently. Up to 50% of the sentences are concerned (Yoshimoto 1997). Secondly, there are examples where the topic occurs in the middle of the sentence, unlike the English topics that occur sentence-initially. Yoshimoto (1997) gives the example:

Example 171

ビル	が	東京	へ	は	行く
Bill	ga	Toukyou	e	wa	iku
Bill	NOM	Tokyo	E	TOP	go

(Bill goes to Tokyo.)

There are also examples in the Verbmobil dialogue corpus:

Example 172

来週中	に	打ち合わせ	は	したい	んです	けれども
raishuuchuu	ni	uchiawase	wa	shitai	ndesu	keredomo
next week	NI	meeting	TOP	want to do	COP	SAP

(I would like to hold a meeting in the next week.)

Thirdly, Japanese verbal arguments are optional. Suppressing of verbal arguments could be called more a rule than an exception in spoken language. The SLASH approach would introduce traces in almost every sentence. This, in connection with scrambling and suppressed particles, could not be restricted in a reasonable way. If one follows Gunji's interpretation of those cases, where the topic-NP can be interpreted as a noun modifying phrase, a genitive gap has to be assumed. But this leads to assuming a genitive gap for every NP that is not modified. Further, genitive modification can be iterated.

Fourth, two or three occurences of NP-*wa* are possible in one utterance:

Example 173

ご予定	の	ほう	は	来週	は	先生	は
go-yotei	no	hou	wa	raishuu	wa	sensei	wa
HON-plan	GEN	side	TOP	next week	TOP	Prof.	TOP

いかが	でしょう	か
ikaga	deshou	ka
good	COP	QUE

(Concerning your plans: Would next week suit you?)

Thus, we decided to assign topicalized sentences the same syntactic structure as non-topicalized sentences and to resolve the problem on the lexical level. Still, there is a problem of massive ambiguity (where topics can be linked to arguments or not) that asks for a decision:

---

[47] The Binding Feature Principle says:

The value of a binding feature of the mother is identical to the union of the values of the binding feature of the daughters minus the category bound in the branching.

We have the possibility to introduce ambiguous readings in many cases and leave the disambiguation to a disambiguation module, or analyse all topics as modifiers to the sentence and leave the linking to a zero pronoun resolution module. In both cases, there is the necessity to rely on a natural language processing module that has access to a different type of information than the HPSG grammar processing. The introduction of ambiguity is useful when parsing not too long sentences and building up treebanks with interfering human evaluation, such as being done in the Hinoki project (Bond et al. 2004). The underspecification of information is useful when parsing large amounts of data containing long sentences and much topicalization, such as was done in the Verbmobil project (Wahlster 2000). As there are different demands for different kinds of processing, we decided to insert the possibility for ambiguous readings and set a switch to the root node of the grammar that constraints the application of the lexical entry which replaces the case particle if required.

The topic particle gets three lexical entries. The first one is for the verb modifying topic variant, as in Example 168, Example 169 and Example 170. The second entry is for the case marking variant of *wa,* as in Example 167, where the case is assigned *ga*. It gets the same head as the other case marking particles and does not add to the semantics, just like case particles. In the case of a topic particle *wa* replacing *wo*, there is furthermore empathy set to the entity marked by the particle, as Watanabe (2000) states. This topic case particle as well does not add to the semantics, but to CONTEXT: it adds empathy setting to the entity it attaches to.

6.4.3.3.2 Other topic particles

mo is similar to *wa* in some aspects. It can mark a predicative adjunct and can follow *de* and *ni*. But it can also follow *wa*, an adjective and a sentence with question mark:

Example 174

午後	で	も	けっこう	です
gogo	de	mo	kekkou	desu
afternoon	DE	TOP	good	COP

(It would also be good in the afternoon)

Example 175

忙しい	も	ので。。。
isogashii	mo	node…
busy	TOP	because

(Because I am busy...)

Example 176

できる	か	も	知れません
dekiru	ka	mo	shiremasen
can	QUE	TOP	do not know

(I don't know if I can)

The particle *dake* can replace *ga* or *wo*, as in Example 177 and adds to the semantics. We have thus an entry for *dake* that is a topic particle.

Example 177

学生	の	半分	だけ	参加	しました。
gakusei	no	hanbun	dake	sanka	shimashita
students	GEN	half	DAKE-TOPIC	participate	light verb

(Only half of the students have participated)

Other topic particles in the lexicon that attach to nouns, replace case particles or are adjuncts to the sentence and add to the semantic content are *demo, koso, nado, nanowa, nomi, notame, sura, tteiunowa, tte, ja* and *shika*.

6.4.3.3.3 Topic particle types

The topic particles can attach to (i.e. subcategorize for) nominal, verbal, particle, conjunctional, cardinal and adverbial heads. The subtypes of topic-p-lex reflect this variation (see Table 14).

Table 14: Types and instances of particles

particle type	type name in the hierarchy	particles in this type
topic particle that takes a nominal	*plain-topic-nobj-lex*	*wa, ga, demo, koso, mo, nanowa, nomi, notame, sura, tteiunowa, tte, ja, shika, nado*
topic particle that takes a modifying particle	*topic-pobj-lex*	*wa, ga, koso, mo, shika*
topic particle that takes a complementizer	*topic-cobj-lex*	*wa*
topic particle that takes a verb	*topic-vobj-lex*	*wa, mo, shika*
topic particle that takes an adverb	*topic-advarg-lex*	*wa, mo, dewa, made*
topic particle that takes a cardinal	*topic-cardarg-lex*	*shika*

Most occurrences are topic particles attaching to nominal heads, such as in Example 172. They insert a relation to the MRS and link this with the subcategorized entity and the modified event. See in Figure 94 the MRS for a noun *hon* (book) and the topic particle *wa*.

```
h4: _hon_n_rel(x5)
h6: udef_rel(x5, h7)
h9: _wa_p_rel(e10, x5, e2)
h7 qeq h4
```

Figure 94: MRS for *hon wa*

As already detected in Section 6.1, topic particles can follow modifying particles and complementizers. This is accounted for in the types topic-pobj-lex and topic-cobj-lex. In this case as well, a topic relation is added, which links the nominal entity with the verbal event. See in Figure 95 the MRS for *hon kara wa* (book – from – topic).

```
h4: _hon_n_rel(x5)
h6: udef_rel(x5, h7)
h9: _kara_p_rel(e10, x5, e2)
h11: _wa_p_rel(e12, x5, e2)
h7 qeq h4
```

Figure 95: MRS for *hon kara wa*

Topic particles can attach to verbs in *te* form, as in Example 178.

Example 178

寝て　は　ならない
nete　wa　naranai
sleep　TOPIC　not become

(It is not allowed to sleep)

The topic particle links the subcategorized event and the modified event with its arguments in the MRS. The same happens when attaching topic particles to adverbs like *hayaku* (fast).

The topic particle *shika* attaches to cardinals, such that we have a type *topic-cardarg-lex* as well.

Example 179

百　しか　ない
hyaku　shika　nai
100　TOP　NEG

(It is only 100.)

6.4.3.3.4 Ga-adjuncts

One can find several examples with *ga* marked adjuncts in the Verbmobil data. On the level of information structure it is said that *ga* marks neutral descriptions or exhaustive descriptions (c.f. Gunji 1987, Kuno 1973). Gunji analyzes these exhaustive descriptions syntactically in the same way as he analyzes his `type-I topicalization'. They build adjuncts that control gaps or reflexives in the sentence. He designs *ga* marked adjuncts without control relations as relying on a very specialized context. Gunji's lexical entries for exhaustive *ga* are:

a. {POS P; PFORM ga; SUBCAT {PP [PFORM *pf*; SEM α]}; ADJUNCT V [SLASH {PP [PFORM *pf*; SEM α]}]}; where *pf* is not *ga, wo*, *ni* or *no*.

b. {POS P; PFORM ga; SUBCAT {NP [SEM α]}; ADJUNCT V [SLASH {PP [PFORM *pf*; SEM α]}]}; where *pf* is *ga*, *wo*, *ni* or *no*.

c. {POS P; PFORM ga; SUBCAT {NP [SEM α]}; ADJUNCT V [REFL {PP [SBJ; SEM α]}]}

However, this treatment leads to the following problems:

1. In all cases, where *ga* marks a constituent that is subcategorized as *ga*-marked by the verb, a second reading is analyzed that contains a *ga* marked adjunct controlling a gap. This is not reasonable. The treatment of the different meaning of *ga* marking arguments and *ga* marking adjuncts belongs to the semantics and not into the phrase structure.

2. This treatment assumes gaps. We already criticized this in connection with topicalization.
3. The Verbmobil dialogue data contains mostly examples with *ga* marked adjuncts without syntactic control relation to the rest of the sentences.

On the level of syntax, we do not decide whether a *ga*-marked subject or object is a neutral description or an exhaustive listing. This decision must be based on context information, where it can be found out whether the noun phrase is generic, anaphoric or new. We distinguish occurrences of NP+*ga* that are verbal arguments from those that are adjuncts.

The examples for *ga*-marked adjuncts in the Verbmobil dialogues can be classified into two kinds:

a. The NP describes a temporal entity:

Example 180

私	の	ほう	の	都合	は	二十八日	が
watakushi	no	hou	no	tsugou	wa	nijuuhachinichi	ga
I	GEN	side	GEN	circumstances	TOP	28th	NOM

午後	に	会議	が	一見	入って	おります
gogo	ni	kaigi	ga	ikken	haitte	orimasu
afternoon	NI	meeting	NOM	at first	inserted	HON-AUX

(On our side, there is at first a meeting inserted at the afternoon of the 28th.)

Example 181

こちら	は	月曜日	が	ちょっと	スケジュール	が
kochira	wa	getsuyoobi	ga	chotto	sukejuuru	ga
we	TOP	Monday	NOM	somewhat	schedule	NOM

いっぱい	なんです	けれども
ippai	nandesu	keredomo
full	COP	SAP

(On our side, the schedule is full on Monday.)

b. The NP describes a personal entity:

Example 182

私	が	十二時	に	会議	が	終わります
watakushi	ga	juuniji	ni	kaigi	ga	owarimasu
I	NOM	12 o'clock	NI	meeting	NOM	end

(As far as I am concerned, the meeting ends at 12 o'clock.)

The adjunctive *ga* is restricted to those cases where the subject of the sentence is saturated and not a zero pronoun. It therefore restricts the XARG of the modified event to be of type *full_ref_ind*, as can be seen in Figure 96.

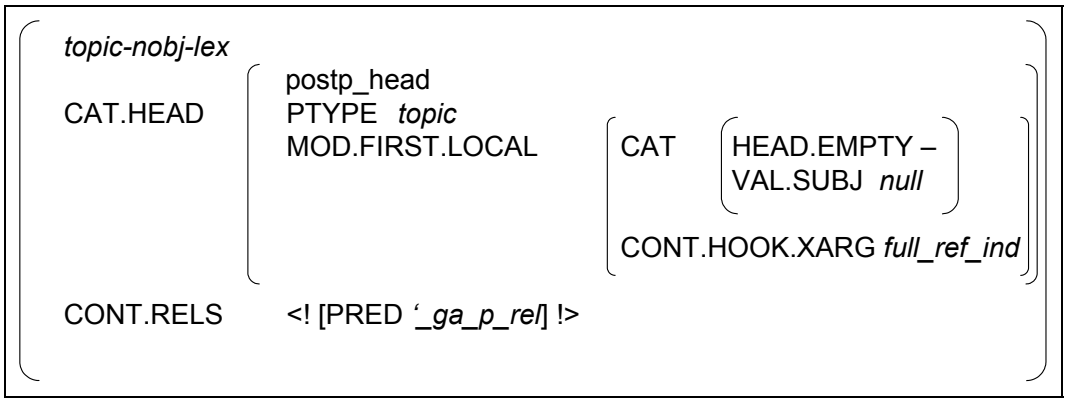
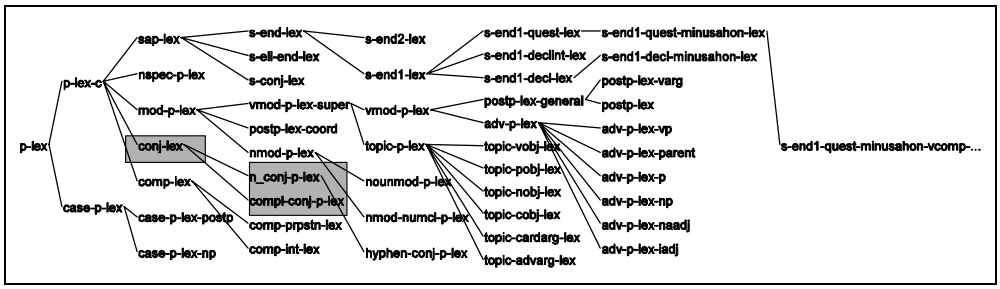


Figure 96

6.4.4 Noun phrase conjunctions



Conjunctions are part of the particle type hierarchy, because they share several peculiarities with particles: They attach to nouns and postpositions, they are head-final, and they have the same distributional behaviour as (other) particles.

A basic question of coordinative structures is what kind of information is unified in the coordination and if it is underspecified in some way. Sag (2003) discussed this topic for non-parallel coordination. Japanese noun phrase coordination can be non-parallel in subject honorific information, as can be seen in Example 183. In these examples, honorification information is different on the two conjuncts. The subject honorification agreement with the main verb refers to the last conjunct.

Example 183[48]

a) 花子 と 田中 先生 が 本 を お買いになった
Hanako to Tanaka sensei ga hon wo o-kai-ni-natta
Hanako CONJ Tanaka Prof. NOM book ACC buy - HON

(Hanako and Prof. Tanaka bought the book.)

b) *田中 先生 と 花子 が 本 を お買いになった
Tanaka sensei to Hanako ga hon wo o-kai-ni-natta
Tanaka Prof. CONJ Hanako NOM book ACC buy - HON

(Prof. Tanaka and Hanako bought the book.)

Four solutions to this can be thought of:

1. The entire HEAD information on the conjunct is the HEAD information of the second conjunct. If this information is inconsistent, there is no coordinative N-N structure.

[48] Thanks to Akira Kusamoto for verifying the examples.

2. The HEAD information on the conjunct is the unification of the HEAD information of the two conjuncts minus the syntactic honorific information in FORMAL.SHON.
3. There is no such thing as an NP conjunction in Japanese. The construction rather contains two PPs modifying the verb, as being proposed by Kasai and Takahashi (2001).
4. Use Sag's (2003) idea of introducing disjunctive types, such that the honorification information is neutralized in coordinations.

Before we discuss these possibilities, let me give another example of non-parallelism in the scope of numeral classifiers:

Example 184

犬	と	猫	が	三匹	います。
inu	to	neko	ga	sanbiki	imasu
dog	CONJ	cat	NOM	3 animals	exist

There are two possibilities to interpret this example:[49]

1. There are cats and dogs, amounting to three animals.
2. There are three cats with some dogs.

As we describe in Chapter 5.6, the floating classifier scopes over the verbal subject in this example. Thus, in the first interpretation, the subject is the coordination of the animals, while in the second interpretation, only the cats – the second conjunct – is counted.

Using Sag's idea of neutralization of the information in disjunctive types would lead to problems when we want to rule out Example 183 b) above. Further, the formal properties of such an approach are quite unclear and hard to restrict. The idea that there is no coordinated construction leads to an unintuitive semantic interpretation of the examples. Further, in the case of Example 184, we would not get the second interpretation where the scope of the numeral classifier includes both conjuncts. The idea that the coordination unifies the HEAD information except for the honorific agreement information would not explain the two different interpretations that we get for Example 184 with the numeral classifier.

Thus, we propose to unify the entire HEAD information of the conjuncts in nominal coordinations. If this information is inconsistent, as in the given examples on honorification, we do not have a case of coordinated noun phrases. Rather, the particle *to* is then interpreted as a modifying particle with the meaning of "with". In the example of numeral classifier interpretation, we can clearly see the ambiguity of analyses, where interpretation 1) results from the coordination analysis and interpretation 2) results from the "with" interpretation.

Nominal conjunctions are part of the particle type hierarchy, because they share several peculiarities with particles: They attach to nouns and postpositions, they are head-final, and they have the same distributional behaviour as (other) particles.

Coordinative structures are handled by binary tree structures in our grammar, as described in Chapter 2. We repeat the basics of the phrase structure rule analysis here:

- The *binary-type-conj* inherits from the general type for (binary) modification *binary-modification-type.*
- The conjunction rule type (*conj-rule-type*) with its rule instance conj-rule makes use of the HEAD feature C-MOD.
- C-MOD takes a list value of a list with no or one item on it.

[49] Thanks to Chikara Hashimoto for evaluating this example.

- Coordinative inflections or particles get the information about the type they combine with in C-MOD.
- The conjunction rule accesses this information and unify CAT, CONT, BAR, NUCL, LEX and NON-LOCAL in C-MOD of the conjunction with the next conjunct.
- The conjunction rule takes Head and Valence of the right conjunct and Index and LTOP of the first conjunct, combines the CONTEXT information and restricts both conjuncts to be saturated.

In order to provide this structure, we need a conjunctional particle that adds conjunctional semantics, linking the nominal indices that participate in the conjunction. This is provided by the type *n_conj-p-lex*. It contains the information in its head that it can participate in a nominal conjunction:

C-MOD: < [LOCAL.CAT.HEAD *noun_head*] >

The Head information on the two noun phrases it conjoins is unified. The semantics it adds contains three indices:

1. C-ARG which is the main index of the conjunct and therefore accessible to subcategorization, modification and another conjunction.
2. L-INDEX which refers to the index of the left conjunct NP.
3. R-INDEX which refers to the index of the right conjunct NP.

The example with non-parallel honorification above (Example 183) gets a correct analysis: In a), there is only the "with" reading of the particle *to*, as the SHON information of the two NPs does not unify. In b), the conjunctive structure is not applied either for the same reason. Further, there is no "with" reading for *to*, as the second NP does not unify with the restrictions on the honorific main verb.

For the example that includes a numeral classifier (Example 184), we give two readings: The first reading is a conjunctive structure, where the conjunction *to* combines the noun phrases. This refers to the meaning 1: "*There are cats and dogs, amounting to three animals*.", where the conjoined index is counted. The subject floating numeral classifier identifies the external arguments XARG of the cardinal it takes as a specifier and the main verb it modifies. The cardinal adds a *const-relation* to the MRS that contains an ARG1 showing to its XARG (and through the identification by the numeral classifier to the external argument of the main verb). The case particle *ga* sets its XARG to its complement's Index, which is the Index of the conjunction. Therefore, the subject floating numeral classifier identifies the predicate's XARG with the conjunction Index, such that the cardinal counts just this.

The second reading assumes that *to* is not a conjunction, but a preposition with the meaning of "with". This refers to meaning 2: "*There are three cats with some dogs*".

Example 185: Nominal coordination

猫	と	犬	が	います
neko	to	inu	ga	imasu
cat	CONJ	dog	NOM	be

(There are a cat and a dog.)

Example 186: Nominal coordination with second conjunction

猫	と	犬	と	が	います
neko	to	inu	to	ga	imasu
cat	CONJ	dog	CONJ	NOM	be

(There are a cat and a dog.)

Japanese has the special behaviour of allowing a second conjunctive particle attached to the second conjunct, as in Example 186. This second conjunctive particle is a sub-type of *mod-p-lex*, *postp-lex-coord*. It contains a Head of type *postp_head* and can therefore be subcategorized by a case particle (*ga* in Example 186). It subcategorizes for a noun head with a coordination index and adds nothing to the MRS. All coordination indices are instances of the type *conj-ref-ind*, such that they can be semantically subcategorized. The resulting MRS for both examples is the same.

```
h4: _inu_n_rel(x5)
h6: udef_rel(x5)
h9: _to_p_and_rel(x11, x5, x10)
h14: udef_rel(x11)
h17: _neko_n_rel(x10)
```

Figure 97: MRS for nominal coordination and nominal coordination with two conjunctions

In the case of more than two conjuncts, we keep the general binary construction policy and let the second conjunction refer to the C-ARG of the first conjunct in its L-INDEX.

The conjunction *matawa* can combine complement sentences, thus taking a complementizer and referring to its top handle LTOP, as in Example 187. This entry gets the type *compl-conj-p-lex* in the type hierarchy of conjunctional particles.

Example 187

申し込み	が	されていない	のか、	または、	事務	手続き
moshikomi	ga	sareteinai	noka,	matawa,	jimu	tetsuzuki
request	NOM	not be	COMPL	CONJ	work	plan

上	遅れて	いる	のか、	確認	して	いただきたい
jou	okurete	iru	noka,	kakunin	shite	itadakitai
viewpoint	fall behind	AUX	COMPL	confirm	do	want to

(I want to confirm, whether there is a request or if you fell behind your workplan.)

6.5 Omitted particles

Some particles can be omitted in Japanese spoken language. Here are three examples from the Verbmobil corpus:

Example 188

六月	十三日	の	火曜日	∅	午後	から
rokugatsu	juusannichi	no	kayoubi	∅	gogo	kara
June	13th	GEN	Tuesday	∅	afternoon	from

は	いかが	でしょう	か
wa	ikaga	deshou	ka
TOP	good	COP	QUE

(Would the 13th of June suit you?)

Example 189

先生	∅	ご都合	の	ほう	は	いかが	でしょう	か
sensei	∅	go-tsugou	no	hou	wa	ikaga	deshou	ka
Prof.	∅	HON-circumstances	GEN	sid	TOP	good	COP	QUE

(Would that suit you?)

Example 190

今	の	所	∅	午後	は	なにも	予定	が
ima	no	tokoro	∅	gogo	wa	nanimo	yotei	ga
now	GEN	time	∅	afternoon	TOP	no	plan	NOM

入って	おりませｎ	ので
haitte	orimasen	node
inserted	HON_NEG	SAP

(Up to now I have no plans for the afternoon.)

This phenomenon can be found frequently in connection with pronouns and temporal expressions in the domain of appointment scheduling. Hinds (1977) assumes that exclusively *wa* can be suppressed. Yatabe (1993) however shows that there are contexts, where *ga*, *wo* or even *e* can be omitted. He assigns it as `phonological deletion'. Kuroda (1992) analyses omitted *wo* particles and explains these with linearization: A particle *wo* can only be omitted, when it occurs directly before a verb. Yatabe (1993) however gives examples to prove the opposite. One of these shall be shown here. He assigns it as `slightly awkward but acceptable':

Example 191

どの	学生	∅	おれ	が	殴った	か	覚えてる
dono	gakusei	∅	ore	ga	nagutta	ka	oboeteru
which	student	∅	I	NOM	hit	QUE	remember

(Do you remember which student I have hit?)

The Verbmobil data of Japanese dialogues does not contain information about phonological phenomena of pitch. It is therefore not possible at this stage to include this kind of information into our analysis. However, it is peculiar that quite often pauses occur instead of particles. This hints at a phonological phenomenon.

In the above examples, it can be observed that NPs without particles can fulfil the functions of a verbal argument (Example 191) or of a verbal adjunct (Example 188, Example 189, Example 190). Therefore, the **pp_np_rule_case** is a unary rule that assumes *ga* (nominative) or *wo* (accusative) case and changes the head type of the noun phrase to **empty-case-p_head**, a subtype of **case-p_head**.[50]

.

[50] Though, in current distributions of the grammar, this rule is commented out. It is only useful in parsing spoken language and adds massive ambiguity.

6.6 Evaluation of case and modifying particles

We randomly chose 100 sentences out of the Verbmobil spoken dialogue data on appointment scheduling. 83 of them got a parsing result.[51] We then observed the accuracy of the analysis of the occurred particles. See Table 15.

Table 15: Analysis of particles

occurred particles	167	100%
correctly analyzed	153	91.6%
wrongly analyzed	14	8.4%

The main problem was that adjuncts were bound to the wrong predicate where more than one predicate occurred in the data. The combinations of particles that occurred in the test data were: *kara ga, kara de mo, de wa, kara wa, ni wa, de* wa and *de mo*. As can be seen in Table 16, all combinations of particles were correctly analyzed.

