Gefördert vom
bmb+f
Bundesministerium für
Bildung, Wissenschaft,
Forschung und Technologie

Verbmobil
Verbundvorhaben

Stylebook for the Japanese Treebank in VERBMOBIL

Yasuhiro Kawata, Julia Bartels

Seminar für Sprachwissenschaft, Universität Tübingen

Vm **Report 240**
September 29, 2000

September 29, 2000

Yasuhiro Kawata, Julia Bartels

Computerlinguistik
Seminar für Sprachwissenschaft
Eberhard-Karls-Universität Tübingen
Wilhelmstr. 113
72074 Tübingen

Tel.: (07071) - 29 74279
Fax: (07071) - 55 13 35
e-mail: ykawata@sfs.nphil.uni-tuebingen.de

Gehört zum Antragsabschnitt: 6.7 Baumbanken

Die vorliegende Arbeit wurde im Rahmen des Verbundvorhabens Verbmobil vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) unter dem Förderkennzeichen 01 IV 701 M0 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei den Autoren.

Acknowledgment

We would like to acknowledge the support of our Tübingen colleagues, especially the members of the Verbmobil team, Valia Kordoni, Sandra Kübler, and Heike Telljohann.

The development of the Tübingen VERBMOBIL treebanks was greatly facilitated by a number of Verbmobil partners whose contributions went well beyond the call of duty. Hans Uszkoreit and his colleagues at the Universität des Saarlandes kindly provided us with the graphical annotation tool *Annotate V2.3* which was developed as part of the research project (*Teilprojekt* C3; Principal investigators: Uszkoreit/Smolka) *Nebenläufige grammatische Verarbeitung* (NEGRA) in the Sonderforschungsbereich 378. The *Annotate* tool provides human annotators with a graphical, user-friendly interface for annotating and editing trees and also provides database support for maintaining large treebanks. We would like to express our special gratitude to Thorsten Brants, who has kindly and generously provided us with software support and user assistance for the *Annotate* tool from the very beginning of the Tübingen treebank project.

Contents

Chapter 1

Introduction

This stylebook for the Japanese treebank describes the design principles of the annotation scheme for the treebank of Japanese constructed at the Eberhard-Karls-Universität Tübingen as a part of the VERBMOBIL-II project. VERBMOBIL-II is a joint research project funded by the German Ministry for Education and Research (BMBF) that is conducted by a consortium of universities, research centers, and information technology companies. The initial phase of the project (VERBMOBIL-I) lasted four years from 1993 to 1996, and the second phase (VERBMOBIL-II) commenced in 1997 and concluded in September 2000.

The overriding goal of the VERBMOBIL-II project is to develop a speaker-independent spontaneous speech translation system. To this end, a number of scenarios have been defined as a testbed for the development of software prototypes. During the first phase of the project (VERBMOBIL-I) the scenario consisted of dialogs, in which two discourse participants negotiate business appointments. In VERBMOBIL-II this scenario is significantly extended along various dimensions. In order to obtain realistic and quantitatively significant data for the relevant scenarios, a major data collection initiative for spoken-language dialogs was launched. The dialogs were recorded in a variety of settings and were transcribed according to mutually agreed upon standards. The transcribed data were then further annotated for the purposes of signal processing and linguistic analysis.

The treebank project, carried out by the Division of Computational Linguistics at the Eberhard-Karls-Universität Tübingen (Lehrstuhl Prof. Hinrichs), constitutes part of the overall effort of linguistic annotation within the VERBMOBIL-II project. Treebanks for German, English, and Japanese have been developed. The present report focuses on the Tübingen treebank for Japanese only; the Tübingen treebank for German is described in (Stegmann et al. 2000), the one for English in (Kordoni 2000).

The size for the Japanese treebank has reached approximately 18,000 entries at

the end of the VERBMOBIL-II project phase. The overall annotation scheme for the Japanese treebank was negotiated with relevant partners in the VERBMOBIL-II consortium.

The purpose of the stylebook is to guide the treebank annotators and to provide the users of the treebank with understanding of the data. The Japanese treebank described here has been developed in the context of the research and development of a stochastic parsing method [1].

The treebank is aimed to be one of the useful resources for development of various kind of context-free techniques in NLP, and also for linguistic research with serious orientation to the empirical facts. The treebank provides its users with the data describing the transcribed utterances in the dialogs in terms of the constituency, dominance and linear precedence relations among the constituents, and the grammatical function of each constituent. From a treebank, we can extract the set of context-free phrase structure rules, which generates the language consisting of all the terminal token strings of the treebank together with the reliable statistical data based on the fact. Thus, the treebank will serve as training data for various computational systems with learning mechanisms. One can find specific linguistic string patterns in a corpus without any annotation. One can search for specific tag string patterns in combination with specific tokens in a tagged corpus without structural annotations. On top of them, one would be able to find out more interesting things about the structures of the language in the treebank.

The treebank contains much human insight both explicitly and implicitly. Explicitly, in terms of the formal specifications of the treebank, and also implicitly because human annotators are not restrained from referring to his or her own linguistic knowledge and intuition. For example, assuming the human annotators follow the chronological order of the transcriptions while annotating, they cannot be free from the contextual effects, for example, the word priming effects[2] that they unconsciously receive from the preceding texts. Annotators are expected not only to pay attention to the structural descriptions but also to understand the contents of the text. Although there are limitations for the treebank to be more than a form of syntactic representations, we believe that interesting facts about the language will be discovered in the linguistic forest.

---

[1] In this stylebook, the Japanese strings are not always accompanied by the English equivalent. For, among the readers of this stylebook, those who concern the contents of the Japanese string, that is, annotators and linguists, are assumed to be able to understand basic Japanese, and those who only concern formal specifications do not need to know what the example strings mean.

[2] The effect that previously presented word stimuli cause to the following language processing by the hearer, that is, facilitation and inhibition.

Chapter 2

The Japanese language

This chapter presents a brief overview of the Japanese language. It is not possible to cover all the important issues and discussions in Japanese linguistics within a few pages of this introductory chapter, but instead we try to familiarize readers with some of the specific linguistic characteristics of the Japanese language by picking up some key words.

Typology Japanese is often categorized as an agglutinating type of language. It is not clear where morphology ends and syntax begins. The right edge of the constituent, such as a derivational morpheme or a postposition at the end of the phrase, determines its possible right continuation class. Some of the apparent morphological processes can be recursive. Word compounding is often unpredictable especially among Sino-Japanese words.

Head final Japanese is also known as a head final language. The right hand side of the constituent is generally responsible for determining the nature of the whole constituent. The verb appears at the end, and postpositions instead of prepositions encode grammatical functions (Tsujimura 1996).

Scrambling — free word order — Similarly to other verb final type languages, Japanese is often said to be a free word order language[1] (Tsujimura 1996). Scrambling the phrasal categories within the same predication domain is mostly legitimate as long as the head appears at the end. Since the position of the subject and the position of the object of their predicate are not predictable in most cases, the claim of a subject object asymmetry often found for other languages finds few

[1]German is also known to be a free word order language, and its subordinate clauses are head final (Reape 1994).

supporting evidences in Japanese. The position of the interrogative words is not at all particular in comparison with the position of the noninterrogative words. Therefore, the movement analysis for this language has not established its place.

Support Verb In Japanese, there is a support verb[2] *suru*, which is semantically empty and bound to the predicative noun turning the compound structure into a verb[3].

Complex predicate Semantically complex expressions such as the passive and the causative are materialized in the morpho-syntactic forms. The passive and the causative form of *tabe-ru* (eat-present) would be *tabe-rare-ru* (eat-passive-present) and *tabe-sase-ru* (eat-causative-present) respectively. Various constructions, such as benefactive expressions to show directions of give and take, and honorific expressions to show respect to someone, are fused in the verbal forms. There are different views of the syntactic structure of sentences with such complex predicates depending on how the string is tokenized.

PRO-drop — missing arguments — There are languages called PRO-drop languages in the linguistic literature[4] The subject is normally dropped in those languages when it is recoverable from the context[5]. In the Japanese dialog in the treebank, one finds that, not only the subject, but also the object and other arguments, which grammarians might consider to be obligatory complements, are frequently missing depending on the context. This may be attributed to the characteristic of conversational dialogues, but it seems that missing arguments are a phenomena of the Japanese language in general.

Orthography The orthography does not demand a space between words. With the agglutinating morpho-syntactic characteristics of the language and the orthographic convention, the notion of word and the notion of morpheme are far less clear than those of European languages[6]. Therefore, the tokenization of the given string is still a research issue.

---

[2]Also mentioned as light verb (Tsujimura 1996)

[3]Expressions, whose literal translation would be "do the washing", or "do the travel", are frequently found. Arabic, for example, is known to have a similar construction.

[4]Also mentioned as zero-anaphora or null-anaphora (Bos and Heine 2000), (Tsujimura 1996).

[5]For example, the subject is normally dropped in Spanish because it is recoverable from verbal inflections.

[6]For example, German allows complex word compounding, and one should sometimes tokenize the string instances in order to be able to look them up in the dictionary.

Chapter 3

Basic Principles

This chapter presents some general assumptions for the annotators and the treebank users before moving to the specific levels of descriptions.

3.1 Segmenting dialogs

Dialogs are first mechanically segmented into turns that are taken by the dialog participants. Sentence boundaries are usually taken for granted in the linguistic theories as a primary segment in the human language, however, they can never be detected in the acoustic signals of the actual human dialog. Hence, it is natural to take dialog turns into consideration as preliminary segments.

Thus, a turn may contain more than a few discrete units such as sentences. In this case, annotators may divide a turn into several trees. On the other hand, a would-be sentence may be interrupted by the other speaker snatching his turn in the middle of the utterance. In this case, the turn may constitute only a fragment of a sentence. There are also cases where an utterance by one person is divided into more than one turn which would form a complete and well-formed constituent. In this case, since the turn is the delimiting unit prior to the "sentence" in the treebank, each turn will also constitute only fragments of a sentence. [1]

Turns are farther segmented in the treebank whenever the human transliterator put a period or question mark. Therefore, a tree ends in a period or question mark.

3.2 Form of the tree

The treebank consists of sets of tree diagrams without crossing branches. Each tree representing the syntactic structure of an utterance transcribed from recorded

[1] See also "Transliteration spontansprachlicher Daten" (Burger 1997).

material according to the convention of VERBMOBIL project (Burger 1997). A tree is defined formally as a specific directed acyclic graph being characterized as follows:

1. There is a root vertex that has no dominating vertex. From that root vertex, there is always a path to each vertex of the tree.
2. Each vertex except the root vertex is dominated by only one vertex.
3. Succeeding vertices of each vertex are linearly ordered from left to right.

A context-free language can be defined by a set of trees with these formal characteristics. We assume that that context-free language is an approximation of the language to which the set of terminal symbol strings of the trees belong to.

A tree in the treebank consists of a set of terminal symbols (word tokens), preterminal symbols (POS tags, see Chapter 4), nonterminal symbols (node labels, see Chapter 5), and edges between nonterminal symbols. Each edge is tagged with an edge label (see Chapter 6) in our treebank, which enhances the complexity that the tree diagram can express without changing the generative capacity of the context-free formalism.

The distinguished root category is not unique for the obvious reason that human utterances are not always complete sentences, but often only a sequence of fragments. It is not a problem for the treebank to follow the formal specifications presented above because root categories fall into a finite set for any given set of trees. With these formal characteristics, treebanks serve as recyclable data for any kind of context-free approach to NLP.

3.3 Annotation strategies

There are common strategies in principle among three VERBMOBIL treebanks of German, English[2], and Japanese:

[2]See also the stylebooks for the German (Stegmann et al. 2000) and English (Kordoni 2000) treebank.

1. **Longest match** As many constituents as possible are included in a syntactic structure provided the whole construction is syntactically and semantically well-formed.
2. **Flat clustering** As many constituents as possible are clustered on the same level.
3. **High attachment** Phrase attachment to the higher node in the tree is preferred if the ambiguity cannot be solved by the human annotator.

A linguistic constituent is represented by a nonterminal node spanning over the substring of word-POS pairs. The annotator combines constituents from left to right, in a bottom-up manner, where a constituent is found. Most of the possible structural ambiguities in attaching phrases should be disambiguated by the human annotator by consulting context and other relevant parameters. If any ambiguities remain to give human annotator a real problem in disambiguation, they are attached to the higher level of the constituent node in the tree structure.

On the other hand, there are annotation strategies that apply only to the Japanese treebank but not to the German and the English treebanks:

1. **Unary branching is avoided** because, prior to axiomatic statements, we want to know the distribution of the classes of POS tags and syntactic category nodes with respect to the other classes[3].
2. Parts of the trees result in **left branching** in general because of the head-final nature of the Japanese language.

One of the recurring issues in the annotation of treebank concerns the resolution of structural ambiguity. For example, *keikaN ga jiteNsha de nigeru dorobou o oikaketa.* 'the policeman chased the theft who is running away by bicycle' is ambiguous depending on which node to attach the phrase *jiteNsha de* ('by bicycle').

The human treebank annotators are not restrained from taking the context naturally into considerations. Consequently, the same substring could possibly be described differently. In other words, local ambiguities may be sorted out in one way or the other by the linguistic and nonlinguistic knowledge of the human annotators. The structural ambiguities, which have always been a serious snag in the rationalist symbolic approaches, are disambiguated in each instance in the treebank by the human treebank annotator according to their expert knowledge, world knowledge, and the context in which each instance appears.