Table 16: Analysis of combinations of particles

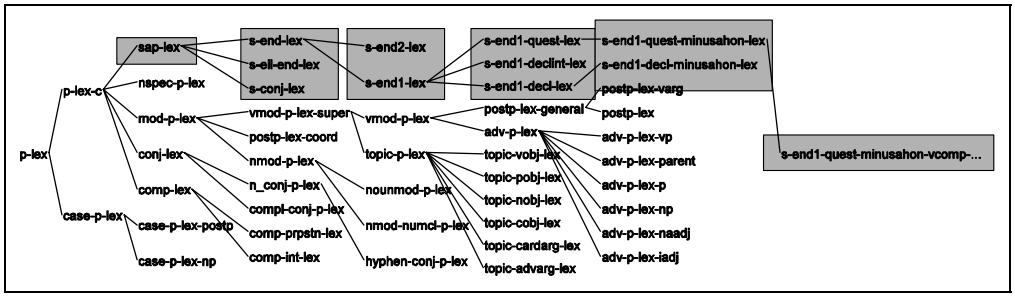
occurred combinations of particles	7	100%
correctly analyzed combinations of particles	7	100%

In 17 cases particles were missing (see Table 17).Three cases were wrongly analyzed. The problem was the same as with particle marked adjuncts: The adjuncts were bound to a wrong predicate in the case were more than one predicate occurred. One parse failed, because of an unexpected missing particle.

Table 17: Analysis of Missing Particles

occurred missing particles	17	100%
correctly analyzed	14	82.35%
wrongly analyzed	3	17.6%

6.7 Sentence particles



In Japanese a quite frequent way to end sentences is by the use of sentence end particles. Sentence end particles can express the speaker's attitudes, such as emotions, doubt, emphasis, caution, hesitation, wonder, or admiration. Some sentence ending particles distinguish male or female speech. Sentence end particles can assign the (declarative or interrogative) sentence mood or be conjunctive.

[51] Failures came from transliterational problems (6 cases), mistakes in lexical entries (5 cases), very complex genitive modification (1 case) and the occurrence of wh-words without question markers (2 cases).

Sentence particles are part of the type hierarchy of particles, located in **p-lex-c**, the semantic contributing particles. Their HEAD and VALENCE types are subtypes of HEAD and VALENCE types for particles as well, taking a **sentence-valid** complement. Thus, they build the head of the sentence.

Three basic types of sentence particles are distinguished in the type hierarchy:

1. Sentence conjunctions (**s-conj-lex**), conjoining sentences in coordinative structures.
2. Sentence end particles (**s-end-lex**), ending sentences and expressing the speaker's attitude.
3. Elliptical sentence particles (**s-ell-end-lex**), ending sentences with a conjunction, leaving the inferring of the second conjunct to the addressee.

Sentence conjunctions, such as *ga*, *keredomo*, *node*, are quite similar to nominal conjunctions. The **conj-rule** that we described above applies to these as well. In the MRS, they refer in L-HNDL and R-HNDL to the propositions on top of the conjoined sentences and add a proposition on top of themselves. The type for these conjunctions in the type hierarchy is **s-conj-lex**.

Sentence end particles add a relation to the MRS. This can be a propositional declarative message (such as *keredomo*), a questional message (such as *ka*) or a tag question (such as *ne*). Furthermore, they add the scope restriction of this message to outscope the main verb. An example output can be seen in Figure 98.

Most of the sentence end particles take the addressee honorification from their complement, i.e., the sentence. Some, though, mark the utterance as honorific with regard to the addressee. Examples for these are *kana, kashira, na*. They are of type **s-end2-lex** and add an AHON + restriction to the HEAD.

```
h1: question_m_rel(h10)
h3: _gohan_n_rel(x4)
h5: udef_rel(x4, h6)
h8: _taberu_v_rel(e2,u9,x4)
h6 qeq h3
h10 qeq h8
```

Figure 98: MRS for *gohan wo tabemashita ka* (rice ACC eat-past question-particle)

Sentence end particles can add sentence mood information, such as declarative (*keredomo, kedo, yo* etc.), tag question mood (*ne, kane, yone* etc.) or interrogative (*no* and *ka*). For these, we have the types **s-end1-decl-lex**, **s-end1-declint-lex** and **s-end1-quest-lex**, respectively. Those particles that are used by male speakers in extremely informal situations get the HEAD information of AHON – and the empathy (in CONTEXT) set to the speaker INDEX. We found examples of declarative sentence end particles such as *i* and interrogatives such as *nokai* or *kai*.

Elliptical sentence particles (**s-ell-end-lex**) leave some inference to the addressee, when ending the sentence. An example for these is given in Example 192. Therefore, we add a subordinate predication to the MRS which is an arg1-relation with the PRED value of "ellipsis". The elliptical particle then functions like a conjunction, conjoining the expressed sentence and an elliptical predication.

Example 192

花子	が	ご飯	を	食べた	ので
Hanako	ga	gohan	wo	tabeta	node
Hanako	NOM	rice	ACC	ate	because

(Because Hanako ate the rice…)

7 Adverbs

Japanese genuine adverbs are a non-inflecting class. An example is given in Example 193.

Example 193

直接　　聞いて　みます
chokusetsu　kiite　mimasu
directly　hear　try

(*I will directly ask*)

Though, other part-of-speech classes can inflect to behave like adverbs. Adjectives in the continuative form can be used as adverbs:

Example 194

弱い　→　弱く
yowai　→　yowaku
weak　　weakly

We therefore need a derivational rule that changes the category of adjective to adverb. A derivational rule called **adj2adv-lexeme-infl-rule** does this job. It changes the adjective ending い (*i*) to く (*ku*), makes the head type of the result to be an **adv_head** (therefore modifying predicates), copies PRED, CONTEXT, CONT.HOOK and NONLOCAL from the adjective stem entry and makes the result an intersective adverb (**isect-adv-lex**). The semantics of the resulting adverb thus contains an ARG0 and an ARG1 (just as the adjective) (see Figure 99).

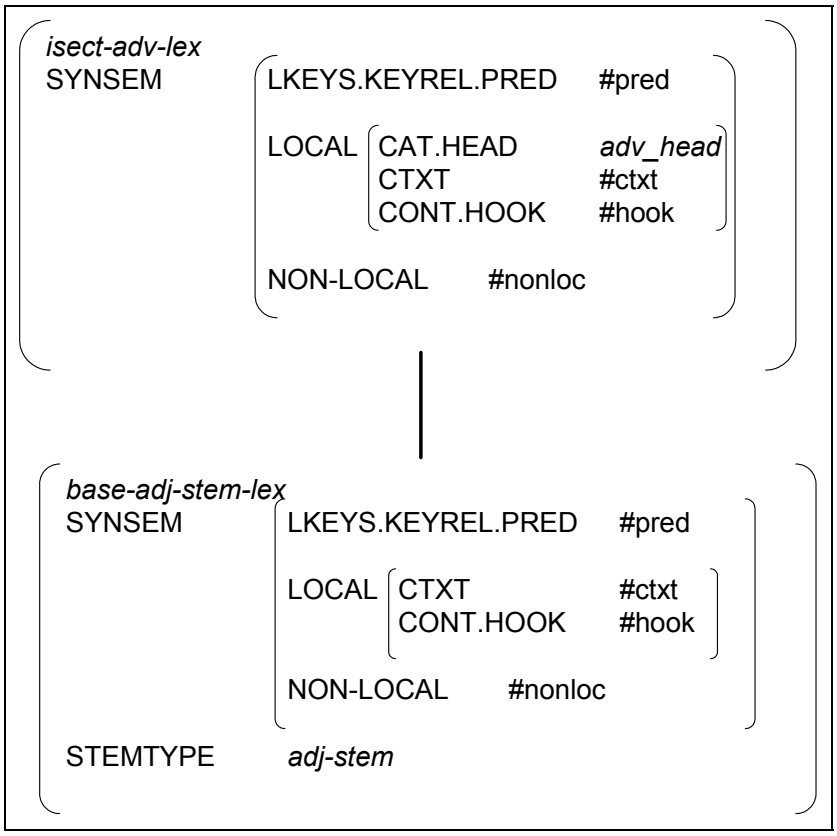


Figure 99: adj2adv-lexeme-infl-rule

Verbs in the infinitive form get the particle *ni* to behave adverbial:

Example 195

見る　→ 見に
miru　→ mi-ni
to see　　in order to see

The particle gets the task of semantically linking the predicated events.

Attaching *ni* gives na-adjectives adverbial behaviour:

Example 196

かんたん　に　できます
kantan　ni　dekimasu

(one can do it easily)

Here as well a particle *ni* semantically links the predicated events.

Temporal nouns can behave just like adverbs:

Example 197

今日　東京　に　行きます
kyou　Tokyo　ni　ikimasu
today　Tokyo　to　go

(Today I go to Tokyo)

A unary rule (**adv_np_rule**) takes these temporal nouns and gives them an adverbial head with intersective predicate modification. It adds an underspecified modification relation between the noun and the modified predicate.

Japanese adverbs typically modify the entire sentence, head-finally. Though, as in other languages, Japanese adverbs can be used to modify verbs, adjectives, other adverbs or sentences.

From a lexical-semantic viewpoint, Japanese adverbs can be grouped into three categories:

1. Adverbs of manner, e.g. *hayaku* (quickly)
2. Adverbs of time, e.g. *ima* (now)
3. Adverbs of degree, e.g. *zuibun* (quite)

Other languages have additionally adverbs of place, which is expressed by nominal categories and particles in Japanese.

Adverbs can be modifiers or complements (the MOD feature can, but does not have to be regarded). As modifiers, they can be intersective or scopal. In order to account for this, we make use of the type hierarchy: the type **adv-lex** contains subtypes **scopal-adv-lex** and **isect-adv-lex**. They add the basic form of the key relation and a modification type which can be called by the head-adjunct-rules to select the appropriate one.

Intersective modification is modification of the sentence. E.g., *sorosoro tabetai* is ambiguous between *I want to eat soon* and *I soon want to eat*. The intersective modifier takes the sentence event as its argument. Scopal modification on the other hand is modification of the predicate. E.g.: *tabun tabetai* clearly means *it's probable that I want to eat* (see Kasper 1995 for further explanations). Scopal modification takes a handle that outscopes the main predicate. Intersective modification is by far the most common.

```
sorosoro taberu
h4: _sorosoro_a_rel (e2)
h4: _taberu_v_rel (e2, x1,x2)
```

Figure 100: Intersective modification

```
tabun taberu
h4: _tabun_a_rel(e2, h5)
h6: _taberu_v_rel (e2, x1,x2)
h5 qeq h6
```

Figure 101: Scopal modification

There are no comparative and superlative forms of adjectives. Rather, there are degree adverbs modifying adjectives (scopal modification) that are comparative and superlative.

Example 198: comparative

もっと　安い
motto　yasui
more　cheap

(cheaper)

Example 199: superlative

一番　　安い
ichiban　yasui
most　　cheap

(cheapest)

8 Head-Initial Constructions in a Head-Final Language[52]

Japanese is generally taken to be strictly head-final in its syntax (Gunji, 1987). Broad claims like this can be tested by implementing grammars for large fragments of the language and testing them against naturally occurring text. In our work on a broad-coverage, precision implemented HPSG for Japanese, we have found a few minor exceptions to the broad trend towards head-final order in Japanese.

8.1 The position of syntactic heads in Japanese

Zwicky (1993) identifies several characteristics which have been taken to differentiate heads and dependents, and points out that they do not correlate all that well.[53]

	Head	Dependent
Semantics	characterizing	contributory
Syntax	required	accessory
	word rank	phrase rank
	category determinant	non-determinant
	external representative	externally transparent
Morphology	morpho-syntactic locus	morpho-syntactically irrelevant

Table 18: Characteristics of head and dependents, from Zwicky 1993

HPSG theory only recognizes some of these characteristics in the identification of syntactic heads,[54] namely required v. accessory, category determinant v. non-determinant, and external representative v. externally transparent. The central intuition is that the syntactic head of a construction is that subconstituent which determines the syntactic distribution of the whole. This notion of head is, of course, fundamental to HPSG and is encoded in the head-feature (Pollard and Sag, 1994) and subcategorization (Borsley 1993) principles. Given an HPSG grammar, the head of any constituent parsed by the grammar is well-defined. The HEAD values encode precisely the kind of part of speech information which determines the syntactic distribution of an element (such as case, preposition form, and modification possibilities) and the head feature principle propagates this information to the mother of the phrase. Likewise, the subcategorization principle distinguishes heads from arguments, in general making the valence requirements of a phrase some function of the valence requirements of its head.[55] Determining which element is the head for the purposes of writing the grammar, on the other hand, can be trickier. Deciding on the head constituent in a phrase requires observing which constituent contributes the head information and the subcategorization information. By this definition, it is true that most heads in Japanese follow both arguments and adjuncts: Verbs appear at the end of clauses, as can be seen in Example 200.

[52] This chapter is an extended version of joint research with Emily Bender, published in Siegel and Bender (2004).

[53] In modifier constructions, the semantic functor is not the head, but the modifier, cf. Zwicky 1993.

[54] Note that the syntactic head need not be the semantic head.

[55] In some cases these `functions' get fairly elaborate and also refer to the valence requirements of the non-head daughter, as in argument transfer and composition in constructions like that combining verbal nouns and light verbs in Japanese.

Example 200

田中　が　本　を　読んだ
Tanaka ga hon wo yonda
Tanaka NOM book ACC read-past

(Tanaka read a book.)

Adjectives, genitives, and relative clauses precede nouns:

Example 201

田中　の　やさしい　友達　が　来た
Tanaka no yasashii tomodachi ga kita
Tanaka GEN nice friend NOM come-past

(Tanaka's nice friend came.)

The language has postpositions, including both contentful elements such as *kara* `from' (3), and the case marking postpositions *ga, wo, ni* (4), which both follow nouns.

Example 202

東京　から　来た
Toukyou kara kita
Tokyo from come-past

(someone) came from Tokyo.

Example 203

何時　から　が　よろしい　です　か
nanji kara ga yoroshii desu ka
What time from NOM good COP QUEST

(From what time would be good?)

That contentful postpositions should head their phrases is relatively uncontroversial. Applying the same treatment to the case markers might be more surprising, especially as they are sometimes considered to be nominal inflection (e.g., Sag et al., 2003). We have, however, already discussed Japanese particles in Chapter 6 and shown that case particles should best be treated as heads as well. We illustrate the argument here with the examples in Example 202 to Example 205, which show that *ga* is crucial in determining the combinatoric potential of its phrase.

Example 204

何時　から　集まります　か
nanji kara atsumarimasu ka
What time from gather QUEST

(From what time are people gathering?)

Example 205

*何時　から　が　集まります　か
nanji kara ga atsumarimasu ka
What time from NOM gather QUEST

In Example 203, there is a single constituent (*nanji kara ga*) containing both a contentful postposition (*kara* `from') and a case-marking postposition *ga*. Constituents ending in *kara*

are verbal adjuncts (Example 202 and Example 204). When *ga* attaches, the result is eligible to appear in an argument (here, subject) position (Example 203), and no longer can appear as a verbal adjunct (Example 205). If *ga* were merely a marker that otherwise preserved the category information of the constituent it attaches to, this behaviour would be hard to explain. Note that on this analysis, the Japanese case particles look fairly similar to English `case-marking prepositions', such as *to* in *Kim gave the book to Sandy.* For our purposes here, the main point is that PPs, with both contentful and case marking postpositions, are also head-final.[56] We now turn to the exceptions we have found to the general head-final trend, which can be classified into two groups: head-initial modification and head-initial complementation.

8.2 Head-initial modification

8.2.1 Data

Using the definition above of the syntactic head in a construction, we can find some elements that behave as non-heads, although they occur final in a construction. In this class, we find the modifiers *dake, nomi, bakari* (in two distinct uses), *goro, kurai, hodo*, and certain instances of numeral classifiers.

8.2.1.1 Dake

The modifier *dake* `only' modifies at least NPs, predicative PPs, and adverbs. The noun-modification use is illustrated in Example 206:

Example 206

a. 野村さん　だけ　が　来た
Nomura-san dake ga kita
Ms. Nomura only NOM come-past

(Only Ms. Nomura came)

b. 野村さん　が　来た
Nomura-san ga kita
Ms. Nomura NOM come-past

(Ms. Nomura came)

c. *だけ　が　来た
dake ga kita
only NOM come-past

The head of the construction *Nomura-san dake ga* is the case particle *ga* (see above). The head of *Nomura-san dake* must be *Nomura-san*, because *ga* selects for a noun. Leaving *dake* out in this construction leads to a grammatical sentence *Nomura-san ga kita*, while leaving *Nomura-san* out gives an ungrammatical sentence. *Dake* is optional in all registers, the noun is obligatory in all, and the case particle is obligatory in some. Therefore we conclude that *dake* in this construction is a modifier to *Nomura-san*, even though it follows the head.

---

[56] In general, distinguishing morphology and syntax is not very clear-cut in this agglutinating language (Shibatani and Kageyama, 1988; Kageyama, 2001). For better or for worse, the orthography does not provide any clues, lacking inter-word spaces. For practical (engineering) purposes, we tend towards regarding syntax over morphology, as ChaSen provides near-morpheme-level segmentation. Along the way, we will point out evidence that the cases presented here involve syntactically separate words (clitics or otherwise).

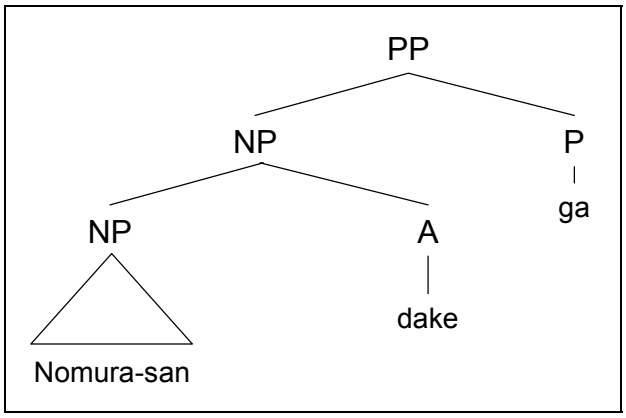


Figure 102: Structure of PP with *dake*

The predicative PPs modifier use of *dake* is illustrated in Example 207:

Example 207

利用者	は	東京	から	(だけ)	ではない
Riyousha	wa	toukyou	kara	(dake)	de-wa-nai
Users	TOP	Tokyo	from	(only)	COP-NEG

(The users were not only from Tokyo.)

The fact that *dake* is optional in this example lends support to the conclusion that *toukyou kara dake* is a head-initial construction. Further support comes from the fact that the order of the particles is flexible, as illustrated in Example 208 (from Makino and Tsutsui, 1986, 95).

Example 208

a.	この	車	は	アルコ・ル	で	だけ	動けます
	kono	kuruma	wa	arukouru	de	dake	arukemasu
	this	car	TOP	alcohol	INST	only	move

(This car runs only on alcohol.)

b.	この	車	は	アルコ・ル	だけ	で	動けます
	kono	kuruma	wa	arukouru	dake	de	arukemasu
	this	car	TOP	alcohol	only	INST	move

(This car runs on alcohol alone.)

As indicated in the glosses, *dake* can modify (semantically as well as syntactically) either the NP or the PP. It can appear in either position without affecting combinatoric potential. Thus, *arukouru de dake* and *arukouru dake* are head-initial.

Finally, adverbs can also be modified (head-initially) by *dake*, as illustrated in Example 209 (from Makino and Tsutsui, 1986, 94).

Example 209

私	は	日本	へ	いちど	(だけ)	行った
Watashi	wa	nihon	e	ichido	(dake)	itta
I	TOP	Japan	to	once	(only)	went

(I went to Japan (only) once.)

To summarize the observations for *dake*, we can say that it combines with (at least) NP, PP, and ADV to form a category of the same type. The relative non-specificity of the host suggests a syntactic rather than a morphological combination. The distributional facts support

treating *dake* as a non-head, even though it is final in its constituent.[57] A second element, *nomi* `only', is very similar to *dake*, except that it cannot follow adjectives and quantifiers. It is used in formal speech and written Japanese, but seldom in the registers found in our corpora.

8.2.1.2 Bakari `only'

Our second example is *bakari* `only'. It can modify PPs and VPs (or possibly Vs). Consider first the example in (Example 210a), from the newspaper *Mainichi Shinbun*. Here, *bakari* is a PP modifier:

Example 210

a. 衝突 に ばかり 関心 が 集まった
Shoutotsu ni bakari kanshin ga atsumatta
collision to only concern NOM collected

(It is only on the collision that concern is concentrated.)

b. 衝突 に 関心 が 集まった
Shoutotsu ni kanshin ga atsumatta
collision to concern NOM collected

(It is on the collision that concern is concentrated.)

c. *衝突 ばかり 関心 が 集まった
Shoutotsu bakari kanshin ga atsumatta
collision only concern NOM collected

In this example, the particle *ni* `to' determines the combinatoric potential of the whole phrase, leaving *bakari* the role of a modifier. There are also examples of head-initial verb modification, including the following attested in *Mainichi Shinbun* in 2002:

Example 211

学校 の 先生 を 怒らせて ばかり いた
Gakkou no sensei wo okorasete bakari ita
school GEN teacher ACC upset only AUX

(The only thing he was doing was upsetting the school's teacher.)

This is one exception to the general rule that nothing should intervene between a verb in the *-te* form and an auxiliary. The exception can be handled if *bakari* modifies *okorasete*. We therefore introduce one instance of *bakari* that can be a post-head modifier of verbs with *-te* inflection.

8.2.1.3 *Bakari* and other forms meaning `about'

There is another post-head modifier *bakari* meaning `about', which modifies temporal expressions. We illustrate it here with another *Mainichi Shinbun* example:

[57] Makino and Tsutsui (1986) also note a use of *dake* where it attaches to verbs and adjectives to make nominal constituents. In this case, *dake* appears to be a nominalizing head and the examples are not relevant to the point at hand.

Example 212

東京	から	車	で	二時間	ばかり	の	混交	の	温泉	に
Toukyou	kara	kuruma	de	nijikan	bakari	no	kinkou	no	onsen	ni
Tokyo	from	car	INST	2 hours	only	GEN	suburb	GEN	hotspring	to

朝	七時	ごろ	出発する
asa	shichiji	goro	shuppatsu-suru
morning	7 o'clock	around	depart

(We depart at about 7 a.m. for a hotspring in the suburbs which is about twohours from Tokyo by car.)

The relevant construction here is *nijikan bakari no*. The head of the construction is *no,* because it carries the information that the construction can modify an NP. *No,* in turn, selects for the temporal noun *nijikan* and *nanjikan* is modified by *bakari*. The sentence would be perfectly grammatical without *bakari*. Similarly, for *goro, kurai* and *hodo* (about), one finds several examples for head-initial modification of temporal expressions, such as Example 213:

Example 213

今日	何時	ごろ	まで	寝て	いました	か
kyou	nanji	goro	made	nete	imashita	ka
today	what time	about	until	sleep	AUX-past	QUEST

(Until about what time did you sleep today?)

Leaving out *goro* in (Example 214a) simply removes the `approximate' meaning from the sentence, while leaving out *nanji* (Example 214b) changes the meaning drastically: *Goro* becomes a modifier of *kyou*. Leaving out *made* (Example 214c) gives the sentence another meaning, `At about what time did you fall asleep today?'. Leaving out both *goro* and *made* gives `At what time did you fall asleep today?'

Example 214

a.

今日	何時	まで	寝て	いました	か
kyou	nanji	made	nete	imashita	ka
today	what time	until	sleep	AUX-past	QUEST

(Until what time did you sleep today?)

b.

今日	ごろ	まで	寝て	いました	か
kyou	goro	made	nete	imashita	ka
today	about	until	sleep	AUX-past	QUEST

(Were you sleeping until about today?)

c.

今日	何時	ごろ	寝て	いました	か
kyou	nanji	goro	nete	imashita	ka
today	what time	about	sleep	AUX-past	QUEST

(At about what time did you fall asleep today?)

d. 今日 何時 寝て いました か
kyou nanji nete imashita ka
today what time sleep AUX-past QUEST

(At what time did you fall asleep today?)

Once again, we see a modifier (*goro*) which can attach to multiple different constituents. Unlike *made*, *goro* does not affect the way the constituent it is attached to interacts with the rest of the sentence. Therefore, we propose the structure in Figure 103 for *nanji goro made.*

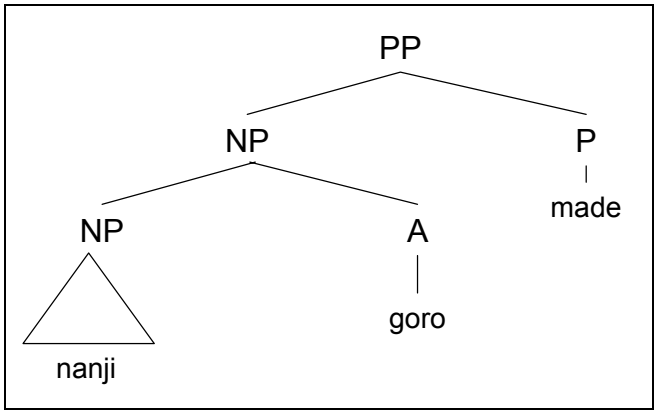


Figure 103: Structure of *nanji goro made* `until about what time'

8.2.1.4 Numeral classifiers

Finally, on our analysis, numeral classifier phrases appearing between a noun and its case particle or immediately after a case particle are post-head modifiers. Some examples are given in Example 215. See Bender and Siegel (2004), as well as Section 5.6 for further details.