[3]In other words, instead of dictating, for example, pronouns are noun phrases, proper nouns are noun phrases, common nouns are noun phrases, and so forth, we would like to find out which classes of words occur, for example, immediately on the left of a particular postposition. Then later, we will know better what noun phrases may consist of.

Chapter 4

POS tags

The Japanese POS tagset used in the VERBMOBIL-II Japanese treebank was originally designed in Tübingen for the Japanese treebank. It has 72 distict POS tags. Familiar major syntactic categories, such as nouns, verbs, adjectives, and adverbs are farther distinguished into subcategories in terms of morphological and semantic features. The basic ideas of the part of speech tagset are taken from common assumptions in grammar books in general and from one of the tokenizing taggers for Japanese available in public domain[1]. In the course of developing the treebank, the tagset has been modified in order to accommodate the specifically romanized and tokenized transcriptions following the convention of VERBMOBIL-II (Burger 1997) [2]. The POS tags are designed to be maximally mnemonic. In general, nominals begin with **N**, postpositions with **P**, verbals with **V**, adjectivals with **A**, and adverbials with **ADV**. Other examples are **CNJ** for conjunctions and **CD** for cardinal numbers.

Token strings are tagged automatically in the first place by the rule-based part of speech tagger "Brill Tagger" (Brill 1992), and are later corrected manually. Further corrections are made during the treebank annotation. If the annotator finds a simple tagging error while working on the annotation tool, the error should be corrected.

In this chapter, we present the POS tags by categorizing and subclassifying words. We will also illustrate the subclassification with examples. For a complete list of part of speech labels, see Appendix A.

[1]Most Japanese taggers are built in with a tokenizer because of the morpho-syntactic nature of the Japanese language mentioned in Chapter 2.

[2]Though we assume that the transcribed data is tokenized consistently and correctly, mistokenized strings remain. In these cases the annotators should mark the mistokenized string(s) with the temporary POS tag **xxx**, that can later be detected and fixed by manipulating the exported data.

4.1 POS categories

Japanese is not special among other languages in the point that the major part of speech categories are recognized, such as verbs, adjectives, nouns, adverbs, interjections, and conjunctions. In addition, Japanese has a prominent class of postpositions.

The POS categories consist of **inflecting** and **non-inflecting** ones. In Japanese, verbs and the so-called i-adjectives, whose present tense ending form is *-i*, are inflecting. Verbs inflect in the -Ru/-Ta paradigm, and i-adjectives in the -i/-ku paradigm. In contrast with i-adjectives, the so-called na-adjectives[3] and the attributive adjectives are non-inflecting. The other non-inflecting wordclasses are nouns, postpositions, adverbs, conjunctions, and interjections.

Most major POS categories consist of **unbound forms** (free forms) and **bound forms**[4]. Thus, there are verbs and particle verbs. There are adjectives and particle adjectives. There are adverbs and particle adverbs. There are nouns, noun prefixes, and noun suffixes, and so on. Conjunctions and interjections are unbound forms, whereas the postpositions constitute a group of bound forms.

In the following sections, each POS category will be defined and subclassified in more detail. Then, the set of POS tags of each category is presented with examples in the tables.

4.2 Verbs

Verbs are inflecting words which end in the -Ru/-Ta paradigm. Verbs are distinguished according to their morphology whether they are in finite (fin), imperative (imp), conditional (cnd), participle (te), base (bas) form, or appear with another inflectional ending.

Table 4.1 gives an overview of the Japanese verbal inflection system of the forms in the treebank[5].

In Japanese, we distinguish full verbs, support verbs, auxiliary verbs, and particle verbs. Accordingly, the complete POS tagset for verbs shown in Table 4.2 devides into four subclasses.

---

[3]The na-adjective is called "adjectival noun" (Tsujimura 1996), because it bears characteristics similar to both nouns and adjectives.

[4]Due to the agglutinating nature of the language, the distinction between free form and bound form is problematic and depends on the theory applied. We define boundness here as not being able to appear by itself in that function or meaning.

[5]Capital letters in the representations of the inflectional endings indicate that there are allomorphes to the ending, for example, *-Ru* stands for *-ru, -u,* and so forth. See (Rickmeyer 1995) for a more detailed description.

Classes	Inflectional ending	Examples
Vfin	*-Ru* (Present)	*taberu, nomu*
	-Ta (Perfect)	*tabeta, noNda*
	-You (Future)	*tabeyou, nomou*
Vimp	*-e/ro/i* (Imperative)	*nome, kimero, kudasai*
Vcnd	*-Reba* (Conditional)	*tabereba, nozokeba*
	-Tara(ba) (Conditional perfect)	*kiitara, shimashitaraba*
Vte	*-Te* (Participle)	*aite, noNde*
Vbas	*-I* (Base form)	*kimari, tabe*
V	*-Azu* (Negation)	*tomarezu, kikazu*
	-Tari (Exemplativ)	*shirabetari, noNdari*

Table 4.1: Verbal inflection system

POS tag	Description	*Examples*
Vfin	Verb finite	*kimeru, kimemashita*
Vcnd	Verb conditional	*kimetara, kimereba*
Vimp	Verb imperative	*kimero*
Vte	Verb participle	*kimete*
Vbas	Verb base form	*kime*
V	Verb	*kimetari, kimezu*
VAUXfin	Auxiliary verb finite	*iru, ita*
VAUXcnd	Auxiliary verb conditional	*ireba*
VAUXte	Auxiliary verb participle	*ite*
VAUXbas	Auxiliary verb base form	*itadaki*
VAUX	Auxiliary verb	*mitari*
VSfin	Support verb finite	*suru, shita*
VScnd	Support verb conditional	*shitara, sureba*
VSimp	Support verb imperative	*shiro*
VSte	Support verb participle	*shite*
VSbas	Support verb base form	*shi*
VS	Support verb	*shitari, sezu*
PVfin	Particle verb finite	*da, desu, deshita*
PVcnd	Particle verb conditional	*deshitara*
PVte	Particle verb participle	*deshite, de*
PV	Particle verb	*dattari*

Table 4.2: Verbal POS tags

Full verbs (V) are verbal categories which function by themselves as predicates, for example,

> *shucchou no keN* <u>*kimemashou*</u> *ka* ('Shall we decide the business trip affair?')
> <u>*wakarimashita*</u> ('I understand').

Auxiliary verbs (VAUX) immediately follow verbs in the participle (*-Te* form)[6]. Auxiliary verbs differ from their full verbal counter parts in both valence and meaning. They express aspect (e.g., *iru, oru, irassharu, iku, kuru, shimau*), benefactivization (e.g., *morau, itadaku, kureru, kudasaru*), or 'trying to do something' (e.g., *miru*). In the phrase *"opera o mite miru"* ('trying to see an opera'), for example, the first *miru* means 'to see (an opera)' as a full verb, and the second *miru* as auxiliary verb expresses that the act of seeing an opera is a trial and done for the first time.