Example 215

a. 猫 二匹 を 飼う
neko nihiki wo kau
cat 2 NumCl ACC raise

((I) am raising two cats.)

b. 猫 を 二匹 家 で 飼う
neko wo nihiki ie de kau
cat ACC 2 NumCl house LOC raise

((I) am raising two cats in my house.)

8.2.2 Summary

In this section, we have seen post-head modification of nominal, postpositional, adverbial and verbal constituents. Many of the modifiers can modify multiple different parts of speech. Others (numeral classifier phrases) are internally complex (potentially containing arbitrarily large number names) and further more can appear before or after the phrases they modify, or `floated' away from them (Bender and Siegel, 2004). These properties suggest that we are dealing with a syntactic rather than morphological phenomenon.

8.2.3 Analysis

Our analysis for head-initial modification consists of:

1. A lexical type hierarchy containing types that allow for head-initial constructions.

2. Grammar rules for head-initial modification and head-initial complementation.

3. A head feature POSTHEAD that is referenced by head-adjunct rules.

Figure 104 shows part of the type hierarchy of lexical signs, containing lexical items that modify nouns, postpositions and verbs, and which are divided into left-modifying and right-modifying items.

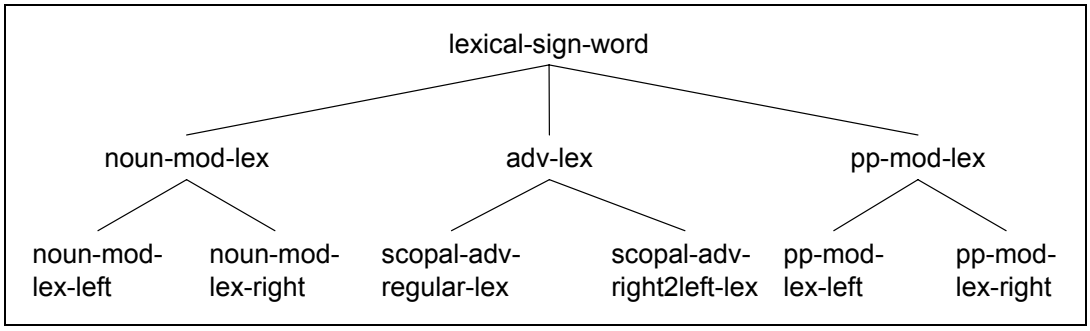


Figure 104: Partial hierarchy of lexical types for modifiers

The inventory of grammar rules contains rules for both head-initial and head-final complementation, which differ in the order of the daughters. The rules reference the HEAD.POSTHEAD value of the modifier daughter in order to constrain the distribution of lexical items across the constructions. POSTHEAD can be *left* or *right*, or can be left unspecified for those items that can modify in both directions.[58]

Head-initial modifier rules (scopal or intersective) bear these constraints, where the feature ARGS encodes the daughters of the rule and the order in which they appear:

```
HEAD-DTR        #1
NON-HEAD-DTR #2 [… CAT | HEAD | POSTHEAD right]
ARGS    < #1,#2 >
```

Figure 105

Modifiers of type *pp-mod-lex-right*, etc., are constrained to be [POSTHEAD *right*], and are compatible with head-initial modifier rules. In contrast, *pp-mod-lex-left*, etc., are [POSTHEAD *left*] are thus incompatible with head-initial modifier rules. In principle, modifiers could be underspecified for POSTHEAD, thus appearing on either side. Our lexicon does not currently contain any such modifiers.[59]

8.3 Head-initial complementation

8.3.1 Data

We have found two clear cases of head-initial complementation, the first in number names and the second in numeral classifiers. In both cases, one optional argument follows the head. We argue that number names like *ni hyaku juu* `210' are head-medial on the basis of examples like Example 216 and Example 217. Example 216b and Example 216c each share one

[58]We also use POSTHEAD for the selection of relative clause constructions, coordinated structures and the head selection of nominal compounds (see Radford, 1993 for criteria on head selection in nominal compounds).

[59]It might appear that numeral classifiers would constitute a case of modifiers attaching either to the left or the right of their heads. However, in pre-head uses of numeral classifiers there is always an intervening no (genitive) particle. We treat this particle as a head which selects for a numeral classifier phrase and mediates the modification of the noun by the numeral classifier. For details, see Bender and Siegel (2004).

element in common with Example 216a. The examples in (Example 217) show that the external distribution of these phrases differs.

Example 216

a. 二 百 十
ni hyaku juu
two hundred ten

b. 五 百 三
go hyaku san
five hundred three

c. に 千 三
ni sen san
two thousand three

Example 217

a. 六 千 に 百 十
roku sen ni hyaku juu
six thousand two hundred ten

b. 六 千 ご 百 三
roku sen go hyaku san
six thousand five hundred three

c. *六 千 に 千 三
roku sen ni sen san
six thousand two thousand three

d. *六 千 ご 千 十
roku sen go sen juu
six thousand five thousand ten

Expressions with *hyaku* (Example 216a and Example 216b) have the same combinatoric potential. Expressions without *hyaku* differ. The other elements of Example 216 *ni `two'* and *juu* `ten' are not relevant. Thus, we take *hyaku* to be the head of Example 216. If we forget for the moment that Japanese is supposed to be head-final, this isn't very surprising: English number names work the same way (see Smith 1999). So do number names in another SVO language: Chinese, the source from which Japanese borrowed this system. One might argue that this is actually a morphological process, in which case the head-medial structure is less surprising. However, Martin (1987) finds that while some local combinations within number names (e.g., the names for 11 through 19, 20, 30, 200, 300, etc.) form single phonological words, longer combinations made up of these pieces (such as *sanbyaku juuichi* `311') show phrasal phonology.

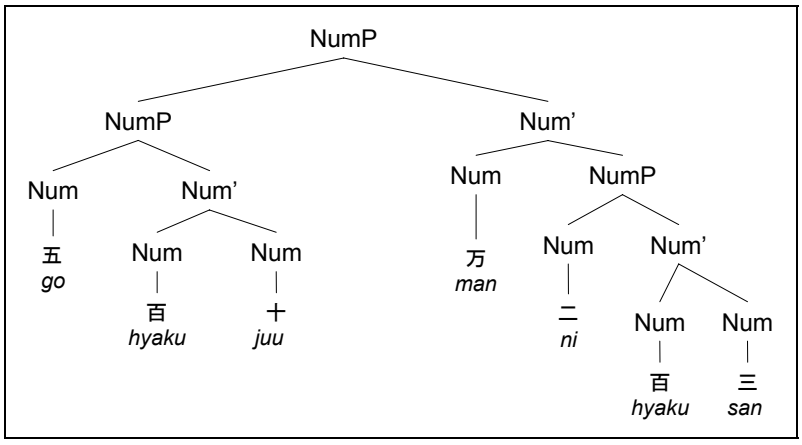


Figure 106: Center recursion in number name expressions

Moreover, number names show the sort of center recursion that distinguishes context-free languages from regular languages (see Figure 106). This kind of recursion is (to our knowledge) unattested elsewhere in morphology. The analysis presented here was developed within the context of an application that takes text-based input. As such, it was most convenient to apply the phrasal analysis uniformly. A similar analysis could be developed that provides lexical entries for every combination that forms a phonological word. It would still involve head-initial structures: In a phrase like *sanbyaku juiuichi*, the phonological words are *sanbyaku* (`three hundred') and *juuichi* (`eleven'). Following the same argumentation as above, *sanbyaku* (and within it, *hyaku*, meaning `hundred') determines the distribution of the phrase within larger number names.

Example 218

a. 一千 [三百 十一]
issen [sanbyaku juuichi]
one thousand three hundred eleven

b. *五百 [三百 十一]
gohyaku [sanbyaku juuichi]
five hundred three hundred eleven

c. 五百 十一
gohyaku juuichi
five hundred eleven

The second type of head-initial complementation involves numeral classifiers. All numeral classifiers combine with a number name to their left, but certain mensural numeral classifiers such as *nen* `year' can also take the word *han* `half' to their right Example 219. Syntactically, the numeral classifier determines the combinatorics of the phrase (being able to modify nouns, not being able to show up as the specifier of a larger number name). The presence or absence of *han* has no effect on the distribution. The numeral classifier is also in a better position to integrate the semantics of *han* than vice versa (Bender and Siegel, 2004).

Example 219

a. 二 年 半
ni nen han
two years half

b. 二 年
ni nen
two years

8.3.2 Analysis

Our analysis of both of these instances of head-initial complementation consists of:

1. Two head-complement rules, differing in the order of the daughters, and sensitive to the HEAD type of the head.

2. A high-level distinction in the sub-types of *head* into *init-head* and *final-head.*

The two head-complement rules are sensitive to the head type of their head daughter. Most head types are subtypes of *final-head,* giving the general pattern, while numeral classifiers and number names are given subtypes of *init-head.*

In the case of head-initial complementation, we do not posit an additional feature, but instead take advantage of the type hierarchy and posit a split between initial heads and final heads. Most head types inherit from *final_head*, including *noun-or-case-p head* (subsuming nouns and the case particles), *verb_head*, and *p_head*, for the contentful post-positions. The two subtypes of *init_head* are *int_head* (for number names) and *num-cl_head* (for numeral classifiers). The latter point is a bit subtle: The only numeral classifiers that take complements at all are those that can appear with *han* (as a complement).[60]

As the classification into *final head* and *init head* is only referenced by the head-complement rules, it is simplest to make them all *init head*. The following constraints on the two head-complement rules capture the necessary contrast:

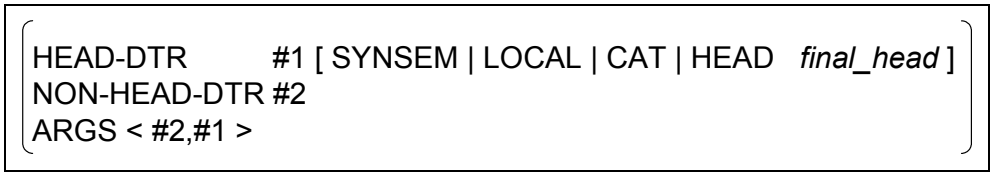


Figure 107: Head-complement-head-final-rule

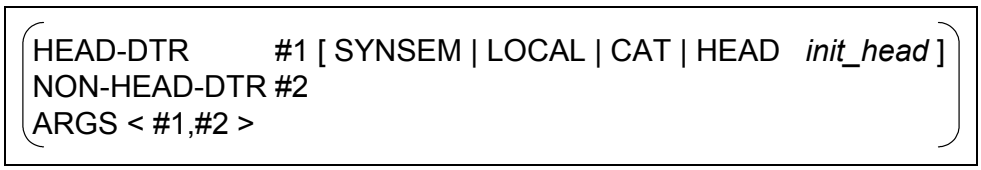


Figure 108: Head-complement-head-initial-rule

The ordering constraints relating HEAD-DTR, NON-HEAD-DTR, and ARGS are inherited from a supertype that is also applicable to the head-modifier cases. In our current implementation, there are no head types which are indeterminate between *init head* and *final head.* All head

[60] We have actually found it convenient to posit one more kind of numeral classifier which takes a complement: namely currency symbols such as `$', which appear to the left of a numerical expression but otherwise function syntactically and semantically like currency words such as doru and en, which appear to the right of a number name. Most numeral classifiers select their dependent number name.

types inherit from exactly one of these. It would of course be possible to cross-classify the ordering dimension with the part of speech dimension, should this be necessary, if some elements of a certain head type preceded their complements and others followed or if all elements of some head type could appear in either order with respect to their complements. Our investigations so far suggest that this is not the case for Japanese. It might be relevant for another language with relatively free order in general, but with some heads showing a more fixed order.

9 Honorification

Spoken language encodes references to the social relation of the dialogue partners. The utterances can express social distance between addressee and speaker and third persons, which are mentioned. Honorifics can even express respect concerning entities of the world. Consider the following examples from Japanese, German and French:

Example 220: German

```
wann  haben  Sie  Zeit
when  have   you  time
```

Example 221: French

```
quand  est-ce que  vous  avez  du   temps
when   do          you   have  the  time
```

Example 222: Japanese

```
いつ  ご都合          が   よろしい  でしょう  か
itsu  go-tsugoo       ga   yoroshii  deshoo    ka
when  HON-conditions  NOM  good      COP       QUE
```

(When are the conditions (hon, i.e. your conditions) good?)

The semantic content of these utterances is: 'When does it suit you?'. But there is an additional pragmatic content: The speaker expresses social distance concerning the addressee. This is expressed by the honorific pronouns *Sie* and *vous* in German and French. In the Japanese example it is expressed by the following attributes:

- The honorific prefix *go* in front of *tsugoo*
- The honorific adjective *yoroshii*
- The honorific copula *deshoo*

A Japanese utterance with the same semantic content in – for example – a family context could be:

Example 223

```
いつ  時間   が   ある  の
itsu  jikan  ga   aru   no
when  time   NOM  have  QUE
```

Information about honorification is – on the one hand – necessary for the description of syntactic phenomena like honorific agreement or relative sentences and – on the other hand – necessary for correct translation. In order to understand the whole meaning of the Japanese utterances it is important to represent the different honorific attributes in the analysis structure. The information can be used to resolve zero pronominalization and topicalized structures. It is even more important for the adequate *generation* of the Japanese utterances. In other investigations on zero pronoun resolution in task-oriented dialogues (Siegel 1996b) we calculated that 23.9% of the zero pronouns can be solved using lexical pragmatic restrictions about honorification.

9.1.1 Honorific forms in Japanese

Honorifics in Japanese express the social relation of familiarity or distance between speaker, addressee and third persons. Consider the situation where the speaker waits for Ms. Tanaka. In a familiar context (s)he would say:

Example 224

田中　　さん　を　待つ
Tanaka　san　wo　matsu
Tanaka　Ms　　ACC　wait

(I wait for Ms. Tanaka.)

In a more formal situation with more social distance between speaker and addressee the utterance would be:

Example 225

先生　　を　お待ちします
sensei　wo　o-machi-shimasu
Prof.　　ACC　hon-wait-hon

(I wait for the professor.)

The person that is waited for is referred to with her title. The predicate gets the 'humble' extension *o…shimashita*.

The social relationships that can be expressed are threefold: The first one is the relation between speaker and addressee, in the above example expressed by *no* and *ka* and the verbal endings *–ta* and *–mashita.* The second one is the relation between the speaker and the subject of the utterance, in the above example expressed by the verbal form and the prefix. The third one is the relation between the speaker and other arguments in the sentence. For example, the noun book (*hon*) can get the honorific *go*-prefix, if it is a book belonging to the addressee being honoured by the speaker.

Familiarity or distance between speaker and addressee can be expressed by verbal endings and/or the lexical choice of self-referring pronouns. Verbal endings encoding a relation of distance between speaker and addressee can be, for example, *-masu*, *-mashita* or *-n-deshoo-ka*. Those encoding a familiar relation can be, for example, *-ru*, *-ta* or *-no*. The choice of self-referring pronouns also depends on the gender of the speaker. A self-referring pronoun uttered by a woman in a familiar context could be *watashi*, a self-referring pronoun uttered by a woman in a distant context could be *watakushi*. Parallel, the appropriate self-referring pronoun for a man in a familiar context would be *boku,* one in a distant context would be *watashi*. I will call the relationship of honorifics concerning the relation between speaker and addressee **AHON** and give it a polarity [AHON -] for the plain form in a family context and [AHON +] for the expressions in a context of social distance.[61]

The social relation between the speaker and a subject that is not referring to the speaker is expressed by the lexical choice of verbs, by the expression *o-VERB-ni-naru*, by the honorific prefix *o/go* at nouns referring to entities belonging to the subject and by the lexical choice of pronouns. I will call this relation between speaker and subject SHON. A relation of distance between speaker and subject (where the subject is the addressee or a third person) can be – for example – expressed by the verb *irassharu* (to go), while in a familiar situation the verb *iku* with the same semantic content is used. This is expressed by [SHON +] and [SHON -],

[61] [AHON +] is the relation traditionally referred to as “Teineigo”.

respectively.[62] Possible referring expressions for the second and third person can be, for example, *sochira* and *X-san* in relations of distance and *kimi* or *X-kun* in relations of familiarity.

The third relation is the one between speaker and other entities in the sentence (other than subject). I will call this relation **EHON**. It is expressed by the lexical choice of these entities and by the honorific prefixes *o* and *go*.[63]

9.1.2 Interaction of different kinds of honorification in Japanese

The relationship between speaker and addressee can be one of three possible constellations:

1. The addressee is the subject of the utterance.
2. The speaker is the subject of the utterance.
3. A third person is the subject of the utterance.

When the addressee is the referent of the sentence subject, the relationships AHON and SHON must have the same polarity. Thus, in this case, in a sentence with AHON there must also be SHON.

In the situation, where the speaker is the subject of the utterance, (s)he uses humble forms of the verbs (a matter of lexical choice), if the AHON relation is a distant one. An example is *mairu* (to go). In this case, both relationships (SHON and AHON) are concerned.

In many cases utterances contain multiple honorifications as can be seen in the following example:

Example 226

私	が	お電話	いたしました
watakushi	ga	o-denwa	itashimashita
I	NOM	telephone	do(hon)-hon-Past

The verbal stem *itashi* expresses subject honorification (with negative polarity), the verbal ending *mashi* and the pronoun *watakushi* addressee honorification.

Japanese honorification undergoes different kinds of restrictions. The first kind to mention is called 'pragmatic agreement' by Pollard and Sag (1994). There must be agreement between the SHON honorification of the subject and the verb, as the following examples show:

Example 227

私	が	先生	に	お電話	いたしました
watakushi	ga	sensei	ni	o-denwa	itashimashita
I	NOM	professor	NI	telephone	do(humble-shon)-ahon-Past

Example 228

*先生	が	私	に	お電話	いたしました
*sensei	ga	watakushi	ni	o-denwa	itashimashita
professor	NOM	I	NI	telephone	do(humble-shon)-ahon-Past

[62] [SHON +] is traditionally referred to as “Sonkeigo” and [SHON -] as “Kenjôgo”.

[63] [EHON +] is part of the traditional category “Sonkeigo”.

Example 229

先生	が	私	に	お電話	なさいました
sensei	ga	watakushi	ni	o-denwa	nasaimashita
professor	NOM	I	NI	telephone	do(honorific-shon)-ahon-Past

The pronoun *watashi* can be used with a humble verb form, but not the noun *sensei*, which refers to a honourable person. This must be used with an honorific verb form.

Another kind of restriction concerns relative sentences as opposed to complement sentences. See the following examples from Harada (1976):

Example 230

太郎	は	花子	が	来ました	と	言った
Taro	wa	Hanako	ga	kimashita	to	itta
Taro	TOP	Hanako	NOM	come-hon-Past	COMPL	say-Past

Example 231

太郎	は	花子	が	来た	と	言った
Taro	wa	Hanako	ga	kita	to	itta
Taro	TOP	Hanako	NOM	come-Past	COMPL	say-Past

Example 232

*太郎	は	花子	が	来ました	こと	を	しらなかった
*Taro	wa	Hanako	ga	kimashita	koto	wo	shiranakatta
Taro	TOP	Hanako	NOM	come-hon-Past	NOM	ACC	not know - Past

Example 233

太郎	は	花子	が	来た	こと	を	しらなかった
Taro	wa	Hanako	ga	kita	koto	wo	shiranakatta
Taro	TOP	Hanako	NOM	come-Past	NOM	ACC	not know - Past

Complement sentences allow an honorific predicate (addressee honorification, as expressed by the verbal ending), while relative sentences do not.

9.1.3 Previous approaches

Investigations of Japanese honorification have been made from the sociolinguistic, the grammatical and the machine translational viewpoint. For the sociolinguistic viewpoint see for example Ide (1986), Coulmas (1987), Hori (1986), Hill et al. (1986) and McGloin (1976). The authors state that honorification is an expression of the social distance or 'perceived distance' (Hill et al. 1986) between speaker and addressee and the belonging to a social group (Coulmas 1987). They investigate the relation between gender and the use of honorificational expressions (Hori 1986). Examples for a grammatical investigation of Japanese honorification are Ikeya (1983), Kuno (1973) and Harada (1976). Hori (1986) uses honorification for a definition of `subject' in Japanese. Kuno (1973) classifies honorification concerning style and honorification concerning respect. In our approach, these classes are AHON and SHON, respectively. He shows that there are differences of grade in various expressions of honorification. Harada (1976) gives a classification of honorificational forms that at first sight seems complementary to ours. It can be seen in Figure 109.

A closer look shows that the difference is only a question of naming (see Figure 110). Harada's `Subject honorifics' is [SHON +] in our approach, the `Object honorifics' is [SHON -] and the `Performative honorifics' would correspond to our [AHON +]. What we call

[EHON] turns into [SHON], if the entity is used as a subject in the utterance. Ikeya (1983) gives a GPSG account for honorification, where [SHON +] and [SHON -] (called EHON in his approach) are head features, with the head feature principle accounting for the agreement restrictions on subject honorification. Gunji (1987) also gives examples for syntactic restrictions on honorification and introduces HON as a head feature.

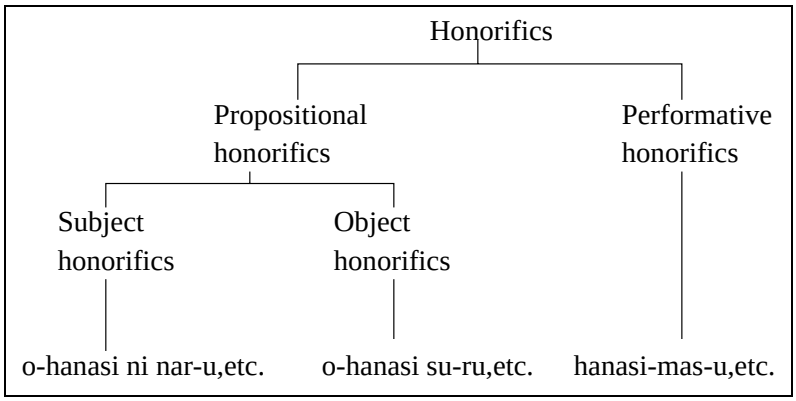


Figure 109: Classification of honorifics by Harada

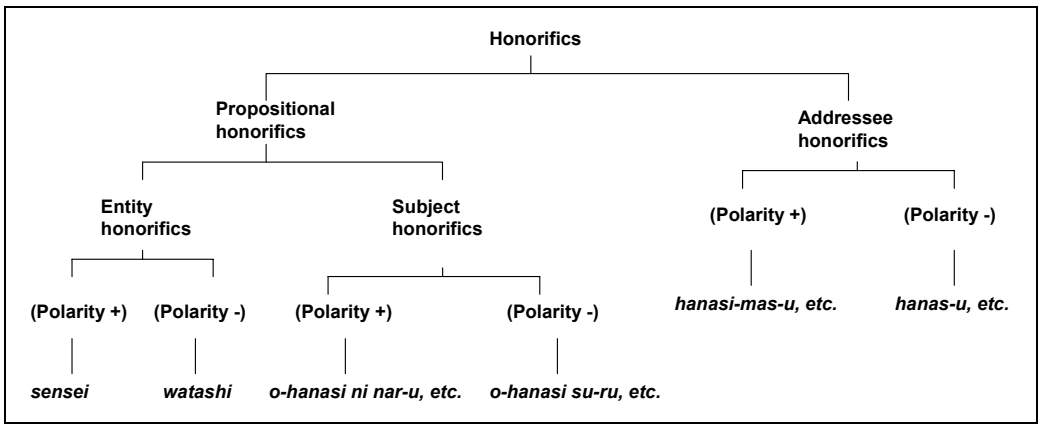


Figure 110: Classification of honorifics in JACY

The machine translational viewpoint is shown by Dohsaka (1990). He describes how information about honorification can be used in a machine translation system to resolve zero pronominal references to human entities. He builds up a model of social relations during processing the dialogue, where the pragmatic relations honorification, speaker's point of view and territory of information is on the one hand extracted from the dialogue and on the other hand restricts the interpretation of zero pronouns in the dialogue. This approach shows that the extraction of information about honorification from the dialogue is urgently needed for the interpretation of zero pronouns.