Support verb (VS) *suru* has a mere supportive function, when it follows a verbal noun. There are a lexical humble form *itasu*, a respective form *nasaru*, and a potential form *dekiru* in the same category. When they follow verbal nouns (see Section 4.3) which are the logical predicates, they are classified as support verbs. They may also function as full verbs similarly to 'do' in English as a transitive verb. Compare the full verb *suru* in *"kore ni* <u>*shimashou*</u>*"* ('Let's go for this.') with the support verb *suru* in *"kaNkou* <u>*shimashou*</u>*"* ('Let's do some sightseeing.').

Particle verbs (PV) are plain *da* and polite *desu*. They have several different functions. On the one hand, they function as copula verbs, as in *"kaeri wa getsuyoubi* <u>*desu*</u>*"* ('The return will be on Monday.'). On the other hand, similarly to the auxiliary verbs, they follow predicates to add politeness and/or mark tense, for example, *"nani ga ii* <u>*deshou*</u> *ka ne"* ('Which one would be better?': polite and future); *"dame* <u>*datta*</u>*"* ('It didn't work.': perfect) [7].

4.3 Nouns

Unbound nouns are defined as words that can be modified by the demonstrative adjectives *kono, sono, ano* and/or form a phrase with case postpositions, such as

[6] A limited group of focus postpositions may appear between main verb and auxiliary; for example, *do nichi hasaNde* <u>*wa*</u> *imasu* ('The weekend <u>does</u> come inbetween' — focus on the action).

[7] For a different treatment in terms of annotation in order to account for the functional difference see the trees of Figure 5.47 and Figure 5.48 in Chapter 5.

ga and *o* (Rickmeyer 1995), (Tsujimura 1996).

Nominal bound forms (formal nouns, noun suffixes, prefixes) do not normally fulfill the above requirement, but when they follow or precede a phrase, the resulting phrase may occur as a nominal phrase either with *ga* or *o*, as in *"touchaku suru no ga ..."* ('the arrival').

Table 4.3 supplies a list of all nominal POS tags together with a short description and examples.

POS tag	**Description**	*Examples*
NN	Common noun	*hoteru, hikouki*
NF	Formal noun	*no, hou, koto, bun*
PRON	Personal pronoun	*watashi, anata, kare*
NAME	Proper noun	*doNjobaNni, buNkanohi*
NAMEper	person	*matsumoto, yoshikawa*
NAMEloc	location	*hanoofaa, doitsu, nihoN*
NAMEorg	organization	*rufutohaNza, jaru, ana*
NT	Temporal nouns	*kyou, kayoubi, gogo*
Ndem	Demonstrative noun	*sore, kochira, sochira*
Nwh	Interrogative noun	*dochira, naNji, dore*
CD	Cardinal numbers	*ichi, juuhachi, nijuu*
CDtime	with time unit	*nijuuji, juppuN*
CDdate	with date unit	*juuichigatsu, yokka*
CDU	with other unit	*itsukakaN, ichijikaN, futaheya*
UNIT	Unit	*maruku, biN, meetoru*
PreN	Noun prefix	*yaku, dai*
Nsf	Noun suffixes	*hatsu, ikou*
PNsf	Personal name suffix	*saN, sama*

Table 4.3: Nominal POS tags

Nouns are subclassified according to their syntactic and semantic features in the following way:

Common nouns (**NN**) are nouns without special semantic or syntactic features, for example,

deNsha, hoteru, joukeN, machi, nanika, okusaN, ryoukiN, sauna.

Verbal nouns (**VN**) are morphologically nouns. They function as the logical predicate that selects the subject and other complements, for example, *kakuniN,*

onegai, shuppatsu, yoyaku. They are often followed by a form of the support verb *suru* (see Section 4.2.): *hanoofaa o shuppatsu suru biN* ('the flight that departs from Hannover').

Regardless of the fact that verbal nouns also function as common nouns as in *shuppatsu wa naNji desu ka* ('What time is the departure?'), they are always tagged as **VN**.

Formal nouns (**NF**) are nouns that usually cannot stand on their own. They are mostly semantically empty and only used for nominalization (e.g., *no, koto, tokoro*), or other rhetorical devices (e.g., *naN, hou*). Some of the formal nouns can be modified by the demonstrative adjectives *kono, sono, ano* (e.g., *sono buN, sono hou, kono koro, kono keN, dono tame*), but they do not appear alone.

Proper nouns include nouns describing personal names (**NAMEper**), names of locations (**NAMEloc**), and names of organizations (**NAMEorg**). If a proper noun falls into neither of these three categories, it is tagged simply as **NAME**. For example,

NAMEper *abe, daisuke, guroosu*

NAMEloc *amerika, hanoofaa, hoteruruiizeNhoofu, raiNgawa, sutifutogyararii*

NAMEorg *ana, buNdesuriiga, zeNnikuu*

NAME *buNkanohi* ('Culture Day', name of a Japanese holiday), *kurisumasu* ('Christmas'), *doNjobaNni* ('Don Giovanni' — name of an opera).

Pronouns The POS tag **PRON** comprises personal pronouns. Most of the examples in the treebank refer to the first person, 'I' and 'we'. For example,

1st	**sing.**	*watashi, watakushi, atashi, atakushi, boku*
	pl.	*watashitachi, watakushitachi, watashidomo, watakushidomo, atashitachi, atakushitachi, atashidomo, atakushidomo, bokura, wareware*
2nd	**sing.**	*anata*
	pl.	*anatagata, anatatachi*
3rd	**sing.**	*kare, kanojo*
	pl.	*karera*

Temporal nouns (**Ntmp**) express temporal relations or units, but contain neither cardinal numbers nor suffixes (see also under "Cardinal numbers" below). For example,

mae, koNgo, haNnichi, mukashi, koNseiki, kayoubi, gogo

Demonstrative nouns [8] (**Ndem**) are deictic expressions. Some refer to things, for example,

kore ('this' close to the speaker),
sore ('that' close to the hearer),
are ('that' far from both),

and others refer to directions or places, for example,

kochira, kocchi, koko ('here' close to the speaker),
sochira, socchi, soko ('there' close to the hearer),
achira, acchi, asoko ('there' far from both).

Interrogative nouns (**Nwh**) are a special kind of demonstrative nouns that are used to form an interrogative statement. For example,

nani, dore, doko, izure, ikura, itsu.

See also the footnote about demonstatives.

[8] Demonstaratives are in the ko-so-a-do paradigm, which refers to the beginning sounds of the paradigmatic sets of nouns, adjectives and adverbs; the ones beginning with *ko-* refer to the speaker's sphere, whereas the ones with *so-* point to the hearer's sphere. Both forms can be used anaphorically as well. The ones beginning with *a-* belong to a sphere that is out of reach to speaker and hearer and words beginning with *do-* are interrogative words.