9.1.4 Japanese honorification in HPSG

Pollard and Sag (1994) analyze honorification as a pragmatic fact. They describe the problem as 'pragmatic agreement' and introduce a relation **owe-honour** in the BACKGR feature, as can be seen in Figure 111.

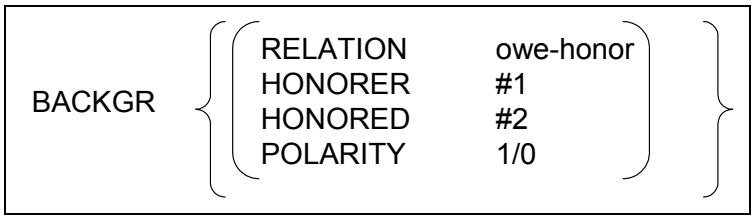


Figure 111

Still, the approach lacks for the fact that there are different kinds of honorification as we described above. It describes only subject honorification. Green (1997) elaborates the CONTEXT feature and introduces information about social ranking of the participants. We would, though, propose to leave the inference of the social relations to other components of, e.g., a machine translation system. The reason is that all necessary information is not always directly accessible in the analysis procedure. An example is given by Coulmas (1987): The secretary in a company is asked by an employee, when the boss comes back from a business trip. He or she would answer:

Example 234

来週	帰って	いらっしゃいます
raishuu	kaette	irasshaimasu
next week	come back	SHON +, AHON +

If the same secretary would be asked by a customer, the answer would be:

Example 235

来週	帰って	まいります
raishuu	kaette	mairimasu
next week	come back	SHON −, AHON +

In this example, we would represent the fact that the secretary honours the boss in the first example, but not in the second one. The interpretation of the complex social relations must be left to a module that has access to the information about the actual social relations of the participants in the context.

To account for the fact that Japanese honorification has more dimensions, we propose the following CONTEXT feature structure:

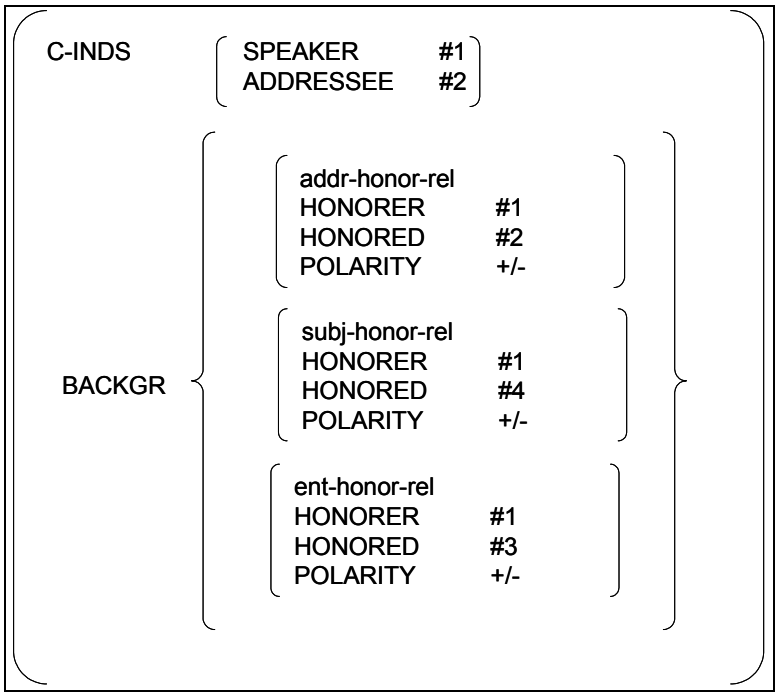


Figure 112

The C-INDS contain indices for speaker and addressee, as proposed by Pollard and Sag (1994). The value of BACKGR is a difference list that sums up the occurring honorificational relations in the utterance. Each occurring relation gets classified into **addr-honor-rel**, **subj-honor-rel** and **ent-honor-rel**. The HONORER is co-indexed with the speaker in all cases here. This must be different in cases of indirect speech that we will describe later. The HONORED value is co-indexed with the addressee in C-INDS in the **addr-honor-rel** case, with the subject's CONTENT.INDEX value in the **subj-honor-rel** case and with the CONTENT.INDEX value of the argument that introduces the relation in the **ent-honor-rel** case.

The relations all get a value of POLARITY, to account for the fact that there can be forms that are honorific, humble or neutral. A negative SHON polarity, e.g., reflects the situation where the speaker or a third person that socially belongs to the inner circle of the speaker is the subject of the utterance. McGloin (1987) describes this situation socio-linguistically as "positive politeness", because it expresses social closeness.

The question is: how does the information enter into the BACKGR? Let us start with the **ent-honor-rel**. This relation is encoded in the nouns that express honorification. The entry for *o-uchiawase*, e.g., contains:

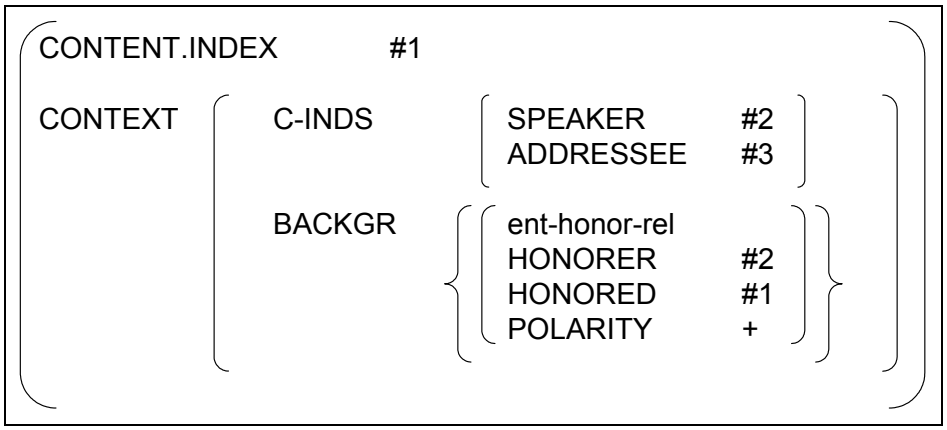


Figure 113: Honorification in the lexical type of *o-uchiawase*

A verb that restricts the honorification of its subject contains the following in its lexical entry:

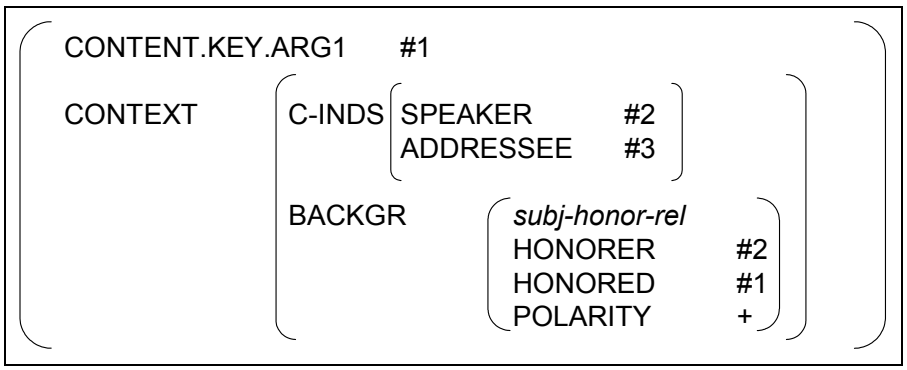


Figure 114: Honorification of a shon verb

If it happens to be the case that an entity with an **ent-honor-rel** in its BACKGR becomes the subject of the sentence, the mother node must get the **subj-honor-rel** from the predicate, identified with the index of that entity. It is, though, necessary to prove the honorification restrictions on predicate and subject. Since this is a syntactic process, we decided on representing honorification on the syntactic level, too.

Gunji (1987) gives reasons for the syntactic approach.[64] He describes in his JPSG-account of Japanese syntax honorification as a kind of agreement:

"Since Japanese does not have syntactic agreement phenomena such as number, person, etc., the honorification system is more or less a counterpart."

He introduces the feature HON as a head-feature (with values +/-), underlying the head feature principle. This accounts for the fact that the value of the HON-feature passes from the head to the mother node. Gunji's HON, though, is only a representation of subject honorification. Honorification concerning the addressee or sentence entities is not considered. The values of SHON can be either plus or minus, but neutral forms also exist.

Therefore, we expanded the syntactic part of the representation of honorification. The lexical entries get a HEAD feature called FORMAL:

[64] See also Ikeya (1983).

Figure 115

Only the connection of representing honorification on the syntactic **and** contextual level makes it possible to account for all phenomena. The pure syntactic representation cannot account for the representation of honorification relations between speaker and addressee, for EHON and for multiple honorifications, while the pure contextual representation cannot account for the syntactic restrictions on subjects and relative sentences. The CONTEXT level gives information about felicity of an utterance, while the CAT level gives information about syntactical correctness of an utterance. For honorification in Japanese, we need both. With the fundamental concept of HPSG, the sign, it is possible to incorporate **both** levels of linguistic analysis.

Being a HEAD feature, the value of FORMAL is passed up from head daughters to mother daughters. An honorific noun therefore contains the value SYNSEM|LOCAL|CAT|HEAD|FORMAL|SHON +, as well as a verb with subject honorification. For Japanese, we set up the principle of subject honorification:

In an honorific lexical structure, the FORMAL|SHON value of the HEAD is identical to the FORMAL|SHON value of the subject's HEAD and the polarity of the subj-honor-rel in BACKGR. The values of the subject's CONTENT|INDEX and the HONORED of the subj-honor-rel in BACKGR are identical.

This principle accounts for the compatibility of the information on the syntactic (CAT) and contextual (CONTEXT) levels. While the agreement of subject and verb is checked on the syntactic level, the contextual level gets the information on subject honorification and links it to the semantic entities.

Honorification concerning the addressee inside the sentence is seen as a purely syntactic restriction. As non-addressee-honorific and addressee-honorific verbs may combine, it is not useful to introduce the relation into the context during processing the sentence. The syntactic restriction is needed for relative sentences, as shown above. At the top-most node (**utterance-type** in our grammar), the **addr-honor-rel** is introduced into the CONTEXT|BACKGR. Its polarity is co-indexed with the value of HEAD-DTR| SYNSEM| LOCAL| CAT| HEAD|FORMAL| AHON. The HONORER is co-indexed with the speaker, while the HONORED is co-indexed with the addressee. Also here it can be seen that it is meaningful to represent the honorification on both levels. Inside the sentence, it is a purely syntactic relation, but outside, it is a contextual relation.

While the syntactic information goes up the tree via the head feature principle, the contextual information underlies different principles.

The HPSG principle of contextual consistency (Pollard and Sag 1994, p.333) says:

The CONTEXT|BACKGR value of a given phrase is the union of the CONTEXT|BACKGR values of the daughters.

This must be modified for our approach, since the head-subject rule takes the CONTEXT|BACKGR value of its head daughter. It can be hold for all structures besides the head-subject rule and the utterance rule, as shown before.

Let us take an example for multiple honorifications, Example 226, which shall be repeated here as Example 236:

Example 236

私	が	お電話	いたしました
watakushi	ga	o-denwa	itashimashita
I	NOM	telephone	do(hon)-hon-Past

The self-referring pronoun *watakushi* introduces an **ent-honor-rel** with POLARITY -, where HONORER and HONORED are co-indexed with the speaker and the CONTENT|INDEX. This is passed up the tree in the head-complement structure of watakushi ga}. At the same time, the values of HEAD|FORMAL are introduced: AHON + and SHON -. As particles are assumed to be heads (see Siegel 1999 and Chapter 6), they must take their SHON value from their complements (which is defined in the lexical type of particles).

The honorific form *o-denwa itashi-mashi-ta* introduces a **subj-honor-rel** in the context with POLARITY -. The HONORED is co-indexed with the subject's CONTENT|INDEX. The HEAD|FORMAL values are the same as the ones of *watakushi*. The principle of subject honorification sets up the restrictions for the predicate's subject. As this is found in *watakushi ga*, the HEAD|FORMAL values are unified and the **subj-honor-rel** is introduced. The **utterance-rule** introduces an **addr-honor-rel** with POLARITY +, since the value of HEAD|FORMAL|AHON is +.

This was an example of the special case where the speaker is the subject. Another example with the addressee being the subject is:

Example 237

あなた	が	お電話	を	くださいました
anata	ga	o-denwa	wo	kudasai-mashi-ta
you	NOM	telephone	ACC	do(shon)-ahon-Past

All three types of honorification relations are introduced here: subject honorification by the addressee-referring pronoun *anata*, entity honorification by the honorific noun *o-denwa* and addressee honorification by the *-mashita* ending of the verb. The polarity is + in all cases.

9.1.5 Effects

The CONTEXT|BACKGR value passes up the tree, independent of which daughter is the head of the phrase. It is even possible to represent the honorification in cases of embedded phrases. There can be more than one relation of **ent-honor-rel** in an utterance, as there can be more than one honoured constituents. An effect for the machine translation system is that lexical pragmatic restrictions for zero pronouns can be directly accounted for in the analysis. They are essential to find referents for many zero pronouns, as is shown by Metzing and Siegel (1994). See for example:

Example 238

お待ち	しております
omachi	shite-orimasu
wait	do-hon

(I am waiting.)

This is part of the structure for this utterance:

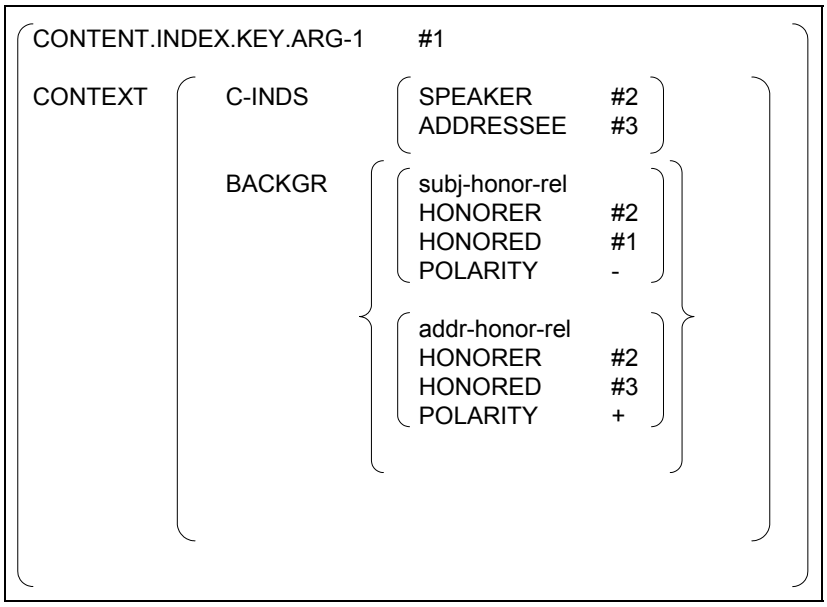


Figure 116

The structure restricts the subject to one with a subject honorification with negative polarity. That is, only the speaker or a person of the speaker's social group can be the antecedent of the subject.

Two occurrences of subject honorification can also be possible: Imagine a sentence where the complement sentence has a different subject honorification from the matrix sentence. E.g. in a sentence with indirect speech:

Example 239

私	が	スケジュール	を	立てたい
watashi	ga	sukejuuru	wo	tatetai
I	NOM	schedule	ACC	want to set up

と	おっしゃいました
to	osshaimashita
COMPL	say-addrhon-past

(I said that I wanted to set up a plan.)

Syntactic restrictions for relative sentences can easily be formulated in a way that only verbs with a non-addressee-honorific form can modify nouns.

9.1.6 Evaluation

We randomly chose 100 utterances from the Verbmobil corpus. Then we tagged these with expected values for SHON, EHON and AHON. The utterances contained 170 occurrences of honorification, with 99 AHON, 32 EHON and 39 SHON. We parsed the utterances and compared the human-made tagging with the parsing result. Then we calculated precision and recall in the following way:

$$\text{Precision} = \frac{\text{number of correct assigned honorifications}}{\text{number of assigned honorifications}}$$

$$\text{Recall} = \frac{\text{number of correct assigned honorifications}}{}$$

number of honorifications in the corpus

The results can be seen in Table 19.

	precision	**recall**
AHON	1	1
SHON	1	0.86
EHON	1	0.79
HON (sum)	1	0.93

Table 19

9.1.7 Honorification in other languages

Honorification in German concerns only the relation between speaker and addressee, as the following example shows:

Example 240

```
Sie  sind  nett
you  are   nice
```

The sentence is ambiguous, because it allows a first interpretation where *Sie* is a third person plural pronoun and therefore refers to a group of people and a second interpretation where it is a polite second person singular pronoun and refers to a single person. Honorification in German thus introduces EHON with honorific pronouns, but no special treatment of subjects and no AHON relation. The agreement between subject and verb is a purely syntactic one.

French honorification shows different habits in agreement (as is shown by Pollard and Sag 1994, p.96f.), but as well concerns only the EHON dimension.

Honorification in Korean, as it is described by Lee (1996), is distinct from Japanese honorification in one point: There are no neutral forms of NPs and VPs in respect to subject honorification.

Our approach thus seems to work for different kind of languages that express honorification.

10 JACY in Different Application Domains

During the development history of the grammar, different applications with different application domains were modeled:

- Appointment scheduling in machine translation of spoken language.
- Emails in the banking domain for an automatic email correspondence system.
- Parallel multilingual grammar development embedded in hybrid natural language processing.
- Dictionary definition sentences for treebanking and ontology extraction.

Each of these required specialized vocabulary and grammatical constructions, had different processing needs and got quite different input. In order to address these needs, the grammar had to be highly modular, easy to extend and flexible to configure in usage.

10.1 Appointment scheduling in machine translation

The grammar was first developed for the purpose of machine translation of spoken dialogs. Therefore, it had to deal with spontaneous spoken dialogue language, erroneous speech recognition input and unclear sentence boundaries. Utterances are relatively short, often fragmentary. Speakers make heavy use of sentence end particles and honorification to express his/her empathy and thoughts and to reflect the social situation.

There are three main features that are characteristic for turns in spoken dialogue language, as opposite to sentences in written language:

- They can be sequences of sentences.
- They can be fragments.
- They can obey syntactic rules that are not valid for written language.

The syntax distinguishes between sentences that contain the main syntactic components (mainly PP and VP) and utterances that have some reference to the hearer. This can be an honorific form or inflection or a sentence particle like *keredomo* or *ne*. A turn - on the other hand - cannot only consist of one utterance but also of a sequence of utterances:

Example 241

今日	は	です	ね	ちょっと	夕方	まで	授業	も	入ってます
kyou	wa	desu	ne	chotto	yuugata	made	jugyou	mo	haittemasu
today	TOP	COP	TAG	a bit	evening	till	lessons	too	inserted

し	で	午前中	も	鈴木	さん	と	打ち合わせ	が
shi	de	gozenchuu	mo	Suzuki	san	to	uchiawase	ga
CONJ	INTERJ	morning	too	Suzuki	Mrs./Mr.	with	meeting	NOM

入って	おります	ので	ちょっと	今日	という	の	は	むり
haitte	orimasu	node	chotto	kyou	toiu	no	wa	muri
inserted	HON-AUX	because	a bit	today	COMPL	NOM	TOP	bad
なんです	けれども							
nan desu	keredomo							
COP	SEP							

(Because today there are lessons until evening and in the morning there is a meeting with Mr./Mrs. Suzuki, today is a bit difficult)

There are two ways of sequencing: One is, to combine a finite sentence and a conjunction with other sentences. The whole conjunct must be an utterance in referencing to the dialogue situation. In the above example the finite sentence (*kyoo wa desu ne chotto yuugata made jugyoo mo haittemasu*) is conjoined with the rest via the conjunction *shi*. The other way is, to combine an utterance with others without a conjunction. As 'utterance' is defined for spoken language, it refers to the dialogue situation. Consider the following example:

Example 242

十七	日	の	火曜日	です	ね	そう	です	ね	一時	まで
juunana	nichi	no	kayoubi	desu	ne	sou	desu	ne	ichiji	made
17	day	GEN	Tuesday	COP	TAG	so	COP	TAG	1 o'clock	till

会議	そのた	あ り ま す	ので	一時	すぎ	から	でしたら
kaigi	sonota	arimasu	node	ichiji	sugi	kara	deshitara
meeting	and so on	exist	because	1 o'clock	after	from	COP-conditional

なんとか	予定	が	取れる	んです	が	そちら	の	ご都合
nantoka	yotei	ga	toreru	ndesu	ga	sochira	no	go-tsugou
somehow	plan	NOM	can take	COP	but	you	GEN	HON-convenience

は	いかが	でしょう	か
wa	ikaga	deshou	ka
TOP	good	COP	QUE

(That's Tuesday the 17th, isn't it? Well, until one o'clock there are meetings and so on, if it would be after one o'clock, I could somehow take some time, how is that for you?)

The first segment (*juu nana nichi no kayoobi desu ne*) has the honorific form *desu* of the copula and the tag-particle *ne*. So is the second segment (*soo desu ne*). The third one has the honorific form *ndesu* and the sentence particle *ga*.

The grammar accepts fragmentary input and is able to deliver partial analyses, where no spanning analysis is available. A complete fragmentary utterance from the Verbmobil corpus could, e.g., be:

Example 243

イナー シチィー ホテル
intaashitiihoteru
intercity hotel

This is just a noun, but there is still an analysis available that assumes a non-expressed predication. If another utterance is corrupted by not being fully recognized, the parser delivers analyses for those parts that could be understood. An example is the following best hypothesis from the speech recognizer in a system test:

Example 244

そう	です	ね	私	の	ほう	は	大じょうぶ	です	だが
sou	desu	ne	watakushi	no	hou	wa	daijoubu	desu	daga
so	COP	TAG	I	GEN	side	TOP	okay	COP	but

この	日	は	火曜日	です	ね
kono	hi	wa	kayoubi	desu	ne
this	day	TOP	Tuesday	COP	TAG

(lit.: Well, it is okay for my side, but this day is Tuesday, isn't it?)

Here, analyses for the following fragments are delivered (where the parser found *opera wa* in the word lattice, but not in the hypothesis):

Example 245

そう	です	ね	私	の	ほう	は	大じょうぶ	です	ね
sou	desu	ne	watakushi	no	hou	wa	daijoubu	desu	ne
so	COP	TAG	I	GEN	side	TOP	okay	COP	TAG

(Well, it is okay for my side.)

オペラ	は
opera	wa
opera	TOP

(The opera)

この	日	は	火曜日	です	ね
kono	hi	wa	kayoubi	desu	ne
this	day	TOP	Tuesday	COP	TAG

(This day is Tuesday, isn't it?)

Another necessity for partial analysis comes from real-time restrictions imposed by the Verbmobil system. If the parser is - due to time restrictions - not allowed to produce a spanning analysis, it delivers best partial fragments (see Kiefer et al. 2000 for further details).

The grammar must further be applicable to distinct phenomena of spoken language. A typical problem is the extensive use of topicalization and even omission of particles. Also serialization of particles occurs more often than in written language, as we described in Siegel (1999). A well-defined type hierarchy of Japanese particles is necessary here to describe their functions in the dialogues, as we described in Chapter 6.

Extensive use of honorification is another significance of spoken Japanese.

Compare the following sentences:

Example 246[65]

a) 十一　　日　　の　日　は　セミナー　が　一日中
juuichi nichi no hi wa seminaa ga ichinichijuu
11 day GEN day TOP seminar NOM whole day

入って　いる
haitte iru
insert progressive-AUX

(There is a whole-day seminar on the 11th)

b) 十一　　日　　の　日　は　セミナー　が　一日中
juuichi nichi no hi wa seminaa ga ichinichijuu
11 day GEN day TOP seminar NOM whole day

入って　おります
haitte orimasu
insert honorific-AUX

(There is a whole-day seminar on the 11th)

The first one is a syntactically correct sentence. In a dialogue though, it cannot be uttered in isolation, because it contains no reference to the dialogue situation. The second one refers to the dialogue situation via honorifics: The speaker as the agent of the utterance refers to himself with the 'humble' verbal form *orimasu* and such defines the social interaction between the dialogue participants as a distant one. *-masu* is a verbal flexion that also expresses social distance to the hearer. *keredomo* is a weakening particle.