Cardinal numbers occur either by themselves (**CD**), or in connection with a suffix expressing time (**CDtime**), a suffix expressing date (**CDdate**), or another sort of unit (**CDU**). For example,

CD	*hachi, juuni, hyakugojuu*	8, 12, 150
CDtime	*juuichiji, juppuN*	'11 o'clock', '10 minutes'
CDdate	*saNgatsu, mikka*	'March', 'the third'
CDU	*ippoN, saNshurui*	'one [flight]', 'three kinds of'

Note that suffixes of numericals (e.g., *-ji, -fuN, -ka/nichi, -gatsu*) are sometimes separated from the preceding numeral and depicted in a single token (see Section 4.3.) depending upon the tokenizing conventions.

Units (**UNIT**) are nominals which function as units similar to suffixes, for example,

biN, kiro, meetoru, maruku, shitsu.

Noun affixes are prefixes and suffixes. Although the usage of prefixes (**PreN**) is rather restricted, suffixation with nominals is very common in Japanese. Nominal suffixes include ordinary nominal suffixes (**Nsf**) and personal names suffixes (**PNsf**) which usually attach to personal names as a polite address form. For example,

PreN	*dai, yoku, maru*
Nsf	*hatsu, chaku, hodo, dake, kurai, nado, tomo*
PNsf	*saN, sama*

4.4 Postpositions

Postpositions are bound forms which attach to different phrases. Their functions are as follows:

- mark them as nominative, accusative, or genitive attribute (case postpositions),
- organize information structure (focus postpositions),
- mark quotations (quotative postpositions),
- introduce additional semantic attributes (semantic postpositions),
- mark coordinations or items to be coordinated (conjunctive postpositions),

- connect clauses (subordinate clause postpositions),
- express speaker's attitude (sentence final postpositions).

Table 4.4 summarizes the classification of the postpositions.

POS tag	Description	Examples
P	Common postposition	*de, e, kara, made, ni, shika*
Pacc	Accusative case	*o*
Pnom	Nominative case	*ga*
Pgen	Genitive case	*no*
Pfoc	Focus postpositions	*wa, mo, koso, sae*
PQ	Quotative postpostions	*to, te, tte, toka*
Pcnj	Conjunctive postpostions	*ka, to, toka, ya*
PSSa	Subordinate clause postpositions (conditional/causal)	*to, nara, node, kara*
PSSb	(adversative)	*ga, keredomo, kedo*
PSSq	(interrogative)	*ka*
PSE	Sentence end postpositions	*ka, kana, mono, ne, yo*

Table 4.4: Postpositional POS tags

Case postpositions are the three postpositions *ga, o,* and *no*. Nominative *ga* (**Pnom**) usually marks the subject[9], accusative *o* (**Pacc**) the object, and genitive *no* (**Pgen**) describes possessive relations, however, genitive *no* can be replaced with *ga* in relative clauses.

Focus postpositions (**Pfoc**), most prominently *wa*, organize the information structure. The postposition *wa* usually marks the topic in contrast with the comment that usually follows and is focused on: *hoteru no namae wa amushutat-topaaku desu*, ('the name of the hotel (marked with *wa* = topic) is "Am Stadtpark" (focus)'). On the other hand, *demo, koso, mo,* and *sae*, usually set the focus on the marked phrase: *resutoraN mo arimasu*, ('there is (also) a restaurant.') Here the focus is on '(also) a restaurant' which is equivalent to the phrase which is marked with *mo* in the original.

[9]See also Section 6.2 about the notion of subject.

Quotative postpositions (**PQ**) are special in that they can follow every phrase. They delimit all kinds of utterances to the right, and mark them as the contents of verbs of verbal action, thinking, or writing. Besides *to*, the more colloquial forms, *te, tte, toka, naNte*, can be found in the treebank.

Semantic postpositions (**P**) carry some semantic information . Yet, their semantic interpretation often depends on the nature of the predicate or of the phrase they mark. The postposition *de* for example, describes:

means e.g., *"hayai biN de ikimashou ka"*, ('Shall we go with the early plane?'),

location e.g., *"roNdon de norikae"*, ('changing planes in London'),

time span e.g., *"ichinichi haN de owarimasu"*, ('finish in one and a half day').

Some of the semantic postpositions also mark complements. Especially *ni* is often seen as the dative case postposition (refer to Section 5.3 for a discussion about the postpositional *ni*). Examples of semantic postpositions are,

de, e, kamo, kara, made, ni, shika, to, yori, yorika.

Conjunctive Postpositions (**Pcnj**) are all particles which can mark coordinations of (usually noun) phrases, or phrases to be coordinated, most prominently,

to, toka, ka, ya.

Subordinate clause postpositions are classified into the following three groups:

PSSa marks conditional or causal subordinate clauses.
e.g., *kiNyoubi da to* ('if it is a Friday'),
kayoubi desu node ... ('since it is a Tuesday ...').

PSSb marks adversative subordinate clauses, such as *ga, keredomo, kedo.*
e.g., *aru mitai desu keredomo ...* ('seems to exist, but ...').

PSSq marks interrogative subordinate clauses.
e.g., *ii bijutsukaN aru ka ...* ('...whether there is a good art museum').

Sentence final postpositions (**PSE**) are frequently used in spoken Japanese, also in polite form. They usually appear at the end of the sentence following one of the final verbs or adjectives. They convey a kind of finiteness of the utterance. Their function is to convert a phrase into a question (e.g., *ka*) or to express the speaker's attitude and/or empathy (e.g., *ne, yo, na, kana, mono*).

4.5 Adjectives

There are three classes of adjectives in our Japanese treebank, the i-adjectives, the na-adjectives and the attributive adjectives. Table 4.5 summarizes the adjectival POS tags.

POS tag	Description	Examples
ADJifin	i-adjectives finite	*takai, chikai, hayakatta*
ADJiku	i-adjectives adverbial	*takaku, arigatou*
ADJite	i-adjectives participle	*takakute, chikakute*
ADJicnd	i-adjectives conditional	*takakereba, takakattara*
VADJi	Verb i-adjectives	*inai, kaeritai*
ADJ_n	na-adjectives	*beNri, hitsuyou*
ADJteki	na-adjectives, ending in -teki	*gutaiteki, jikaNteki*
VADJ_n	Verb na-adjectives	*ikesou, owarisou*
ADJ	Attributive adjectives	*taishita, iroNna*
ADJdem	Demonstrative adjectives	*kono, sono, ano*
ADJwh	Wh-adjectives	*dono, doNna*
ADJsf	Adjective suffix	*na*
PADJ	Particle adjectives	*mitai, rashii*

Table 4.5: Adjectival POS tags

i-adjectives are adjectives that inflect in the -i/-ku paradigm. Finite (fin) i-adjectives are distinguished from adverbial (ku), participle (te), and conditional (cnd) ones. Table 4.6 gives an overview of the inflection system of i-adjectives in the current treebank:

na-adjectives (**ADJ_n**) are non-inflecting words. They are usually attributed adnominally with the ADJsf *na*, and adverbially with the postposition *ni*. Some of them may function as predicates, for example,

Subclass	Inflectional ending	Examples
ADJifin	*-i* (Present)	*takai*
	-katta (Perfect)	*yokatta*
ADJiku	*-ku* (Adverbial)	*takaku*
	-U (Adverbial, historical)	*ohayou, arigatou*
ADJite	*-kute* (Participle)	*takakute*
ADJicnd	*-kereba* (Conditional)	*takakereba*
	-kattara (Conditional perfect)	*takakattara*

Table 4.6: Inflection system of i-adjectives

watashi mo biiru ga suki nanode ... ('...since I also like beer').