A detailed description of honorification is necessary for different purposes in an MT system: honorification is a syntactic restrictor in subject-verb agreement and complement sentences. Furthermore, it is a very useful source of information for the solution of zero pronominalization, as was described in Metzing and Siegel (1994). It is finally necessary for Japanese generation in order to find the appropriate honorific forms. The sign-based information structure of HPSG is predestined to describe honorification on the different levels of linguistics: on the syntactic level for agreement phenomena, on the contextual level for anaphora resolution and connection to speaker and addressee reference, and with co-indexing to the semantic level. Our solutions to the generation and representation of honorific knowledge are described in Chapter 9.

Connected to honorification is the extensive use of auxiliary and light verb constructions that require solutions in the linked areas of morpho-syntax, semantics, and context.

Finally, a severe problem of the Japanese grammar in the MT dialogue translation task is the high potential of ambiguity arising from the syntax of Japanese itself, and especially from the syntax of Japanese spoken language. For example, the Japanese particle *ga* marks verbal arguments in most cases. There are, though, occurrences of *ga* that are assigned to verbal adjuncts, especially occurring in spoken language. Allowing *ga* in any case to mark arguments or adjuncts would lead to a high potential of (spurious) ambiguity. Thus, a

[65] Actually, in the Machine Translation project, the input to the grammar was Romaji (the romanized Japanese encoding), coming from the speech recognizer. Here we give the Japanese annotation for reading convenience as well.

restriction was set on the adjunctive *ga*, requiring the modified verb not to have any unsaturated *ga* arguments.

The Japanese language allows many verbal arguments to be optional. For example, pronouns are very often not uttered. This phenomenon is basic for spoken Japanese, such that the syntax urgently needs a clear distinction between optional and obligatory (and adjacent) arguments. We therefore used a description of subcategorization that differs from standard HPSG description in that it explicitly states the optionality of arguments, as being described in Chapter 3.

After Verbmobil, the grammar contained about 3000 lexical entries (full forms) in Latin writing, rules for the basic Japanese constructions (such as utterances and phrases, Nouns, Particles, Verbs, Adjectives, Copula, Adverbs, Honorification, Empathy, Zero Pronouns, Topicalization and Light Verb Constructions) and for special needs in Japanese spoken language processing.

Recently, there has been a new attempt to use JACY in Machine Translation: in Bond et al. (2005), we describe the use in an open-source MT prototype.

10.2 Emails in the banking domain

The email language contains short sentences and often fragments. Additionally, email language contains special abbreviations, greetings, tabular-like language and some punctuation. As usual with a shift of domain, there is a lexicon extension involved. Many idiomatic expressions and abbreviations had to be included, many of them as multiword expressions.

But first of all, we had to replace the Latin orthography in the lexicon with Japanese characters and further develop inflection and derivation rules in order to replace the fullform lexicon.

As Japanese written text does not have word segmentation, a preprocessing system is required. We integrated ChaSen (Asahara & Matsumoto 2000), a tool that provides word segmentation as well as POS tags and morphological information such as verbal inflection[66]. As the lexical coverage of ChaSen is higher than that of the HPSG lexicon, default part-of-speech entries are inserted into the lexicon. These are triggered by the part-of-speech information given by ChaSen, if there is no existing entry in the lexicon. These specific default entries assign a type to the word that contains features typical to its part-of-speech. It is therefore possible to restrict the lexicon to those cases where the lexical information contains more than the typical information for a certain part-of-speech. This default mechanism is often used for different kinds of names and 'ordinary' nouns, but also for adverbs, interjections and verbal nouns (where we assume a default transitive valence pattern). The ChaSen lexicon is extended with a domain-specific lexicon, containing, among others, names in the domain of banking.

For verbs and adjectives, ChaSen gives information about stems and inflection that is used in a similar way. The inflection type is translated to an HPSG type. These types interact with the inflectional rules in the grammar such that the default entries are inflected just as 'known' words would be.

Grammar extensions were done in collaborative work with Emily Bender, Stephan Oepen, Ulrich Callmeier and Daniel Flickinger (see Oepen et al. 2002a). The grammarians were working on different sides of the world, contributing to the grammar. There was a continuous

[66]Most of the technical implementation work in integrating ChaSen was done by Stephan Oepen and Ulrich Callmeier.

demand for improving coverage and quality of analyses, as the grammar was used in a commercial product of automatic email correspondence. Thus, the grammar was put into a CVS system in order to make submission and tracking of changes possible. A strong focus was set on collaboration with the grammar developers of the English and Spanish grammars, requiring a careful design and discussion of MRS output structures, such that the multilingual application could make use of them. The grammar was already in a good state of complexity, such that one had to be careful about side effects of changes. The adding of an auxiliary to the lexicon could have the effect of a massive increase of overall parsing ambiguity, for example. Thus, phenomena-oriented testsets were set up and an integrated competence and performance profiling was extensively used: The [incr tsdb()] system (Oepen and Carroll 2000).

The banking domain contains a variety of numeral expressions; such that it was necessary to extend and refine the analysis of numeral classifiers (see Section 5.6).

A phrase type that occurred quite regularly was two noun phrases with a colon, such as:

Example 247

ＩＤ ： １０

Very similar to this is the expression with a topic marker:

Example 248

ＩＤ は １０

This is analyzed by adding special rules for fragments into the grammar, which introduce underspecified events to the MRS.

Furthermore, an analysis of date expressions was added. Symbols like dash, arrow or brackets suddenly occurred in the data, such that we had to give an account for these. For example, we inserted "~“ and ":" as **case-p-lex-np** and **adv-p-lex-np** into the grammar, carefully observing their syntactic behaviour and semantic functions in the corpus.

Email language, which is a kind of semi-spoken language contains contracted verb forms such as *chatta, teru*, etc. and interjections.

As the grammatical coverage grew, the ambiguity rate did as well. So, a focus had to be set on reduction of spurious ambiguity, which arose in compounds and conjuncts and zero pronoun processing.

10.3 Parallel multilingual grammar development embedded in hybrid language processing

Two basic ideas of the DeepThought project (http://www.dfki.de/deepthought) had a high influence on the Jacy grammar development:

1. The development of grammars in a multilingual environment, where a focus was set on collaboration with developers of other grammars, in order to provide compatible and comparable grammar output from multiple languages.
2. The embedding of these grammars in a hybrid architecture (the Heart-of-Gold, see Callmeier et al. 2004) that allowed the connection of various NLP modules of different preciseness and robustness in different modes, but with a common and compatible output format, RMRS.

The idea of parallel multilingual grammar development was taken up in the DeepThought project. First, a Grammar Matrix was extracted from the Japanese and English grammars (see Flickinger and Bender 2003). The aim of this Grammar Matrix is to provide a common set of

lexical and rule types that can build the basis for the set-up of new grammars and furthermore guarantees that existing grammars being adapted to the Matrix obey MRS principles and provide a valid and useful output structure.

In the project (and connected to the project), new grammars of Norwegian (see Hellan and Haugereid 2003), Italian and Greek (Kordoni and Neu 2003) have been set up. The existing grammars of Japanese, English and German have been adapted to Matrix principles by inserting and connecting the Matrix types to the grammars. This process was bidirectional: In some cases, a grammar writer came up with a phenomenon that could not be described when using the Matrix types, such that the Matrix had to be revised.

Some changes to the JACY grammar that were necessary when including Matrix types were:

- **Naming Conventions for MRS Feature Names.** The feature naming conventions were made consistent with standard reference on MRS. Therefore, feature name replacements were necessary, as well as some re-orderings of features.
- **Definition of Connection Points between Matrix and Language Grammars**. The Matrix files were directly integrated into the Japanese grammar, such that connection points between matrix and language-specific grammars had to be defined. Discrepancies between definitions in the Matrix and the Japanese grammar have been identified and were subject to discussion about matrix revisions.
- **Introduction of the HOOK Attribute.** The externally visible attributes of an MRS are now grouped within a single attribute called HOOK, which is consistently used in constructions to identify the properties of the semantic head daughter with those of the phrase. The features in HOOK include the previously used LTOP (formerly TOP), INDEX, and E-INDEX, as well as a new feature XARG which is unified with the semantic index of the controlled argument of a phrase (to simplify the definition of e.g. equi and raising types).
- **Naming of Argument Roles (ARG1, ARG2, ARG3, ARG4).** Each relation now assigns its first (least oblique) argument to ARG1, its next argument to ARG2, and so on. The major change from previous grammar versions was to assign objects of transitive verbs to ARG2 rather than ARG3, and similarly for objects of prepositions.
- **Basic Relation Types**. The inventory of basic relation types has been simplified. New relation types had to be introduced to the grammars, such as a relation type for quantifiers (*quant-rel*). The basic relation type *`named-rel'* has also been incorporated into the grammars, and its inherent constant is now the CARG.
- **Ambiguity Packing**. To allow more efficient grammar processing, especially for complex input, the grammars were adopted to allow ambiguity packing (Oepen & Carroll 98) in the parser. This required defining and tuning suitable restrictors for each grammar that strike a good balance between the degree of packing in the parsing phase, and the number of failures in the unpacking phase.
- **Subcategorization**. A new multilingual approach to subcategorization was introduced into the Japanese and English grammars and tested. In order to give a direct encoding to the division of optional and obligatory arguments, as well as scrambling and adjacent arguments, the argument status is explicitly stated in an attribute OPT. This contains information about the saturation status of subcategorized arguments. It is an advantage of this approach that it provides a straightforward and easy-to-process way of dealing with scrambling and optionality of arguments. There are no lexical rules necessary that move arguments from valence to adjacency or slash lists, there is no

need for traces and slashs. We tested on the Japanese grammar that the treatment is still adequate for the phenomena associated with Japanese subcategorization.

- **ARG-S**. The Japanese grammar introduced the ARG-S feature (that was missing in previous versions), where verbal arguments are stored in lexical items. This is used for a general treatment of binding, argument raising and equi verbs, such that language-specific restrictions come from the mapping of arguments to COMPS and SUBJ lists, where multilingual restrictions can be stated on ARG-S.

The analysis output of the different languages shows the compatibility of the different language's grammars:

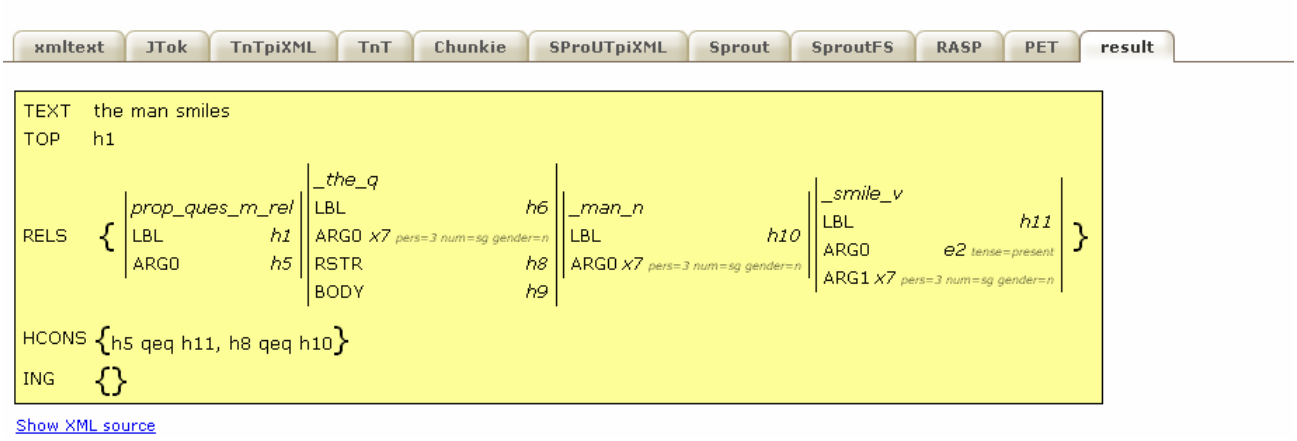


Figure 117: English

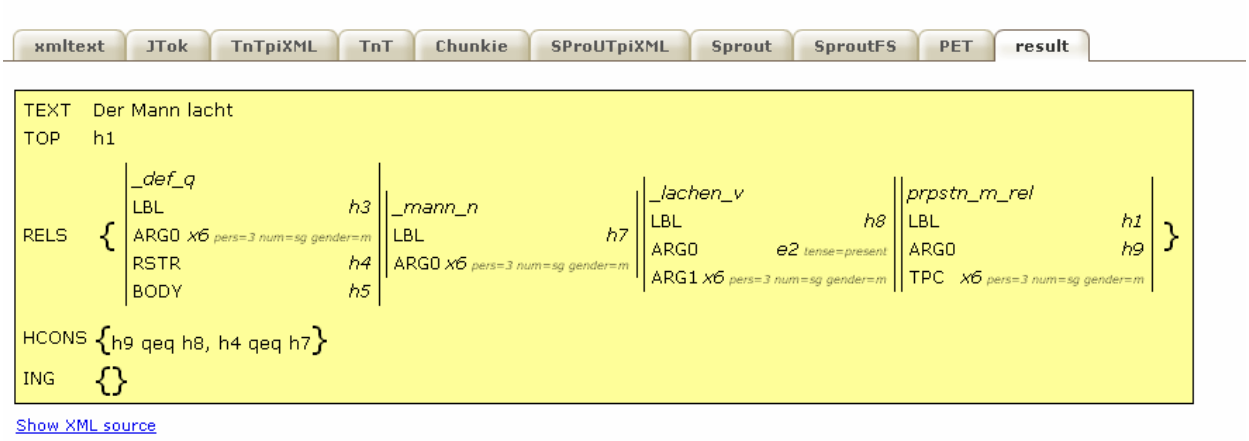


Figure 118: German

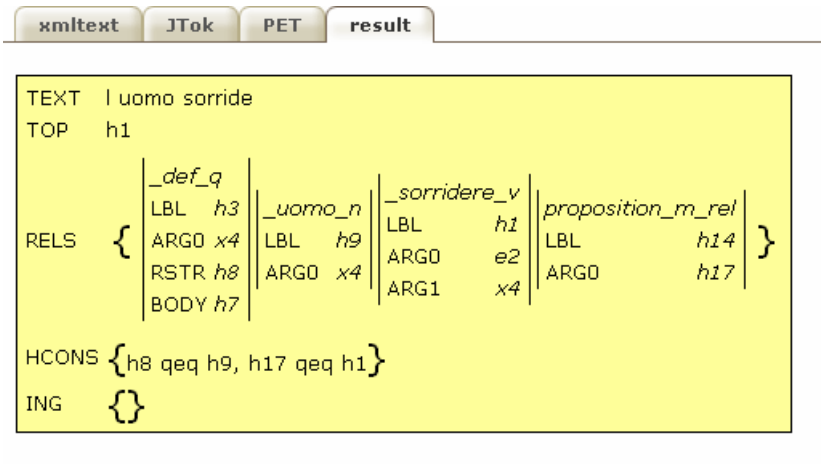


Figure 119: Italian

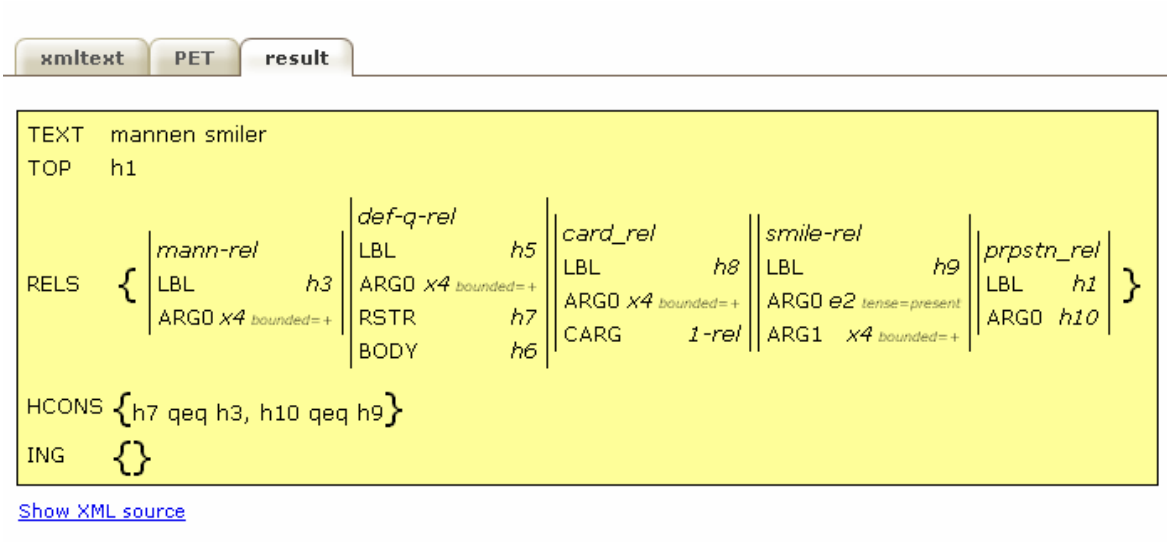


Figure 120: Norwegian

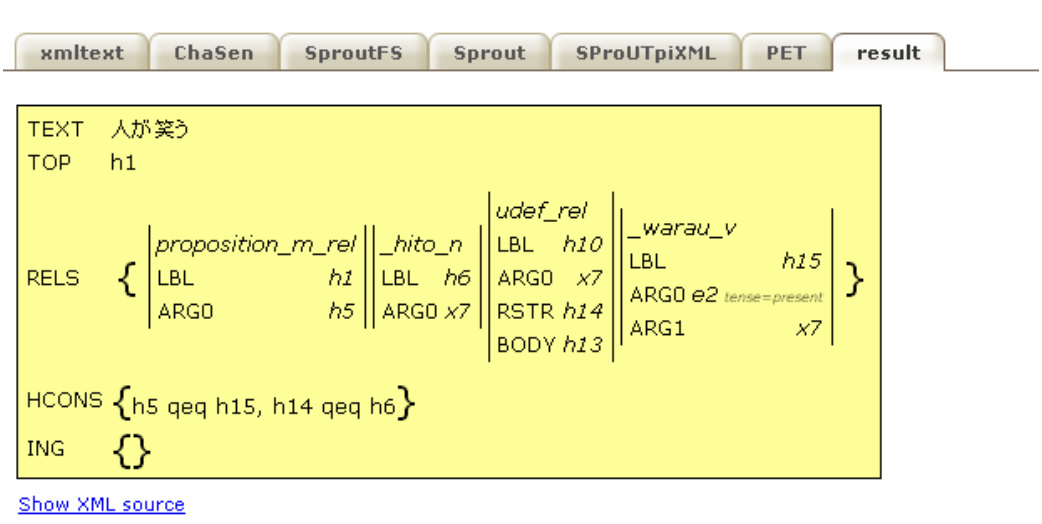


Figure 121: Japanese

Basic to the hybrid multilingual processing approach in DeepThought was the idea of Robust Minimal Recursion Semantics (RMRS). This semantic output structure combines the ideas of compatibility of multiple languages' grammar output and compatibility of Natural Language Processing Modules of different preciseness in analysis. In the multilingual examples above,

we have seen (an HTML presentation of) RMRS output of multiple languages. Although the actual semantic relations are different for each language, there are a lot of common relation names for propositions and determiner relations, as well as arguments, labels and scoping restrictions.

For multiple NLP modules, the same should be valid. For Japanese, the Heart-of-Gold contains a named-entity recognition tool. Consider the sentence in Example 249:

Example 249

花子　　が　成田　　に　行った
Hanako　ga　Narita　ni　itta
Hanako　NOM　Narita　to　went

(Hanako went to Narita.)

The Named-Entity Recognition Tool Sprout (Drizdzynski et al. 2004) delivers output for the recognized place name *Narita*:

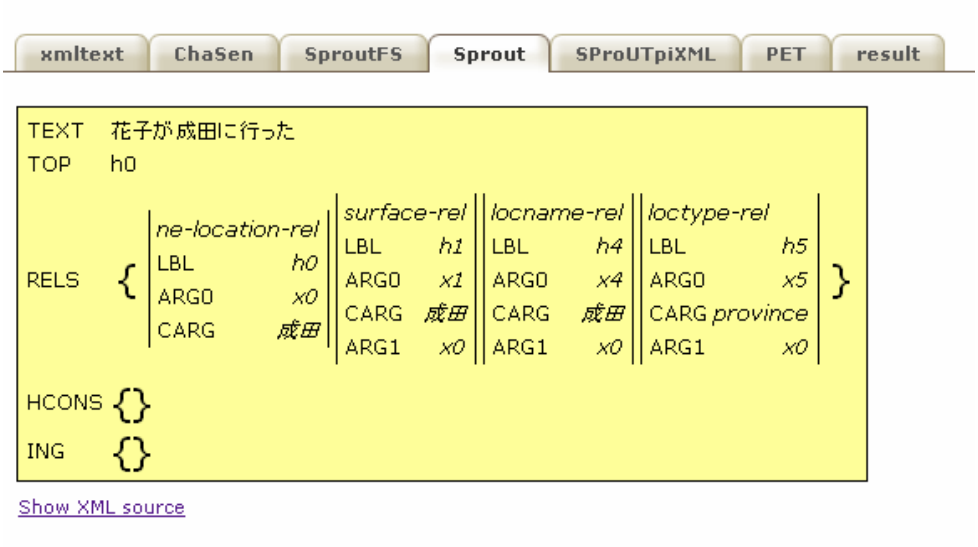


Figure 122: RMRS output of Sprout for *Narita*

The information that Narita is a place name is (via XML) passed to the HPSG processing component and inserted to the grammar processing. The RMRS output of the HPSG processing of Jacy therefore contains a generic location name, as can be seen in Figure 123.

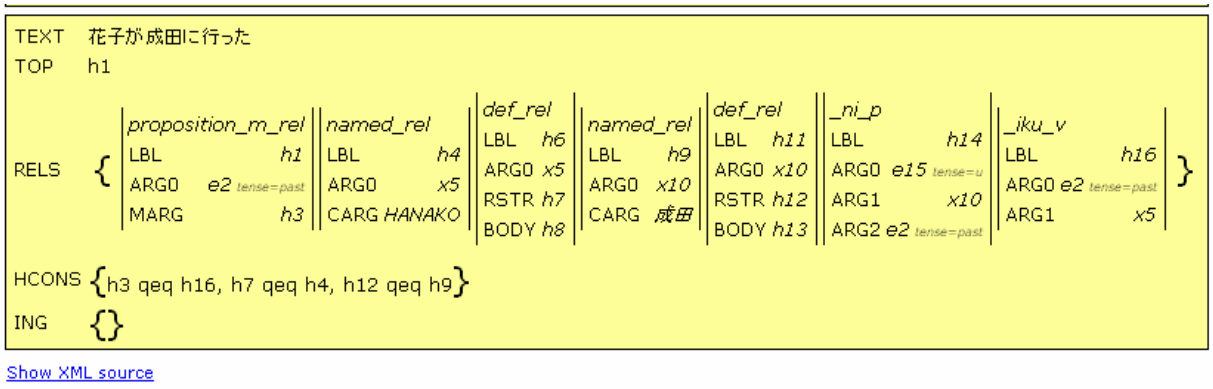


Figure 123: RMRS output of JACY processing of *Hanako ga Narita ni itta*.

This extension that had a high influence on lexical coverage and robustness, as it is now possible to recognize named entities with a separate tool and use the output information in grammar parsing.

10.4 Dictionary definition sentences

In collaboration with the NTT Machine Translation Research Group, the focus of work on JACY was dictionary definition sentences. The NTT group built treebanks based on the grammar and used these for stochastic disambiguation methods and for ontology extraction (Bond et al. 2004, Bond et al. 2005). Therefore, the strategy concerning ambiguity was rather not to under-specify, but rather give all possible readings. Ranking is then done using stochastic models trained on the manually evaluated and selected trees, following the methods proposed by Oepen et al (2002).

Dictionary definitions contain relatively short and well-structured sentences covering a broad vocabulary and domain. Thus, the group brought in a huge amount of lexical items to be added to the grammar. The lexical type system therefore has to be clear and well understandable; and was actually re-organized in order to support this (see Hashimoto et al. 2005). Documentation efforts have started with setting up a web site with the open-source grammar download and an increasing amount of documentation material (http://www.delph-in.net/jacy).

Furthermore, definitions contain lots of nominalizations, such that a precise analysis of these is necessary. Our approach to this is described in Section 5.4.

11 Evaluations

A number of evaluation methods were proposed for Natural Language Parsing systems. We reference some of them here: Parseval (Black et at. 1991) proposed a quantitative system comparison, measuring system performance based on corpora. The evaluation was based on phrase structure tree match (tree topology). The evaluation measure was labeled recall/precision for phrase-structure annotations. Collins (1999) gives an alternative evaluation to Parseval, measuring word-word dependencies. Carroll et al., (1999) proposed to evaluate dependency relations, based on a manually annotated subset of the Brown corpus. Briscoe et al. (2002) proposed a scheme for evaluating parse selection accuracy based on named grammatical relations between lemmatised lexical heads.

Big evaluation events were set up for NLP tasks and applications: The MUC evaluation tasks focused on Information Extraction (MUC-7, for example, Chinchor 1997), TREC (TREC 13 in Voorhees and Buckland 2004) evaluated Information Retrieval, Question-Answering Systems had standard evaluations in CLEF (Peters et al. 2004), BLEU score (Papineni 2002) and NIST score (NIST 2002) were used to evaluate Machine Translation Systems.

A number of annotated corpora for parse evaluation have been created. To mention here are the treebanks, such as the Penn Treebank of English (Marcus et al., 1993). This treebank, annotated Wall Street Journal material, has become a kind of benchmark for the evaluation and comparison of parsers and English grammars. Further, there is also the Prague Dependency Treebank for Czech (Hajic, 1998), the ATIS Corpus (http://www.ldc.upenn.edu/Catalog/readme_files/atis/sspcrd/corpus.html) and SUSANNE (http://www-2.cs.cmu.edu/afs/cs/project/ai-repository/ai/areas/nlp/corpora/susanne/0.html) for English. The EDR annotated corpus of Newspaper text (EDR 1996) has become a standard for Japanese annotation and evaluation. We plan to evaluate our grammar based on this in the near future.

The research project TSNLP (Lehmann et al. 1996) developed general guidelines for test suite construction. Further, they developed test suites for deep linguistic processing, which are highly structured and annotated. The grammar profiling tool [incr tsdb()] (Oepen and Carroll 2000) which we use for evaluation of our grammar origins from this project. A central component of the project PERFORM (Fouvry 2004) was a common multilingual test data base, i.e. complex annotated bench marks.

The task of evaluating the semantic output of a deep grammar for hybrid multilingual NLP is different from evaluating trees or part-of-speech tagging, as the output structure is farer away from text string and results do not directly correspond to surface strings. For example, semantic annotation can contain a proposition or a linking relation between compound nouns. Further, it might contain undefined quantification in the case, where there are no determiners in the surface string (which is often the case in analyzing Japanese).

The first decision to be made in evaluation here is the decision between intrinsic and extrinsic evaluation, i.e., developer-oriented and operational evaluation. Intrinsic evaluation is part of grammar development at all stages. Further, we need extrinsic evaluation, in order to see whether the analyses are useful for NLP tasks (and for what kinds of tasks). The “usefulness” of the grammar must have a different focus, due to the task it has to perform. For example, if a grammar is used in grammar checking, the focus of evaluation should be on the side of argument structure, as this is the most interesting information here. The information retrieval task would need a grammar which is good at getting the correct negation scope. Generation, for text summarization or machine translation, needs highly precise and formally correct

MRSs. Dialogue systems would need a grammar that can be parsed efficiently, as the application is time critical.

A second decision to be made in evaluation is internal or external evaluation. We do not have a comparable grammar of Japanese and therefore evaluate internally. Quantitative measures, such as coverage and size, shall be given as well as qualitative measures about the behavior of the grammar.

The data the evaluation is based on can be constructed data to show the constructions the grammar can cover, as well as corpus-based data to show the usefulness in real-world applications. We will refer to both kinds of data, though leaving out annotated data due to unavailability.

Evaluation can be performed manually or automatically. Manual evaluation has the advantage to be precise in results, but is very time consuming and cannot be performed on large amounts of data. Automatic evaluation can give information about facts like coverage on large amounts of data, but does not give precise information about output validation. We report results on both kinds of evaluation; manual evaluation by treebanking parsed sentences and inspection of MRS output, as well as automatic evaluation by coverage on constructed and natural data in different domains.

Evaluation of NLP is a complex task, many difficult matters have to be considered and there is no one "magic number" (see Sparck-Jones 1994). Thus, in course of the evaluation, we need to answer the following questions:

1. What is the grammar size? How many rules and lexical entries does it contain?
2. What is the general coverage on what kinds of data?
3. How far is the grammar flexible and useful for applications? Is it domain-adaptable and can be used for different application domains? What does it take to go to new domains?
4. How far can the grammar be used in multilingual applications?
5. Is the output precise and does it correspond to semantic format and content restrictions?

Table 20 shows the evaluation categories the questions belong to.

Table 20: Evaluation typology of leading questions

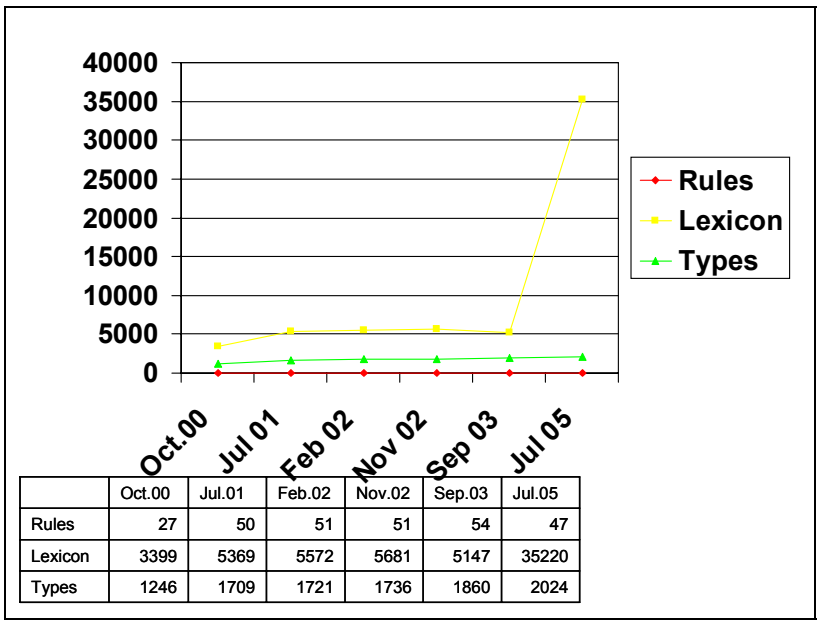
Question	intrinsic - extrinsic	internal - external	quantitative - qualitative	constructed - corpus-based data	manual - automatic evaluation
1. What is the grammar size? How many rules and lexical entries does it contain?	intrinsic	internal	quantitative	—	automatic
2. What is the general coverage on what kinds of data?	extrinsic	mostly internal	quantitative	constructed and corpus-based data	automatic
3. How far is the grammar flexible and useful for applications? Is it domain-adaptable and can be used for different application domains? What does it take to go to new	extrinsic	mostly internal	qualitative	constructed and corpus-based data	manual

domains?					
4. How far can the grammar be used in multilingual applications?	extrinsic	mostly internal	qualitative	constructed and corpus-based data	manual
5. Is the output precise and does it correspond to semantic format and content restrictions?	extrinsic	internal	qualitative	constructed data	manual

We will try to answer these questions by considering the different applications and domains the grammar was part of. All tests were performed using the [incr tsdb()] tool for grammar testing and profiling (Oepen and Carroll 2000).

11.1 What is the grammar size? How many rules and lexical entries does it contain?

The Japanese HPSG grammar in Verbmobil in October 2000 (Siegel 2000) consisted of 27 rule schemata, 1,246 types and a lexicon of 3,399 entries. In the end of the follow-up project with the Californian company YY in September 2003 (Siegel and Bender 2002), there were 5,147 words in the lexicon, 54 rule schemata and 1860 types. In July 2005, there were 47 rules, 35,220 lexicon entries and 2024 types. This development can be seen in Figure 124, Figure 125 and Figure 126. It shows that the number of rules did not increase, but even decreased at bit. This is a sign that they could be grouped and organized in a better way. The number of lexicon entries increased quite a lot, especially over the last two years, when the application domain got more and more open. The types increased quite a lot. In a typed lexicalized formalism, this is exactly the behavior that can be expected.



	Oct.00	Jul.01	Feb.02	Nov.02	Sep.03	Jul.05
Rules	27	50	51	51	54	47
Lexicon	3399	5369	5572	5681	5147	35220
Types	1246	1709	1721	1736	1860	2024

Figure 124: Development of grammar size over a period of five years

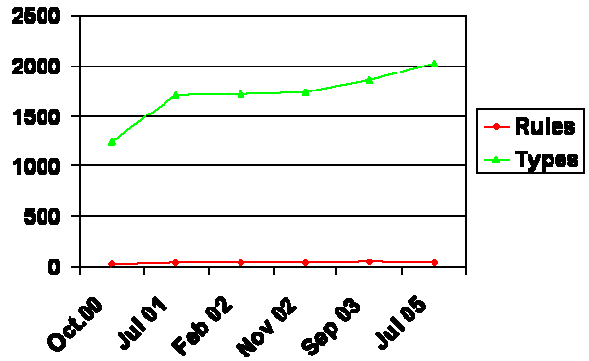


Figure 125 : Development of rules and types

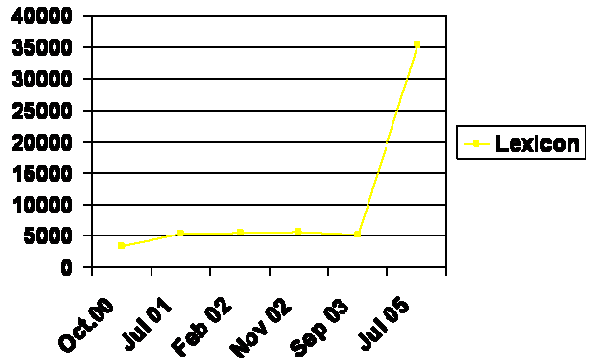


Figure 126: Development of the lexicon

11.2 What is the general coverage on what kinds of data?

Data evaluation can concern constructed data and corpus data. The former gives information about the range of phenomena the grammar covers, while the latter gives information about the coverage on naturally occurring data. In the case of constructed data, it is possible to evaluate the quality of output.

We executed a parsing test on 2607 utterances out of 100 dialogs in the Verbmobil scenario of appointment scheduling (corpus data). The average sentence length was 8.99 words. Parsing was interrupted, when the chart contained more than 20 000 items. We processed only the first reading of every utterance. The result can be seen in Table 21. 2 044 utterances (78.4%) got a parsing result, i.e. a spanning analysis. On average, there were 100.60 parses processed for every utterance.

Table 21: Test of 2607 utterances in Verbmobil scenario I (using [incr tsdb()])

total *Phenomenon*	*positive* *items* #	*lexical* *strings* ∅	*parser* *items* ∅	*total* *analyses* ∅	*overall* *results* #	*coverage* *%*
Total	*2607*	*8.99*	*100.60*	*1.00*	*2044*	*78.4*

Next, a parsing test on the utterances of 50 dialogs in the Verbmobil scenario of travel planning was executed. The test contained 7 969 utterances of spontaneous language. This test processed exhaustive parsing. The average length of the utterances was 6.22 words. There were on average 76.14 parses per utterance. The results can be seen in Table 22. 5 807 (72.9%) got a parsing result, *i.e. a spanning analysis*. Overgeneration was tested with 223 ungrammatical sentences. Table 23 shows that 7.2% of these got a parsing result.

Table 24 shows the performance of the parser in these tests. They were executed on a 44 MHz Ultra Sparc 2 with Solaris 2.6.

The Verbmobil parser already had the possibility to allow partial analyses. Result of this parse was the longest matching fragments that the grammar could parse. A Verbmobil system test with 491 sentences (all words known, string input, allowing partial analyses) showed no cases, where the Japanese syntax delivered nothing. This shows that in almost any case the syntax module delivers at least partial analyses.

Table 22: Test of 7969 utterances in Verbmobil scenario II

total Phenomenon	*positive items #*	*lexical strings Ø*	*parser items Ø*	*total analyses #*	*overall results #*	*coverage %*
Total	***7969***	***6.22***	***76.14***	***25.65***	***5807***	***72.9***

Table 23: Test for overgeneration

negative Phenomenon	*word items %*	*lexical strings Ø*	*parser items #*	*total analyses #*	*overall results #*	*coverage %*
Total	***223***	***7.40***	***80.64***	***11.50***	***16***	***7.2***

Table 24: Performance

	Test 1 (nonexhaustive)	*Test 2 (exhaustive)*	*Test 3 (ungrammatical)*
average cpu time	*1.74s*	*2.13s*	*1.26s*

Evaluation of the coverage of the grammar in the banking domain delivered the following results:

The grammar now covers 93.4% of constructed examples from the banking domain (747 sentences) and 78.2% of realistic email correspondence data (316 sentences), concerning requests for documents. During three months of work, the coverage in the banking domain increased 48.49% overall. The coverage of the document request data increased 51.43% in the following two weeks.

Table 25: Coverage on the development data in the banking domain, generated by [incr tsdb()]

Phenomenon	*total items #*	*positive items #*	*word string %*	*lexical items Ø*	*parser analyses Ø*	*total results #*	*overall coverage %*
Total	*747*	*747*	*101*	*75.24*	*6.54*	*698*	*93.4*

Table 26: Coverage on test data in the domain of document request, generated by [incr tsdb()]

Phenomenon	*total items #*	*positive items #*	*lexical items Ø*	*parser analyses Ø*	*total results #*	*overall coverage %*
Total	*316*	*316*	*83.90*	*39.91*	*247*	*78.2*

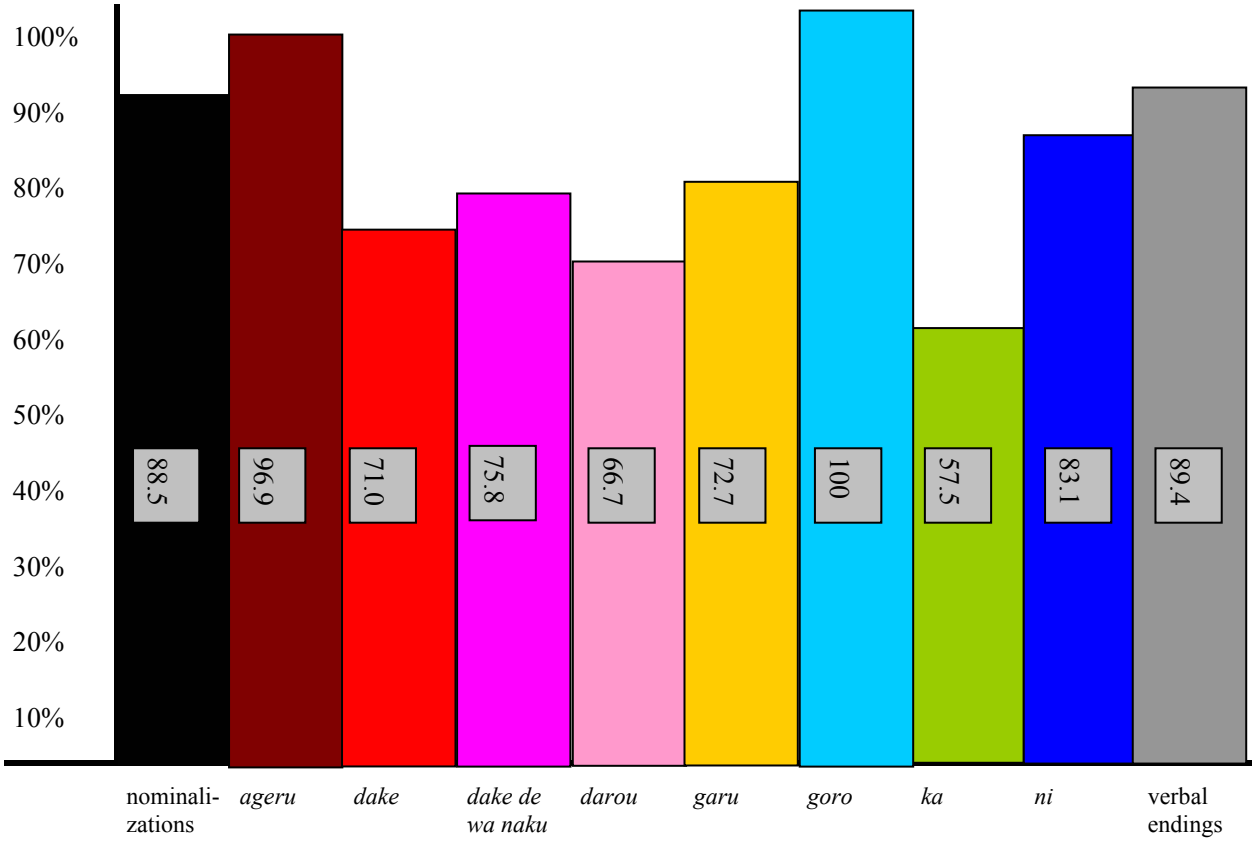
We applied the grammar to completely unseen data in the banking domains, namely the FAQ web site of a Japanese bank. The coverage was 61%. Manual evaluation of the results showed that 91.2% of the parses outputs were associated with all well-formed MRSs. That means that we could get correct MRSs in 55.61% of all sentences (spanning analysis).

The latest test was done on completely unseen data in a new domain: We parsed 1.000 sentences of the Kyoto University Corpus of Mainichi Shinbun newspaper data (Kurohashi and Nagao 1998). Initial coverage was 32.8% on this data, with an ambiguity of 219.85 analyses/sentence. 30% of the parses failed because of edge limit exhausting problems with the parser: They were too complicated to be parsed in a set amount of time and with a

reasonable number of readings. The aim to give more and more precise results is followed by an ambiguity problem. Here, the formalism is required to develop more strategies for packing and underspecification. Only 0.37% of the parses failed because of missing lexicon entries, which shows that the lexicon is in a quite stable state now.

A method to evaluate the grammar coverage on grammatical phenomena we are interested in is to build up test sets for these phenomena using [incr tsdb()], parse and evaluate the results manually (evaluation of constructed data). We therefore set up test sets containing of example sentences of a grammar book (Makino and Tsutsui 1986), ordered by the grammar book chapters. We performed a coverage test on nominalizations (grammar book examples in the chapters of *koto*, *mono* and *tame)*, the chapters of *ageru, dake, dake de wa naku, darou, garu, goro, ka* and *ni* and a manually constructed testsuite of verbal endings. Figure 127 shows the results of coverage evaluation on these selected phenomena.

Figure 127: Coverage on selected phenomena



11.3 How far is the grammar flexible and useful for applications? Is it domain-adaptable and can be used for different application domains?

The grammar is aimed at working with real-world data, rather than at experimenting with linguistic examples. Therefore, robustness and performance issues play an important role. While grammar development is carried out in the LKB (Copestake 2002), processing (both in the application domain and for the purposes of running test suites) is done with the highly efficient PET parser (Callmeier 2000). Table 27 and Table 28 show the performance of PET parsing of manually constructed and real occurring data in the banking domain, respectively. It shows the number of items in the test suite in “items”, the average number of tasks that have been performed in “etasks”, the filter efficiency of the parser in “filter”, the number of edges in the chart in “edges”, the average time (seconds) needed for getting the first parse in

"first", the average number of time (seconds) for getting all parses in "total", the average cpu time in "tcpu" and the average work space needed in "space".

Table 27: Performance parsing banking data, generated by [incr tsdb()]

Phenomenon	items #	etasks Ø	filter %	edges Ø	first Ø (s)	total Ø (s)	tcpu Ø (s)	space Ø (kb)
Total	742	946	95.7	303	0.06	0.11	0.11	833

Table 28: Performance parsing document request data, generated by [incr tsdb()]

Phenomenon	items #	etasks Ø	filter %	edges Ø	first Ø (s)	total Ø (s)	tcpu Ø (s)	space Ø (kb)
Total	316	2020	96.5	616	0.23	0.26	0.26	1819

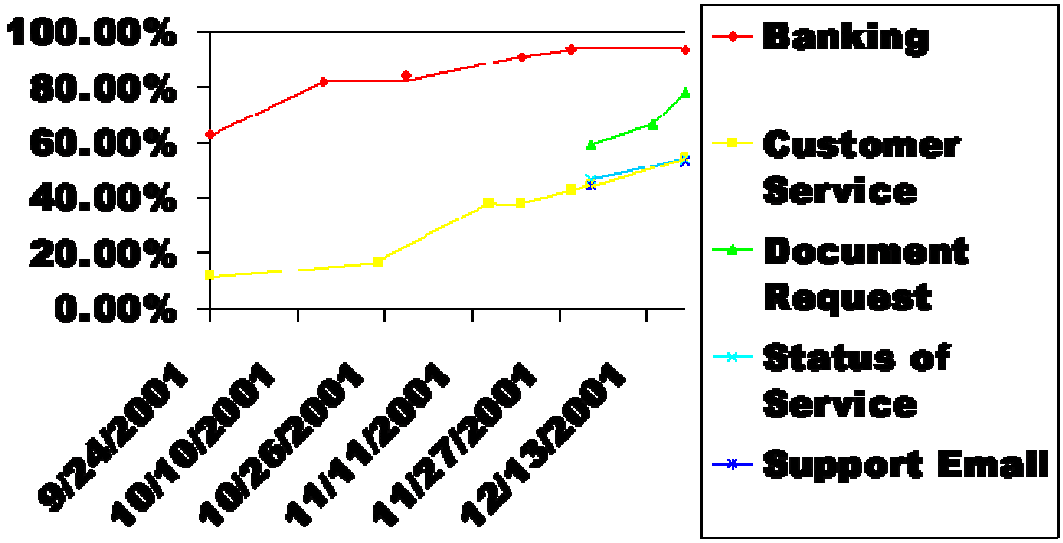
The task of adapting the grammar to a new domain first occurred when going from Verbmobil (spoken dialogues) to YY (emails in the banking domain). Continuous evaluation of coverage and performance gives a picture of the development. We had four different test suites in the new domain, made from data given by customers:

- Banking data (864 items)
- Document request (317 items)
- Customer service (952 items)
- Support email (1982 items)
- Status of service (204 items)

A first test on a smaller testset in the domain in August 2001 showed that lexicon extension and linking to named-entity recognition is a main challenge in extending the coverage of the grammar in a new domain:

- 457 items tested
- 213 parsed
- 202 lexicon errors:
  - 150 numbers, names, placeholders (X), email-addresses
  - 52 words (20 unique)
- 42 failed for reasons of grammatical coverage:
  - you, causative, kata, np fragments, interrogative complement sentences

Development of coverage on this test data over a period of three months can be seen in Figure 128. It shows that it is possible to switch to a completely new domain with work being done over a period of three months.



Banking:	Customer Service:	Document Request:	Status of Service	Support Email:
Sept.2001: 62.9%	Sept. 2001: 11.4%	2001-12-07: 59.5%	2001-12-03: 46.6%	6.12.2001: 44.3%
Dec.2001: 93.4%	Dec. 2001: 54.7%	2001-12-20: 78.2%	2001-12-20: 53.9%	20.12.2001: 53.7%

Figure 128: Development of Coverage over a period of 3 months.

Another experiment to adapt the grammar to a new domain was done by the Japanese company NTT Communication Science Laboratories (Nippon Telegraph and Telephone Corporation). In Bond et al. (2004) they report about the idea to extend the coverage on the Lexeed defining sentences to over 80% in four weeks. Six people were involved in this task, three of them had little experience in HPSG, none of them was previously involved in JACY. First tests on the data gave a coverage of 39.3%. Adding some orthographemic variants to the lexicon extended the coverage to 46.2%.

The coverage increased to 82% in four weeks of work. First of all, the lexicon was expanded (up to 32,000 lexicon entries), which gave an increase of up to 55%. Further, some rules were added to account for compound verbs, increasing the coverage up to 70%. Domain-specific adaptation concerned a structure for dictionary definitions, such as “driver: In Golf, a long distance club”. This experiment has shown that the grammar can be easily adapted to another domain, such that most of the work is in the area of lexicon extension. The task was done by researchers that were not involved in the original grammar development process and had minimal knowledge of HPSG.

An experiment on another data set showed a similar behavior (see Figure 129): On a data set of 3742 test items that were originally set up for evaluation of Machine Translation by NTT[67], we had an initial coverage of 66.1%. After 3 months of mostly lexicon work, we achieved 83.0%. However, it has to be mentioned that the ambiguity went up as well: from 18.62 analyses on average to 68.11 analyses on average. It is thus very important to work on parse selection when extending the grammar coverage.

[67] A description of the test data can be found in Ikehara et al. (1994).