Among this group there is a distinct group of adjectives that are ending in *-teki* (**ADJteki**). They are attributive with the suffix *na*

gutaiteki na sukejuuru ('an explicit schedule'),

but may also function as adverbials with the postpositional *ni*

watashi wa kojiNteki ni wa opera ga suki ('personally I like operas').

Examples of na-adjectives are,

ADJ_n *daijoubu, kekkou, beNri, suki, muri, dame*
ADJteki *jikaNteki, kojiNteki, gutaiteki*

Attributive adjectives are only used as attributes (**ADJ**). Among them, there are two distinct groups, the interrogative adjectives (**ADJwh**) and the demonstrative adjectives (**ADJdem**)[10]. For example,

ADJ *taishita, iroNna*
ADJwh *dono, doNna*
ADJdem *kono, sono, ano.*

Other adjectival POS tags Derivational adjectives with a verbal stem that still maintain the predicate-argument structure of their matrix verb are called **VADJi** or **VADJ_n** depending on the continuation classes. For example, negation of verbs can be expressed by derivation with the suffix-adjective *-(a)nai*, and the voluntative form of verbs can be expressed by derivation with the suffix-adjective *-tai*. The continuation class of *-(a)nai* or *-tai* is identical to that of i-adjectives, therefore the derived forms are called **VADJi**, for instance,

[10]See also the footnote for "demonstrative nouns" in Section 4.3.

POS tag	Description	Examples
ADV	Adverbials in general	*chotto, mou, mata, daitai, dekireba, choudo, moshi, ichiou, zehi*
ADVdem	Demonstrative adverb	*kou, sou*
ADVdgr	Degree adverb	*ichibaN, sukoshi, amari, sugoku, kanari, zeNzeN, zuibuN*
ADVtmp	Temporal adverb	*sassoku, mazu*
ADVwh	Wh-adverb	*dou, doushite, ikaga*
PADV	Particle adverb	*youni, fuuni, shidai, nagara, hodo*

Table 4.7: Adverbial POS tags

iku ('go'-present) → *ikanai* ('not to go'),
→ *ikitai* ('want to go'),

kimeru ('decide'-present) → *kimenai* ('to not decide'),
→ *kimetai* ('want to decide').

Similarly, there is a modal suffix-adjective *-sou* expressing 'appearance', whose continuation class is identical to that of the na-adjectives. Therefore the derived forms are called **VADJ_n**, for instance,

tobi ('fly'-base) → *tobisou* ('seems to fly'),
owari ('end'-base) → *owarisou* ('seems to end').

Adjectival bound forms that appear as separate tokens in the treebank are the adjective suffix (**ADJsf**) *na* and the particle adjectives (**PADJ**), for example, *rashii, mitai.*

4.6 Adverbs

As a summary of adverbials, Table 4.7 shows the POS tags for the different groups of adverbs. Adverbs are words which primarily modify verbs, adjectives, or other adverbs. Some adverbs may be marked with *to* or *ni*. There are some semantically/functionally distinct groups of adverbs that are subclassified, such as demonstrative adverbs of the ko-so-a-do paradigm[11] (**ADVdem**), interrogative (**ADVwh**), and temporal adverbs (**ADVtmp**). All the other adverbs are tagged as **ADV**. Particle adverbs (**PADV**) are bound adverbial forms that typically attach to verb phrases and mark their adverbial dependency, for example, *onedaN wa* [$_{VP}$*dou iu*] *fuuni natte imasu ka* ('What are the prices like?').

[11]See the footnote for "demonstrative nouns" in Section 4.3.

4.7 Others

Conjunctions (**CNJ**) are a morphologically heterogeneous group of words, perhaps because they are derived from different etymological origins. However, they are common in being used to connect clauses and phrases on various levels. They typically appear in the beginning of sentences or between coordinated items, for example,

> *dewa, soredewa, sorekara, soshitara, soshite, soretomo, moshikuwa, aruiwa, ato.*

Interjections (**ITJ**) are words that generally form utterances by themselves. They are usually isolated from the sentence structure, for example,

> *a, aa, ano, ara, are, e, eeto, hai, iya, iza, jaa, uN, yaa.*

Punctuations are tagged as they appear in the terminal string in the transcription. The three punctuation symbols that have appeared in the VERBMOBIL-II transcriptions are ".", "," and "?".

Chapter 5

Node labels

Node labels are spanning over the strings of token-tag pairs to represent the constituency. We have generally chosen to employ classical terms for the name of the node labels (e.g., NP, VP, AP, PP) to maintain theory-neutral descriptions, thus ensuring reusability of the treebank. This chapter explains node labels. Tree diagrams presented in this stylebook are in NEGRA format (Plaehn 1998).

5.1 Errors, Repetitions, Interjections

Spontaneous speech data contains much linguistic noise. False starts, speech errors, and repetitions are transcribed as such. Those fragments are regarded by the transcriber as out of the intended speech, and should be annotated as such also in the treebank. A node label "**err**" is used in those cases as in Figure 5.1. Some

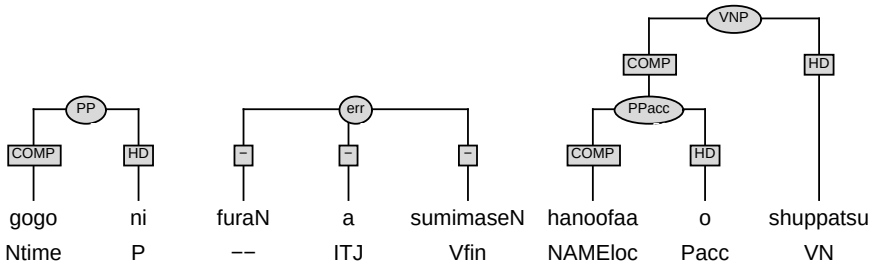

Figure 5.1: Speech Errors

tokens indicate fragments of words and are consequently unidentified in terms of POS tags. They are left without any POS tag, for example *furaN* in Figure 5.1. In this example, the speaker was probably uttering "Frankfurt", interrupted himself

by an interjection in the middle of the word, quickly apologized, and started the correct utterance with "Hannover". Three tokens are transcribed as speech errors. Thus, the node **err** is assigned. False starts and repetitions in the transcription are also treated in the same way as **err**.

Tokens which are tagged as **ITJ** (interjection) are mostly isolated from the node of the trees, for example, *ano* appearing in Figure 5.2. There are, however, several interjectional expressions with a fixed sequence of words very frequently occurring in the treebank, which are described as a node called **ITJ**, for example, *eeto desu ne* ("err, isn't it?") shown in the Figure 5.2.

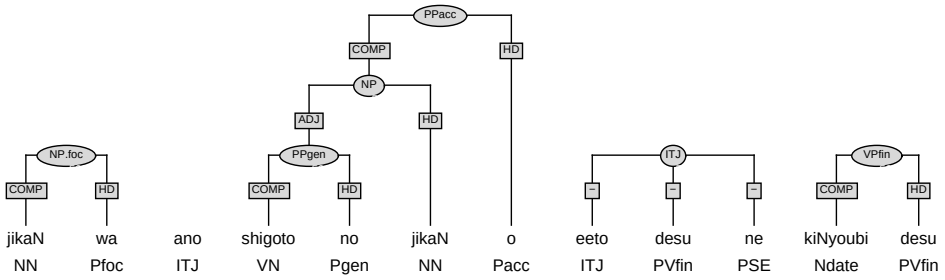

Figure 5.2: Interjections

5.2 Noun Phrases

Noun phrases occur very frequently. There are also many varieties of constructions in noun phrases, from simple naming to complex combinations of nominalized predicate-argument structures.

5.2.1 Name – person and location –

Typical proper noun expressions appearing in the VERBMOBIL-II dialog domain are names of persons and locations.

1. **NPper** stands for the name of a person. Family names precede the first name in Japanese (Figure 5.3). People often address each other with their family name followed by a person suffix, *saN* or *sama*[1]. The suffix may but rarely follows the full name of a person.

[1]Cases of personification may be found. For example, *ana saN*, a company name ANA (All Nippon Airway) is accompanied by the suffix *saN*. This instance is annotated as **NP** with no specific subcategory.

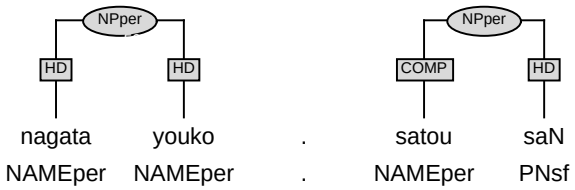

Figure 5.3: Name of Person.

2. **NPloc** stands for the name of a location. Proper nouns indicating geographical locations, hotels, buildings, companies, and so on, may sometimes be sequences of several tokens depending on how the complex name is formed, or depending on the tokenization (Figure 5.4).

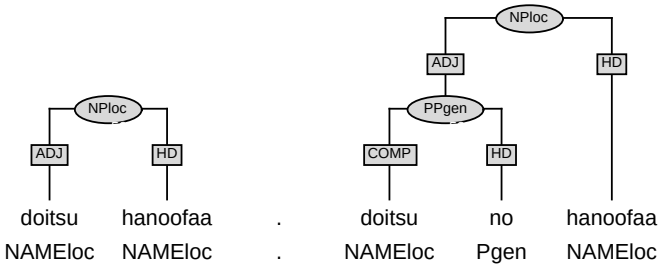

Figure 5.4: Name of locations.

There are empirical motivations to distinguish personal, locative, and temporal nouns, although these features are usually not regarded as syntactic features. Named entity recognition should be made much easier if these distictions are made. Temporal and locative expression have their own local grammar, and to apply for an appropriate local grammar, we should know which subcategories the noun phrase belong to.

5.2.2 Temporal – date and time –

Time and date expressions are typically noun phrases. They are annotated as **NPtmp** to show some semantic selection. They are very frequent and convey crucial information in the conversation domain of the treebank, that is, travel arrangement. A canonical date expression consists of a sequence of year, month, day, and days of the week. A canonical time expression consists of (morning or afternoon,) hour, and minutes (Figure 5.5). The concatenation of nominal

phrases with possessive *no* **Pgen** is also very commonly found in date expressions (Figure 5.6).

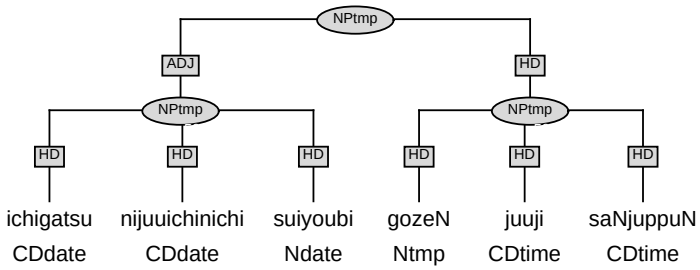

Figure 5.5: Date and Time.

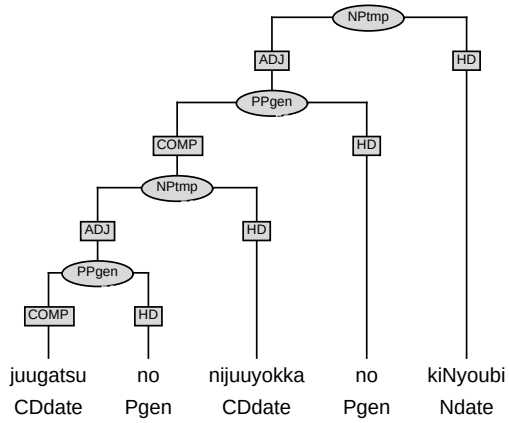

Figure 5.6: Date with *no*

Note that an **NP** whose head is a temporal noun is not always a **NPtmp**. Figure 5.7 shows an example in which the noun phrase is headed by a time expression but refers to a specific aircraft instead of a particular time.

5.2.3 Modified NP

Modified by nouns

A sequence of two or more nouns can form a noun phrase. Noun-noun sequences are common among nominal compoundings of Sino-Japanese words. Following the general characteristics of the Japanese language, nominal compounds are head final, that is, the final noun is responsible for the feature of the whole noun phrase (Figure 5.8).

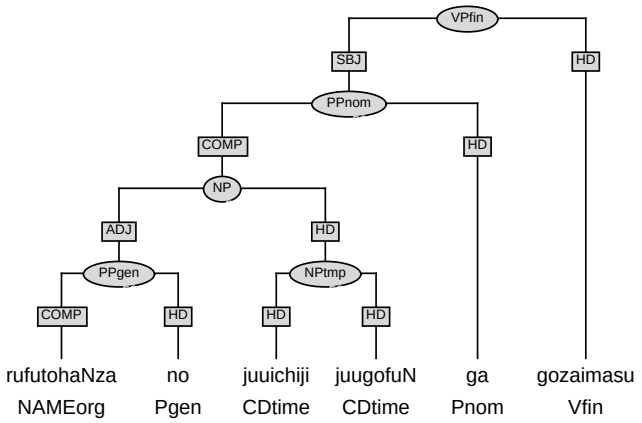

Figure 5.7: Temporal NP?