Phenomenon	(g)old				new			
	lexical Ø	analyses Ø	in Ø	out Ø	lexical Ø	analyses Ø	in Ø	out Ø
Total	123.04	18.62	66.1	100.0	133.43	68.11	83.0	100.0

(generated by [incr tsdb()] at 21−dec−05 (15:21))

Figure 129: Comparing the coverage of MT data September 2005 and December 2005

11.4 How far can the grammar be used in multilingual applications?

The DeepThought project has set a strong focus on multilinguality in grammar development and application. As we have shown in section 10.3: "Parallel multilingual grammar development embedded in hybrid language processing", we included the Grammar Matrix types, in order to make the semantic output compatible and reliable. We further set on RMRS as a compatible semantic output format for languages and NLP modules. Chapter 10: "JACY in Different Application Domains" shows the compatibility of analysis output of the six language's grammars.

Bond et al. (2004) report on using the grammar for automatic ontology extraction by parsing dictionary definition sentences. The fact that it uses MRS as output is explicitly mentioned as a strength for the knowledge acquisition task and multilingual application. Nichols et al. (2005) show how JACY is more useful for the task of ontology acquisitition than robust semantic representations derived from ChaSen.

11.5 Is the output precise and does it correspond to semantic format and content restrictions?

Some NLP applications require robustness and preciseness rather than 100% coverage on spanning analyses. The requirement here is to get as much **correct** information out of the analyzed data as possible. The grammar processing therefore has the option to give partial analyses in the case of the non-availability of spanning analyses. Coverage is not really a topic here, but preciseness of the output representation. This is much more difficult to evaluate and needs some manual work. Manual inspection of the MRS output of the banking domain showed that 91.2% of the parses output were associated with syntactically all well-formed MRS's.

Another possibility to check the output preciseness is to use the treebanking mechanism for HPSG grammars described by Oepen et al. (2002b). We parse test sentences with the [incr tsdb()] tool and choose the best tree on the basis of lexical types, rules and semantic representations. This method combines the evaluation of tree typology with evaluation of semantic output and is therefore an excellent method to evaluate precise and deep grammars. Selection of the correct parse is done by making choices on the application of rules and on the lexical entries, as can be seen in Figure 130. The evaluator further gets presented the trees that are linked with the choices. When the selection process resolves to one tree, the evaluator gets the MRS output for evaluation, as can be seen in Figure 131. It is noted, if the output is fragmentary or cyclic. S/he can further display scoped output, RMRS or parse chart to support her/his selection.

Bond et al. (2004) report that they have treebanked 23,000 sentences; 95% of these could be resolved to one correct parse, while 1% of the analyses were completely rejected.[68]

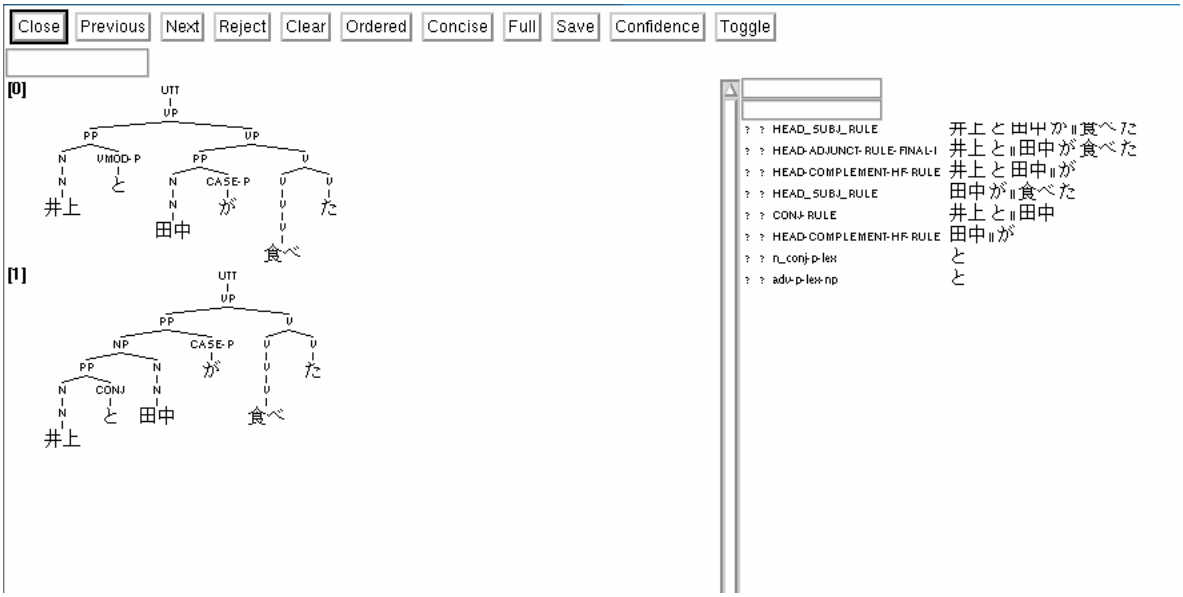


Figure 130: Selection of rules and lexical entries using the [incr tsdb()] treebanking facility

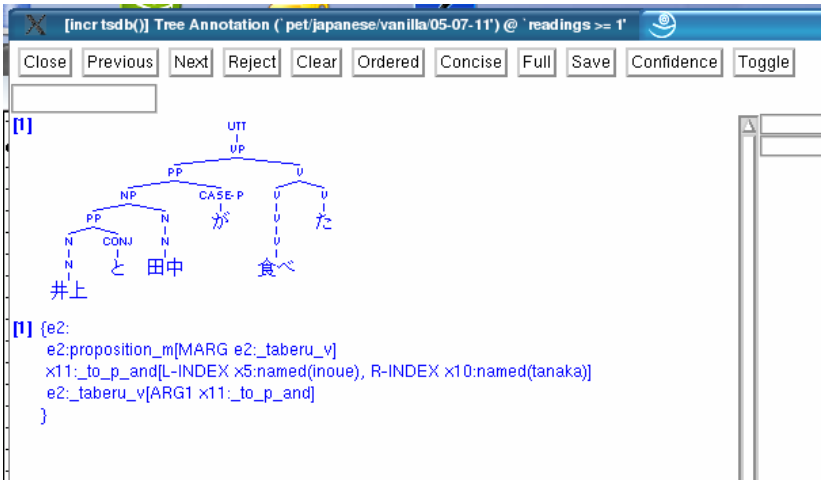


Figure 131: MRS information when down to one tree

[68] The treebanking task was done by Japanese native speakers.

12 Conclusion

We described a broad coverage Japanese grammar, based on HPSG syntactic and MRS semantic theory. It encodes precise morphologic, syntactic, semantic, and pragmatic information in feature structures. The grammar system is connected to a morphological analysis system and uses default entries for words unknown to the HPSG lexicon. The grammar has a long history of being embedded in research projects and is therefore profiled for annotating Japanese language data in applications with real-world data of different domains and language dialects. It covers the basic phenomena of the Japanese language, as well as quite a lot of peripheral ones, and gives morphologic, syntactic, semantic and pragmatic analyses. The grammar is being developed in a multilingual context, where much value is placed on parallel and consistent semantic representations. The development of this grammar thus constitutes an important test of the cross-linguistic validity of the MRS formalism. The consideration of RMRS in grammar development and application draws the line to the usage of the grammar in a hybrid architecture, where shallow analyses (such as named-entity recognition) and deep analysis (such as our grammar) are combined and multilingual resources are compared.

The basic phrase structures of JACY are built on the general and multilingual Matrix phrase structure types. Rules for phenomena on the borderline between morphology and syntax are added, as in Japanese this boarder is not as clear as in other languages.

Subcategorization in Japanese needs careful inspection, because we can find a combination of scrambling and optionality of verbal arguments. We implemented an approach that is Matrix-based (actually, the Matrix approach was based on insights of the Japanese and the English grammars) and therefore general enough to be useful for different languages and special enough for the Japanese phenomena.

Verbal constructions are central for semantic construction, argument and event structure, and in Japanese the central point for inflection. Our approach is to organize lexical types in a type hierarchy and thus highly modular. We have shown how inflectional and derivational morphology can be expressed in this approach of using a type hierarchy and rules operating on it. Argument changing operations are triggered by passive, causative and some auxiliary constructions. The treatment of these proves the subcategorization approach valid.

The description of nominal constructions shows the various relations between syntax, semantics and pragmatics. The HPSG feature structure is very useful for expressing these relations. The connection between named-entity recognition and part-of-speech tagging and JACY grammar could be shown here as well, using a combination of hand-coded lexical information with default lexicon entries. We have described an approach to the syntax of Japanese numeral classifiers which allows us to build semantic representations for strings that contain these prevalent elements — representations suitable for applications requiring natural language understanding, such as (semantic) machine translation and automated email response.

Particles play a central role in Japanese syntax and needed a close investigation and a structured type hierarchy. The syntactic behaviour of Japanese particles has been analyzed using the Verbmobil dialogue data. We observed 25 different particles in 800 dialogues on appointment scheduling. It has been possible to set up a type hierarchy of Japanese particles. We have therefore adopted a lexical treatment instead of a syntactic treatment based on phrase structure. This is based on the different kinds of modification and subcategorization that occur with the particles. We analyzed the Japanese particles concerning to their possibilities of co-occurrence, their behaviour of modification and their occurrence in verbal arguments. We clarified the question which common characteristics and differences between the individual

particles exist. A classification in categories was carried out. After that a model hierarchy could be set up for the HPSG grammar. The simple distinction into case particles and postpositions, as often proposed in theory-oriented research literature, was proved to be not sufficient. The assignment of the grammatical function is done by the verbal valence and not directly by the case particles. The topic particle is ambiguous. Its binding is done by ambiguity and underspecification in the lexicon. The evaluation of the particle treatment showed that 91.6% of the occurred particles in 100 test sentences could be correctly analyzed. All combinations of particles and 82.35% of the missing particles were analyzed correctly.

We believe that the rather peripheral exceptions noted in the chapter about head-initial constructions do not detract from the broad generalization that Japanese has a very strong tendency to be head-final. Rather, they illustrate once again the fact that languages seamlessly combine general tendencies with particular exceptions (cf. Fillmore et al., 1988). In order to build consistent grammars that scale up to ever larger fragments of the languages we wish to model (such as is required for practical applications), we require a framework that allows the statement of generalizations at varying degrees of granularity. Furthermore, we believe that the construction of broad-coverage precision grammars such as JACY in the context of applications which require robustness in the face of real-world language use provides a useful discovery procedure for many of the smaller generalizations and exceptional cases (cf. Baldwin et al., 2004).

The Japanese language has a complicated system to express the social relation between speaker, addressee and subject of an utterance. This relation is expressed by honorification. It concerns verbal forms, verbal conjugations, nominal prefixes and pronouns and undergoes syntactic, semantic, pragmatic and domain-specific restrictions. We have shown that for Japanese it is necessary to distinguish subject honorification, entity honorification and addressee honorification and to introduce polarity for these. The number and kind of the dimensions is language-specific; German and French, for example, have only one dimension, while Korean and Japanese have three. In one sentence, different dimensions of honorification can be expressed. We have given a treatment of honorifics in the HPSG framework that covers all three dimensions of Japanese honorifics and makes it possible to account for honorific agreement as well as restrictions in complement sentences and restrictions for zero pronouns. The approach allows a uniform treatment of honorific dimensions in different languages.

During the long history of the development of this grammar, it has been used in various applications, requiring the work on different types of data. This has put a high demand of robustness and preciseness. The grammar needs to be multilingual, to be able to work with large amounts of data, to be flexible enough to adapt to a new domain and have large and extensible lexica. The embedding in a hybrid approach brought the breakthrough in coverage. The cooperation with external partners in lexicon development brought the breakthrough in the organisation and documentation of lexical types.

The HPSG framework has proven to be very well suited for the task of describing Japanese. Especially the type hierarchy approach makes the grammar modular. The lexicon is easy to extend, once the lexical types are well organized and documented. The idea of the sign as a mean to represent different levels of linguistic information (from morphology to pragmatics) has proven to be right for expressing the complex interactions, as especially in Japanese, pragmatic information relates much with information on the other levels.

The MRS framework has proven to be expressive enough for the Japanese semantics, flexible enough for a wide-coverage grammar used in applications and well defined to generate compatible output in multilingual grammars. It integrated well into the HPSG framework and into the sign- and type-based approach.

The open-source community working with HPSG provides a set of extremely useful tools to make the grammar developers life easy: LKB for grammar development, PET for efficient processing, Heart-of-Gold for hybrid processing and [incr tsdb()] for grammar profiling and treebanking.

The evaluation shows that the grammar is at a stage where domain adaptation is possible in a reasonable amount of time. We have given answers to the five leading questions of grammar evaluation that apply to our approach of multi-level annotation. The JACY grammar is of a reasonable size, has a good coverage on different types of data, is flexible and useful for applications, adaptable to new domains, useful in multilingual applications and gives precise and correct semantic output.

Thus, it is a powerful resource for linguistic applications for Japanese.

In future work, this grammar should be further adapted to other domains, such as the EDR newspaper corpus (including a headline grammar). As each new domain is approached, we anticipate that the adaptation will become easier as resources from earlier domains are reused. The main focus already lies on the lexical level. An important part of the work will be further improvement and support of lexical type documentation efforts, as already started and described by Hashimoto et al. (2005).

Being part of the Heart-of-Gold hybrid processing architecture, we foresee research in the combination of information provided by statistical and chunk parsing and HPSG grammar processing of Japanese. Generation will be a certain topic in the near future, to open up the perspective for new applications and to give further means to evaluate the grammar and increase preciseness.

References

Ahn, Sang-Cheol and Jong-Bok Kim (2000): An Optimal Approach to Korean Nominalization. In Susumo Kuno et al. (eds), Harvard Studies in Linguistics VIII, 211-222. Dept. of Linguistics, Harvard University.

Asahara, Masayuki and Yuji Matsumoto (2000): Extended Models and Tools for High-performance Part-of-speech Tagger. In Proceedings of the 18th International Conference on Computational Linguistics, Coling 2000, 21-27. Saarbrücken, Germany.

Baldwin, Timothy (2004): Making Sense of Japanese Relative Clause Constructions, In *Proceedings of the 2nd Workshop on Text Meaning and Interpretation*, Barcelona, Spain, pp. 49–56.

Baldwin, Timothy, Beavers, John, Bender, Emily M., Flickinger, Dan, Kim, Ara and Oepen, Stephan (2004): Beauty and the Beast: What Running a Broad-Coverage Precision Grammar Over the BNC Taught Us about the Grammar and the Corpus, Paper presented at the International Conference on Linguistic Evidence: Empirical, Theoretical, and Computational Perspectives, Tübingen, Germany.

Bender, Emily M. and Kathol, Andreas. In press. Constructional Effects of Just Because . . . Doesn't Mean . . . . In BLS27.

Bender, Emily M. and Siegel, Melanie (2004): Implementing the Syntax of Japanese Numeral Classifiers. In: Proceedings of IJCNLP-04.

Bender, Emily M., Flickinger, Dan and Oepen, Stephan (2002): The Grammar Matrix: An open source starter-kit for the rapid development of cross-linguistically consistent broad coverage precision grammars. In: Proceedings of the Workshop on Grammar Engineering and Evaluation, Coling 2002, Taipei, 8–14.

Black, E., S. Abney, D. Flickinger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini, and T. Strzalkowski (1991): A procedure for quantitatively comparing the syntactic coverage of English grammars. In *Proceedings of the DARPA Speech and Natural Language Workshop 1991*, Pacific Grove, CA.

Bond, F., Fujita, S., Hashimoto, C., Kasahara, K., Nariyama, S., Nichols, E., Ohtani, A., Tanaka, T., Amano, S. (2004): The Hinoki Treebank: A treebank for text understanding. In: Proceedings of the IJC-NLP-2004.

Bond, F., K. Ogura, and S. Ikehara (1994): Countability and Number in Japanese-to-English Machine Translation. In *15th International Conference on Computational Linguistics: COLING-94* , 32-38.

Bond, Francis (2005): Translating the Untranslatable: A Solution to the Problem of Generating English Determiners. CSLI Studies in Computational Linguistics, CSLI Publications, Stanford.

Bond, Francis and Kentaro Ogura (1998): Reference in Japanese-to-English machine translation. *Machine Translation*. **13**(2-3): 107-134.

Bond, Francis and Kyong-Hee Paik (2000): Reusing an Ontology to Generate Numeral Classifers. In: *Proceedings of Coling 2000*, Saarbrücken, Germany.

Bond, Francis, Ann Copestake, Dan Flickinger, Stephan Oepen and Melanie Siegel (2005): Open-Source Machine Translation with DELPH-IN. Proceedings of the Open-Source Machine Translation Workshop at MT-SUMMIT X, September 2005.

Borsley, Robert D. (1993): Heads in HPSG. In Greville Corbett and N. Fraser and S. McGlashan, editor(s), *Heads in Grammatical Theory*.

Borsley, Robert D. (1993): Heads in HPSG. In Greville Corbett, N. Fraser and S. Mc-Glashan (eds.), *Heads in Grammatical Theory,* pages 186ñ203, Cambridge: Cambridge University Press.

Briscoe, Ted, John Carroll, Jonathan Graham, Ann Copestake (2002): Relational evaluation schemes. In *LREC Workshop on Parser Evaluation*, Las Palmas, Spain.

Callmeier, Ulrich (2000): PET — a platform for experimentation with efficient HPSG processing techniques. Journal of Natural Language Engineering, Special Issue on Efficient Processing with HPSG: Methods, Systems, Evaluation, 99-108.

Callmeier, Ulrich, Eisele, Andreas, Schäfer, Ulrich and Melanie Siegel (2004): 'The DeepThought Core Architecture Framework. '. In: *Proceedings of LREC*.

Carroll, J., G. Minnen and E. Briscoe (1999): Corpus annotation for parser evaluation, *Proceedings of the EACL-99 Post-Conference Workshop on Linguistically Interpreted Corpora (LINC'99),* Bergen, Norway, pp. 35–41.

Chinchor, N. (1997): MUC-7 evaluations metrics, Proceedings of the Third Message Understanding Conference (MUC-3), Morgan Kaufmann, San Mateo, CA, pp. 17-24.

Chung, Chan and Jong-Bok Kim (2002): Differences between External and Internally Headed Relative Clauses In Jongbok Kim and Stephen Wechsler, eds., *Proceedings of the Ninth International Conference on Head-Driven Phrase Structure Grammar, 43--65.* Stanford: CSLI Publications.

Collins, M. (1999): Head-driven Statistical Models for Natural Language Parsing, PhD Dissertation, University of Pennsylvania.

Copestake, Ann (2002): Implementing Typed Feature-Structure Grammars. Stanford: CSLI.

Copestake, Ann, Flickinger, Dan, Sag, Ivan A. and Carl Pollard (2003): Minimal Recursion Semantics. An introduction.

Copestake, Ann, Lascarides, Alex and Dan Flickinger (2001): An Algebra for Semantic Construction in Constraint-based Grammars. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL 2001), Toulouse, France*.

Coulmas, Florian (1987): Höflichkeit und soziale Bedeutung im Japanischen. In: *Linguistische Berichte 107*. Opladen: Westdeutscher Verlag. pp. 44-62.

Dohsaka, K. (1990): Identifying the Referents of Zero-Pronouns in Japanese based on Pragmatic Constraint Interpretation. In *Proceedings of ECAI 1990* , 240-245.

Downing, P. (1996): *Numeral Classifer Systems: The Case of Japanese.* John Benjamins, Philadelphia.

Drozdzynski, Witold, Krieger, Hans-Ulrich, Piskorski, Jakub, Schäfer, Ulrich and Feiyu Xu (2004): Shallow Processing with Unification and Typed Feature Structures --- Foundations and Applications. In: *Künstliche Intelligenz*, vol.1, 2004, pp. 17-23.

EDR (1996). EDR ELECTRONIC DICTIONARY VERSION 1.5 TECHNICAL GUIDE. Technical report, Japan Electronic Dictionary Research Institute, Ltd. (http://www.iijnet.or.jp/edr/E_TG.html).

Farmer, A. K. (1984): Modularity in Syntax: A Study of Japanese and English. Massachusetts: MIT Press.

Fillmore, Charles J., Kay, Paul and Mary Catherine O'Connor (1988): Regularity and Idiomaticity in Grammatical Constructions: The Case of Let Alone. *Language* 64(3), 501-538.

Flickinger, Dan (2000): On building a more efficient grammar by exploiting types. *Natural Language Engineering*, *6 (1) (Special Issue on Efficient Processing with HPSG)*, 15 – 28.

Flickinger, Dan and Emily Bender (2003): *Compositional semantics in a multilingual grammar resource*. In: Bender et al. (eds): A Workshop on Ideas and Strategies for Multilingual Grammar Engineering, ESSLLI 2003.

Flickinger, Dan and Francis Bond (2003): A two-rule analysis of measure noun phrases. In Müller, S., ed.: Proceedings of the 10th International Conference on Head-Driven Phrase Structure Grammar, Stanford CA, CSLI Publications. 111–121.

Flickinger, Dan, Bender, Emily M. and Stephan Oepen (2003): *MRS in the LINGO Grammar Matrix. A Practical User's Guide*. DeepThought deliverable D 3.5 (technical report).

Fouvry, Frederik (2004): PERFORM: Performance modelling for declarative grammar models. Arbeits- und Ergebnisbericht Januar 2001-Januar 2004. Presentation of the research results.

Francis Bond, Sanae Fujita, Chikara Hashimoto, Kaname Kasahara, Shigeko Nariyama, Eric Nichols, Akira Ohtani, Takaaki Tanaka and Shigeaki Amano (2004b): The Hinoki Treebank: A Treebank for Text Understanding, In *Proceedings of the First International Joint Conference on Natural Language Processing (IJCNLP-04)*, Springer Verlag Lecture Notes in Computer Science.

Green, Georgia M. (1997): The Structure of Context: The Representation of Pragmatic Restrictions in HPSG. Studies in the Linguistic Sciences. Proceedings of the 5th annual meeting of the Formal Linguistics Society of the Midwest.

Gunji, T., and K. Hasida (1998b): Measurement and Quantification. In *Topics in Constrained-Based Grammar of Japanese*, ed. T.Gunji and K.Hasida. Dordrecht.

Gunji, Takao (1983): Generalized Phrase Structure Grammar and Japanese reflexivization. In: *Linguistics and Philosophy*, vol. 6, 115-156.

Gunji, Takao (1987): *Japanese Phrase Structure Grammar.* Dordrecht: Reidel.

Gunji, Takao (1991): An Overview of JPSG: A Constraint-Based Descriptive Theory for Japanese. In *Proceedings of Japanese Syntactic Processing Workshop* . Duke University.

Gunji, Takao (1996a): On Lexicalist Treatments of Japanese Causatives. In *Studies in the Universality of Constraint-Based Structure Grammars* , ed. T. Gunji, 61-89. Osaka University.

Hajic, J. (1998): Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. *Issues of Valency and Meaning*. Praha: Karolinum.

Harada, S. (1976): Honorifics. In *Japanese Generative Grammar* , ed. M. Shibatani, Vol. 5 of Syntax and Semantics. New York.

Hashimoto, Chikara and Francis Bond (2005): A Computational Treatment of V-V Compounds in Japanese. In Proceedings of the 12th International Conference on HPSG, pp.143-156.

Hashimoto, Chikara, Tanaka, Takaaki, Bond, Francis and Melanie Siegel (2005): Integration of a Lexical Type Database with a Linguistically Interpreted Corpus. In: *Proceedings of the 6th International Workshop on Linguistically Interpreted Corpora LINC-2005*.

Heine, Julia E. (1998): Definiteness Predictions for Japanese Noun Phrases. In *17th International Conference on Computational Linguistics: COLING-98*, 519-525.

Hellan, Lars and Petter Haugereid (2003): *NorSource - an excercise in the Matrix Grammar building design*. In: Bender et al. (eds): A Workshop on Ideas and Strategies for Multilingual Grammar Engineering, ESSLLI 2003.

Hill, Beverly, Ide, Sachiko, Ikuta, Shoko, Kawasaki, Akiko and Tsunao Ogino (1986): Universals of Linguistic Politeness. Quantitative Evidence from Japanese and American English. In: *Journal of Pragmatics 10*. pp. 347-371.

Hinds, John (1977): Particle Deletion in Japanese and Korean. In: Linguistic Inquiry, vol. 8, number 4. pp 602-604.

Hori, Motoko (1986): A Sociolinguistic Analysis of the Japanese Honorifics. In: *Journal of Pragmatics 10*. pp. 373-386.

Ide, Sachiko (1986): The Background of Japanese Sociolinguistics. In: *Journal of Pragmatics 10*. pp. 281-286.

Ikehara, Satoru, Shirai, Satoshi and Kentaro Ogura (1994): Criteria for evaluating the linguistic quality of Japanese to English machine translations. (in Japanese) Journal of Japanese Society for Artificial Intelligence, Vol.9, No.4, pp.93-103.

Ikeya, A. (1983): Japanese Honorific Systems. In Seoul Papers in Formal Grammar Theory. Proceedings of the 3rd Korean-Japanese Joint Workshop . Seoul: Hanshin Publishing Company.

Inoue, K. (1978): Nihongo no Bunpou Housoku. Tasishuukan, Tokyo.

Kageyama, Taro (2001): Word Plus: The Intersection of Words and Phrases. In J. v.d. Weijer and T. Nishihara

(eds.), *Issues in Japanese Phonology and Morphology*, pages 245-276, Mouton de Gruyter.

Kanayama, Hiroshi, Kentaro Torisawa, Yutaka Mitsuishi and Jun'ichi Tsujii (2000): A Hybrid Japanese Parser with Hand-crafted Grammar and Statistics. In *Proceedings of the 18th International Conference on Computational Linguistics, Coling 2000*. Saarbrücken, Germany.

Kasai, H., and S. Takahashi (2001): Coordination in Japanese. In Proceedings of the Third Formal Approaches to Japanese Linguistics Conference FAJL3.

Kasper, Robert (1995): The semantics of recursive modification. Paper presented at the Workshop on HPSG in Tübingen 1995.ms.

Katagiri, Y. (1991): Perspectivity and the Japanese Reflexive *`zibun '*. In *Situation Theory and its Applications*, ed. J. Barwise, J. M. Gawron, G. Plotkin, and S. Tutiya, Vol. 2, chapter 18, 425-447. CSLI.

Kiefer, Bernd, Krieger, Hans-Ulrich and Mark-Jan Nederhof (2000): Efficient and Robust Parsing of Word Hypotheses Graphs. In: Wahlster, Wolfgang (ed.): *Verbmobil: Foundations of Speech-to-Speech Translation.* Berlin, Springer, 280-295.

Kiss, Tibor (1995): *Infinite Komplementation. Neue Studien zum deutschen Verbum infinitum*. Linguistische Arbeiten 333. Tübingen: Max Niemeyer Verlag.

Kordoni, Valia and Julia Neu (2003): *Deep grammar development for Modern Greek.* In: Bender et al. (eds): A Workshop on Ideas and Strategies for Multilingual Grammar Engineering, ESSLLI.

Kuno, S. (1973): *The Structure of the Japanese Language*. Cambridge, Massachusetts, and London, England: MIT Press.

Kuroda, S.-Y. (1992): *Japanese Syntax and Semantics. Collected Papers.* Vol. 22 of Studies in Natural Language and Linguistic Theory.

Kurohashi, S., and M. Nagao (1994b): A Syntactic Analysis Method of Long Japanese Sentences Based on the Detection of Conjunctive Structures. *Coling* 20(4):507-534.

Kurohashi, S., and M. Nagao (1998): Building a Japanese Parsed Corpus while Improving the Parsing System. Research Report JSPS-RFTF96P00502, Japan Society for the Promotion of Science, March.

Lee, Dong-Young (1996): An HPSG Account of the Korean Honorification System. In: Studies in HPSG. Edinburgh Working Papers in Cognitive Science 12. pp. 165-190.

Lehmann, Sabine, Stephan Oepen, Sylvie Regnier-Prost, Klaus Netter, Veronika Lux, Judith Klein, Kirsten Falkedal, Frederik Fouvry, Dominique Estival, Eva Dauphin, Hervé Compagnion, Judith Baur, Lorna Balkan, and Doug Arnold (1996): TSNLP - Test Suites for Natural Language Processing. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 711-716, Kopenhagen, Denmark.

Makino, Seiichi, and Michio Tsutsui (1986): A Dictionary of Basic Japanese Grammar. Tokyo, The Japan Times, Ltd.

Manning, C., I. A. Sag, and M. Iida (1998): The Lexical Integrity of Japanese Causatives. In *Readings in Modern Phrase Structure Grammar* , ed. R. Levine, and G. Green. Cambridge University Press.

Manning, Christopher D. and Ivan A. Sag (1998): Dissociations Between Argument Structure and Grammatical Relations. In: A. Kathol, J.-P. Koenig, and G. Webelhuth (eds.): Lexical and Constructional Aspects of Linguistic Explanation. CSLI Publications, Stanford.

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

Martin, Samuel E. (1987): A Reference Grammar of Japanese. Tokyo: Charles E. Tuttle Company, second edition.

Matsumoto, Y., Kitauchi, A., Yamashita, T., Hirano, Y., Matsuda, H., Takaoka, K. and Asahara, M. (2000): Morphological Analysis System ChaSen version 2.2.1 Manual.

Matsumoto, Yo (1993): Japanese numeral classifiers: A study of semantic categories and lexical organization. *Linguistics*, 31:667–713.

McCawley, J. D. (1976): Relativization. In *Japanese Generative Grammar* , ed. M. Shibatani, Vol. 5 of Syntax and Semantics. New York.

McGloin, Hanaoka N. (1976): Negation. In *Japanese Generative Grammar*, ed. M. Shibatani, Vol. 5 of Syntax

and Semantics. New York.

McGloin, Naomi Hanaoka (1987): The Role of WA in Negation. In: Hinds, John and Maynard, Senko Kumiya and Iwasaki, Shoichi (eds*.): Perspectives on Topicalization. The Case of Japanese 'WA'*. Amsterdam: John Benjamins Publishing Company. pp. 165-184.

Metzing, Dieter, and Melanie Siegel (1994): Zero Pronoun Processing: Some Requirements for a Verbmobil System. Verbmobil-Memo 46, Universität Bielefeld.

Mitsue, Motomura (2001): Zibun as a residue of overt A-movement. In: *The Third Formal Approaches to Japanese Linguistics Conference*, MIT.

Miyagawa, S. (1986): Predication and Numeral Quantifier. In *Papers from the Second International Workshop on Japanese Syntax*, ed. W. J. Poser, 157-191. CSLI.

Miyagawa, S. (1989): Structure and Case Marking in Japanese. Academic Press, New York.

Miyata, Takashi, Akira Ohtani, and Yuji Matsumoto (2001). An HPSG account of the hierarchical clause formation in Japanese --- HPSG-based Japanese grammar for practical parsing ---. In *Proceedings of the 15th Pacific Asia Conference (PACLIC,15)*, pages 305-316.

Müller, Stefan (1997): *Spezifikation und Verarbeitung deutscher Syntax in Head-Driven Phrase Structure Grammar*. PhD thesis, Saarbrücken: University of the Saarland.

Nariyama, Shigeko, Nichols, Eric, Bond, Francis, Tanaka, Takaaki and Hiromi Nakaiwa (2005): Extracting representative arguments from dictionaries for resolving zero pronouns. In *Machine Translation Summit X*, 3--10, Phuket.

Nichols, Eric, Francis Bond, and Daniel Flickinger (2005): Robust ontology acquisition from machine-readable dictionaries. In *Proceedings of the International Joint Conference on Artificial Intelligence IJCAI-2005*, 1111--1116, Edinburgh.

Nightingale, S. (1996): *An HPSG Account of the Japanese Copula and Related Phenomena*. M.Sc. thesis, University of Edinburgh.

NIST (2002): *Automatic Evaluation of Machine Translation Quality using N-gram Co-Occurrence Statistics*.

Oepen, Stephan and John Carroll (2000): Ambiguity packing in constraint-based parsing. Practical results. In *Proceedings of the 1st Conference of the North American Chapter of the ACL*, pages 162-169, Seattle, WA.

Oepen, Stephan and John Carroll (2000): Performance Profiling for Parser Engineering. Journal of Natural Language Engineering, Special Issue on Efficient Processing with HPSG: Methods, Systems, Evaluation, pages 81-97.

Oepen, Stephan, Emily M. Bender, Uli Callmeier, Dan Flickinger and Melanie Siegel (2002a): Parallel Distributed Grammar Engineering for Practical Applications. In *Proceedings of the Workshop on Grammar Engineering and Evaluation.* Coling 2002 Post-Conference Workshop. Taipei, Taiwan.

Oepen, Stephan, Kristina Toutanova, Stuart Shieber, Christopher Manning, Dan Flickinger and Thorsten Brants (2002b): The LinGO Redwoods Treebank: Motivation and Preliminary Applications. In Proceedings *of the 19th International Conference on Computational Linguistics (COLING 2002),* Taipei, Taiwan, pp. 1253-7.

Ono, Kiyoharu (1996): Syntactic Behaviour of Case and Adverbial Particles in Japanese. In *Australian Journal of Linguistics 16*, 81-129.

Oshima, David Yoshikazu (2002): Out of Control: A Unified Analysis of Japanese Passive. In *Proceedings of HPSG 2002*.

Paik, Kyong-Hee, Bond, Francis (2002): Spatial representation and shape classifers in Japanese and Korean. In Beaver, D.I., Casillas Mart´ýnez, L.D., Clark, B.Z., Kaufmann, S., eds.: The Construction of Meaning. CSLI Publications, Stanford CA (2002) 163–180

Papineni, K., S. Roukos, T. Ward, W.-J. Zhu (2001): BLEU: a Method for Automatic Evaluation of Machine Translation. *IBM Research Report RC22176 (W0109-022)*.

Peters, Carol, Julio Gonzalo, Martin Braschler, Michael Kluck (Eds.) (2004): Fourth Workshop of the Cross-Language Evaluation Forum, CLEF 2003, Trondheim, Norway, August 2003. Revised papers. *Lecture Notes in Computer Science* 3237, Springer.

Pollard, Carl and Ivan A. Sag (1994): *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press.

Radford, Andrew (1993): Head-Hunting: On the Trail of the Nominal Janus. In Greville Corbett, N. Fraser and S. McGlashan (eds.), *Heads in Grammatical Theory*, pages 73-113, Cambridge: Cambridge University Press.

Sadakane, K. & M. Koizumi (1995): On the nature of the "dative" particle *ni* in Japanese. *Linguistics* 33.

Sag, Ivan (2003): Coordination and Underspecification. In *Proceedings of the 9th International Conference on HPSG*. Stanford University.

Sag, Ivan A., Wasow, Thomas and Bender, Emily M. (2003): Syntactic Theory: A Formal Introduction. Stanford, CA: CSLI, second edition.

Sag, Ivan, and Thomas Wasow (1999): Syntactic Theory: An Introduction, CSLI Publications.

Shibatani, M. (1978): Nihongo no Bunseki. Tasishuukan, Tokyo.

Shibatani, Masayoshi and Kageyama, Taro (1988): Word Formation in a Modular Theory of Grammar: Postsyntactic Compounds in Japanese. *Language 64*, 451-484.

Siegel, Melanie (1996a): Definiteness and Number in Japanese to German Machine Translation. In: Gibbon, Dafydd (ed.): Natural Language Processing and Speech Technology. Berlin: Mouton de Gruyter, pages 137-142.

Siegel, Melanie (1996b): Die maschinelle Übersetzung aufgabenorientierter japanisch-deutscher Dialoge. Lösungen für Translation Mismatches. Ph.D.thesis.

Siegel, Melanie (1999): The Syntactic Processing of Particles in Japanese Spoken Language. In: Wang, Jhing-Fa and Wu, Chung-Hsien (eds.): *Proceedings of the 13th Pacific Asia Conference on Language, Information and Computation,* Taipei 1999.

Siegel, Melanie (2000): HPSG Analysis of Japanese. In: W.Wahlster(ed.): *Verbmobil: Foundations of Speech-to-Speech Translation.*, Springer Verlag.

Siegel, Melanie and Emily M. Bender (2002): Efficient deep processing of Japanese. In: Proceedings of the 3rd Workshop on Asian Language Resources and Standardization, Coling 2002, Taipei.

Siegel, Melanie and Emily M. Bender (2004): Head-Initial Constructions in Japanese. In: Stefan Müller (ed.): *Proceedings of the 11th International Conference on Head-Driven Phrase Structure Grammar*, Center for Computational Linguistics, Katholieke Universiteit Leuven, pages 244-260.

Sirai, Hidetoshi and Takao Gunji (1998): Relative Clauses and Adnominal Clauses. In *Topics in Constrained-Based Grammar of Japanese*, ed. T.Gunji and K.Hasida. Dordrecht.

Smith, Jeffrey D. (1999): English Number Names in HPSG. In GertWebelhuth, Jean-Pierre Koenig and Andreas Kathol (eds.), *Lexical and Constructional Aspects of Linguistic Explanatio*n, pages 145-160, Stanford, CA: CSLI.

Sparck-Jones, K. (1994): Towards better NLP system evaluation. Proceedings of the Second ARPA Workshop on Human Language Technology. San Mateo, CA: Morgan Kaufmann.

Tanaka, Takaaki, Bond, Francis, Oepen, Stephan and Sanae Fujita (2005): High precision treebanking -- blazing useful trees using POS information. In *ACL-2005*, 330--337, 2005.

Tsuda, H., and Y. Harada (1996): Semantics and Pragmatics of Adnominal Particle NO in Quixote. In *Studies in the Universality of Constraint-Based Structure Grammars.*, ed. T. Gunji. Osaka.

Tsujimura, N. (1996): *An Introduction to Japanese Linguistics* . Blackwell, Cambridge.

Uda, Chiharu (1996): ARG-S Feature and Valence Features: More Evidence From Japanese Passives. In *Studies in the Universality of Constraint-Based Structure Grammars*, ed. T. Gunji, 203-215. Osaka University.

Uda, Chiharu (2001): Clausal Complement or Adverbial Clause?: Toward an Integral Account of the Japanese Internally-Headed Relative Clause. In: Proceedings of HPSG 2001.

Uszkoreit, Hans, Callmeier, Ulrich, Eisele, Andreas, Schäfer, Ulrich, Siegel, Melanie and Jakob Uszkoreit (2004): Hybrid Robust Deep and Shallow Semantic Processing for Creativity Support in Document Production. In *Proceedings of KONVENS 2004*, Vienna, Austria.

Voorhees, E.M. and Lori P. Buckland (eds.) (2004): NIST Special Publication 500-261: The Thirteenth Text REtrieval Conference Proceedings (TREC 2004).

Wahlster, W. (ed.) (2000): *Verbmobil: Foundations of Speech-to-Speech Translation.*, Springer Verlag.

Watanabe, T. (2000): Object Topicalization, Passive, and Information Structure in Japanese. In *Proceedings of the 14th Pacific Asia Conference on Language, Information and Computation*, ed. A. Ikeya and M. Kawamori, 339-344. Waseda University International Conference Center, Tokyo. Logico-Linguistic Society of Japan.

Yatabe, S. (1993): *Scrambling and Japanese Phrase Structure*. PhD thesis, Stanford University.

Yoshimoto, Kei (1997): *Tense and Aspect in Japanese and English* . PhD thesis, Universität Stuttgart.

Zwicky, Arnold M. (1993): Heads, Bases, and Functors. In Greville C. Corbett, Norman M. Fraser and Scott McGlashan (eds.), *Heads in Grammatical Theory*, pages 292-315, Cambridge: Cambridge University Press.

Appendix A: Grammar Installation

The JACY grammar can be downloaded from http://www.delph-in.net/jacy under an open-source licence.

This is needed for running the JACY grammar :

- Basic requirements: Installation of ACL6.0 with CLIM and all patches, Linux installed with Japanese, Open Motif (http://www.openmotif.org)
- The LKB grammar development system (http://www-csli.stanford.edu/~aac/lkb.html. You will find detailed installation instructions there)
- The ChaSen morphological analyzer (http://chasen.aist-nara.ac.jp/. You will find detailed installation instructions there)

Although the lkb will run standalone, there are problems with Japanese input. The recommended way to run it is from inside emacs, using the eli interface. Install the lkb and eli (following the instructions in http://www-csli.stanford.edu/~aac/emacslkb.html). Problems or questions concerning LKB in general can be directed to lkb-bugs@csli.stanford.edu.

You need to run Lisp with the EUC locale (ja_JP.EUC-JP) and be sure emacs uses EUC for the process encoding in the *common-lisp* buffer. Use the .emacs.jp file on the JACY download site and adapt the paths. Then, your .emacs must be told that the .emacs.jp exists:

(when (file-exists-p (concat user-home "/.emacs.jp")) (load (concat user-home "/.emacs.jp") nil t t))

You will also need the file .clinit.cl. Finally, for running [incr tsdb()] and PET on the Japanese grammar, you will need .tsdbrc. Both can be downloaded on http://www.delph-in.net/jacy as well.

Now load everything, LKB, MRS, plus [incr tsdb()]:

Open emacs

Start Lisp with M-x japanese

```
:ld ~/src/lkb/src/general/loadup
(pushnew :lkb *features*)
(pushnew :mrs *features*)
(compile-system "tsdb" :force t)
```

Load the grammar with (read-script-file-aux "~/japanese/lkb/ascript") (your path to the grammar).

You can parse a sentence by typing (do-parse-tty "SENTENCE") in the emacs window.

Using JACY with itsdb

Install itsdb from the CSLI ftp site (http://lingo.stanford.edu/ftp/), following the instructions in the manual (http://lingo.stanford.edu/ftp/itsdb_documentation.tgz).

The latest version of JACY and versions of itsdb later than 2003-05-20 should work as is with Japanese.

M-x tsdb

Note:
Japanese test sentences should be in euc-jp.

To get itsdb to count Japanese words, you need to segment the test sentences at some stage. This can be done during import.

> if there is a _global_ `preprocessing hook', [incr tsdb()] import will pipe everything through it and use the _second_ value that it returns as the `i-length' field; e.g.
>
> (setf *tsdb-preprocessing-hook* "lkb::chasen-preprocess-for-pet")
>
> will enable that hook globally, and once you use a definition of this function that counts correctly (no good doing length() on a variable _after_ using the destructive nreverse() on it :-{), you will notice that (i) imports from text files are much slower and (ii) `Browse -- Test Items' will show ChaSen word counts for the `i-length' field.

Note that because the import can now take actual time (half a second per item or so), the [incr tsdb()] progress meter should advance correctly during the import from text file function (this does not work on versions older than 2003-06),

There is an example of `user-fns.lsp' for JaCY that enables the *tsdb-preprocessing-hook*, when [incr tsdb()] is loaded _before_ the grammar. (You could also set this in `~/.tsdbrc', but then it would affect everything you do, no matter which grammar was used.)

from user-fns.lsp:

```
;;;
;;; hook for [incr tsdb()] to call when preprocessing input (going to the PET
;;; parser or when counting `words' while import test items from a text file).
;;;

(defun chasen-preprocess-for-pet (input)

(preprocess-sentence-string input :verbose nil :posp t))

#+(or :pvm :itsdb)

(setf tsdb::*tsdb-preprocessing-hook* "lkb::chasen-preprocess-for-pet")
```

Using JACY with PET

Install PET following the instructions at http://www.coli.uni-sb.de/pet/documentation.php3.

You need to segment the Japanese, for example by preprocessing with chasen:

> chasen -F"%m " | cheap ~/japanese/japanese.grm

reading `pet/japanese.set'...

loading `japanese.grm' (Japanese (jan-03))

16674 types in 1.7 s

Using JACY with itsdb and PET

Install itsdb and PET.

You can run Japanese with a cpu defined in your .tsdbrc (on the JACY download site, substituting your pathnames).

After starting lkb-ja and itsdb in emacs:

Choose the cpu in the normal way by evaluating

(tsdb::tsdb :cpu :nihongo :file t) in the *common-lisp* buffer:

LKB(2): (tsdb::tsdb :cpu :nihongo :file t)

The preprocessor calls a function defined in usr-fns.lisp that runs chasen on the input, the combination of "-yy" "-default-les" takes the output and produces default lexical types for unknown words.

Appendix B: Author Index

A

Ahn 178

Asahara 1, 66, 158, 178, 181

B

Baldwin 81, 176, 178

Beavers 178

Bender 17, 18, 65, 71, 130, 136, 137, 139, 158, 159, 167, 178, 179, 180, 181, 182, 183

Bender et al. 2002 18

Black 165, 178

Bond 1, 2, 45, 65, 69, 81, 116, 158, 164, 172, 173, 174, 178, 179, 180, 182, 183

Borsley 130, 178

Briscoe 165, 178, 179

Buckland 165, 183

C

Callmeier 1, 158, 159, 170, 179, 182, 183

Carroll 1, 159, 160, 165, 167, 178, 179, 182

Chinchor 165, 179

Chung 62, 179, 183

Collins 165, 179

Copestake 1, 3, 18, 170, 178, 179

Copestake 2001 18

Coulmas 145, 147, 179

D

Dohsaka 146, 179

Downing 66, 68, 179

Drozdzynski 54, 179

E

Eisele 179, 183

F

Farmer 98, 179

Fillmore 176, 179

Flickinger 3, 69, 158, 159, 178, 179, 180, 182

Fouvry 165, 180, 181

Fujita 55, 178, 180, 183

G

Green 147, 180, 181

Grimshaw 14

Grishman 178

Gunji 16, 18, 21, 42, 51, 52, 81, 82, 85, 88, 90, 91, 92, 114, 115, 118, 130, 146, 149, 180, 183

Gunji 83 21

Gunji 87 16

Gunji 91 16

H

Hajic 165, 180

Harada 51, 105, 145, 146, 180, 183

Hashimoto 2, 121, 164, 177, 178, 180

Hasida 91, 92, 180, 183

Haugereid 160, 180

Heine 45, 180

Hellan 160, 180

Hill 145, 180

Hinds 124, 180, 182

Hori 145, 180

I

Ide 145, 180

Ikehara 172, 178, 180

Ikeya 145, 146, 149, 180, 184

Inoue 38, 180

K

Kageyama 132, 180, 183

Kanayama 1, 181

Kasai 121, 181

Kasper 17, 128, 181

Katagiri 52, 53, 54, 181

Kathol 178, 181, 183

Kiefer 156, 181

Kim 62, 132, 178, 179

Kiss 92, 95, 181

Koizumi 101, 183

Kordoni 160, 181

Krieger 179, 181

Kuno 88, 96, 114, 118, 145, 178, 181

Kuroda 97, 124, 181

Kurohashi 9, 10, 169, 181

L

Lascarides 179

Lee 153, 181

Lehmann 165, 181

M

Makino 50, 133, 134, 170, 181

Manning 42, 52, 181, 182

Marcus 165, 178, 181

Martin 138, 181, 182

Matsumoto 1, 65, 66, 158, 178, 181, 182

McCawley 51, 181

McGloin 145, 148, 181, 182

Mester 14

Metzing 151, 157, 182

Minnen 179

Mitsue 50, 52, 182

Mitsuishi 181

Miyagawa 85, 90, 91, 182

Miyata 2, 182

Müller 17, 95, 179, 182, 183

N

Nagao 9, 10, 169, 181

Nariyama 35, 178, 180, 182

Nederhof 181

Neu 160, 181

Nichols 173, 178, 180, 182

Nightingale 83, 89, 107, 110, 111, 182

O

Oepen 1, 81, 158, 160, 164, 165, 167, 173, 178, 180, 181, 182, 183

Ogura 45, 178, 180

Ohtani 178, 180, 182

Ono 96, 101, 182

Oshima 38, 182

P

Paik 65, 178, 182

Papineni 165, 182

Peters 165, 182

Pollard 1, 2, 4, 5, 7, 16, 38, 85, 90, 93, 95, 115, 130, 144, 146, 148, 150, 153, 179, 182

R

Radford 137, 183

S

Sadakane 101, 183

Sag 1, 2, 4, 5, 7, 11, 16, 38, 42, 49, 50, 52, 85, 90, 93, 95, 115, 120, 121, 130, 131, 144, 146, 148, 150, 153, 179, 181, 182, 183

Schäfer 179, 183

Shibatani 132, 180, 181, 183

Shirai 180

Siegel *ii*, *1*, *45*, *65*, *130*, *136*, *137*, *139*, *142*, *151*, *156*, *157*, *167*, *178*, *179*, *180*, *182*, *183*

Sirai 17, 81, 82, 183

Smith 138, 183

Sparck-Jones 166, 183

T

Takahashi 121, 181

Tanaka 12, 40, 51, 55, 83, 120, 131, 143, 178, 180, 182, 183

Torisawa 181

Tsuda 105, 183

Tsujii 181

Tsujimura 85, 88, 98, 102, 106, 183

Tsutsui 50, 133, 134, 170, 181

U

Uda 38, 55, 56, 81, 183

Uszkoreit 1, 183

V

Voorhees 165, 183

W

Wahlster 86, 116, 181, 183, 184

Wasow 49, 50, 183

Watanabe 116, 184

Y

Yatabe 124, 184

Yoshimoto 2, 36, 92, 115, 184

Z

Zwicky 130, 184