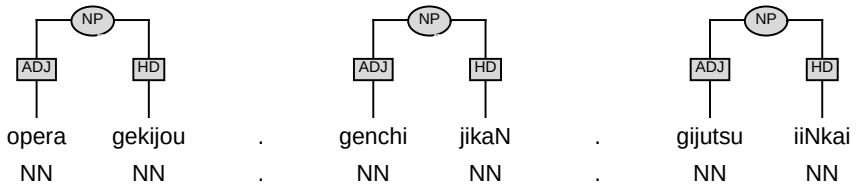

Figure 5.8: Noun-Noun

As mentioned above, a date expression followed by a time expression is frequently found in the conversations on travel arrangements. They often take the form of NP-NP sequence as in Figure 5.5, and the whole constituent is described as a temporal noun phrase.

Sequences of several nouns and NPs might also be coordinations or listing expressions, which will be mentioned in Section 5.2.5.

Modified by Adjectives

Nouns and noun phrases are modified by various preceding adjuncts. Adjective phrases typically modify noun phrases to make more specific noun phrases. The head noun phrase as well as the modifying adjective phrase can either be simple or complex (Figure 5.9 and 5.10). An AP may be a structured predicate as we will see in Section 5.4, and such an AP can also modify a noun (phrase) following it. A series of demonstrative expressions, *kono*, *sono*, *ano*, are also adjectivals in Japanese because they modify nominals to make more specific noun phrase, but are not obligatory as determiners in European languages.

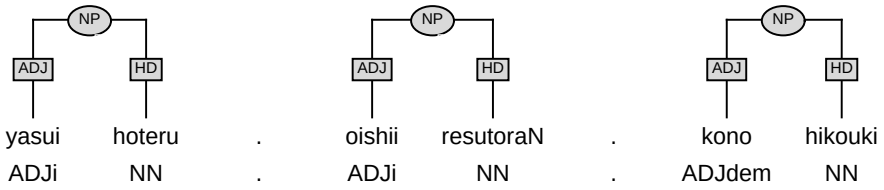

Figure 5.9: Nouns modified by adjectives.

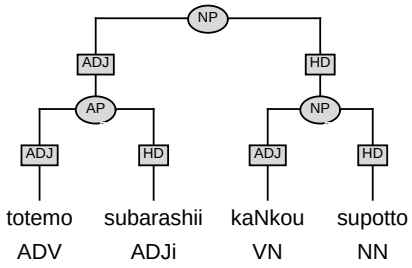

Figure 5.10: NP modified by an AP.

Modified by a PP

Nouns and noun phrases are sometimes modified by postpositional phrases to make more specific noun phrases (Figure 5.11).

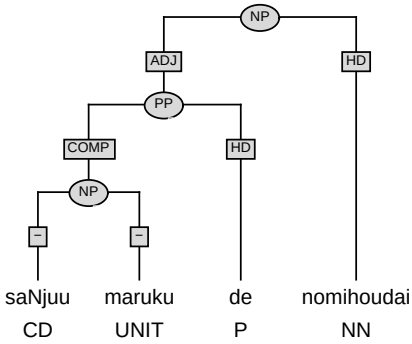

Figure 5.11: NP modified by PP

The postpositional phrase, whose righthand side constituent is the genitive case postposition *no*, functionally indicates the possessor of the following head noun phrase (Figure 5.12). The possessive construction in Japanese, however, is used differently, and in a broader context. We will talk about it in Section 5.3.

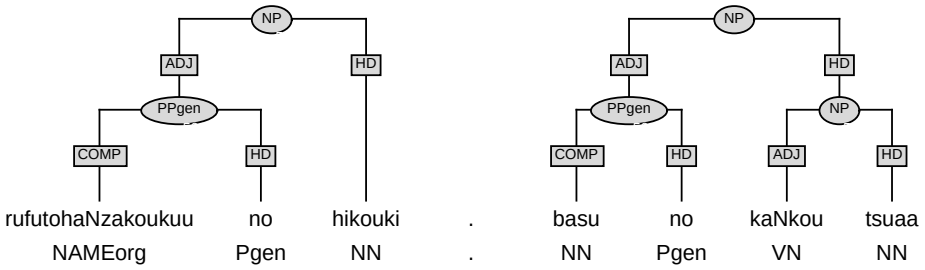

Figure 5.12: NPs modified by genitive PPs

Modified by VP

It is also commonly found that a finite verb phrase modifies an immediately following noun or noun phrase. This is seen as a relative clause, even though there is no explicit relative pronoun, in Japanese (Figure 5.13).

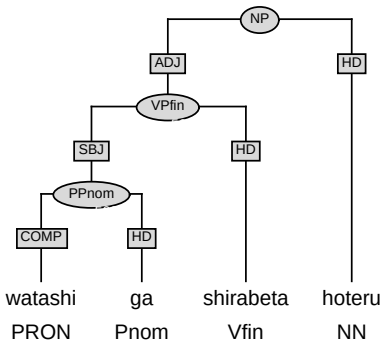

Figure 5.13: NP modified by a VP

5.2.4 Complemented NP

Formal noun phrases

There is a class of formal nouns **NF**, for example *koto* and *no*, whose semantic content is empty and whose function is to form a nominal structure together with other expression. In Japanese, this is a very common way of nominalizing phrases of other categories. Though English has other common means of nominalization, the Japanese formal noun may be seen as similar to the English "fact", which is almost always complemented with the following that-clause. The formal nouns

NF are typically complemented with a relative finite verb phrase (Figure 5.14), an adjective, or an adjective phrase (Figure 5.15).

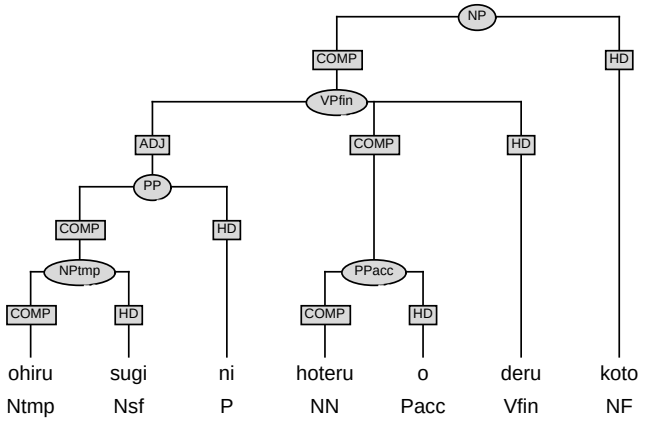

Figure 5.14: VP and formal noun

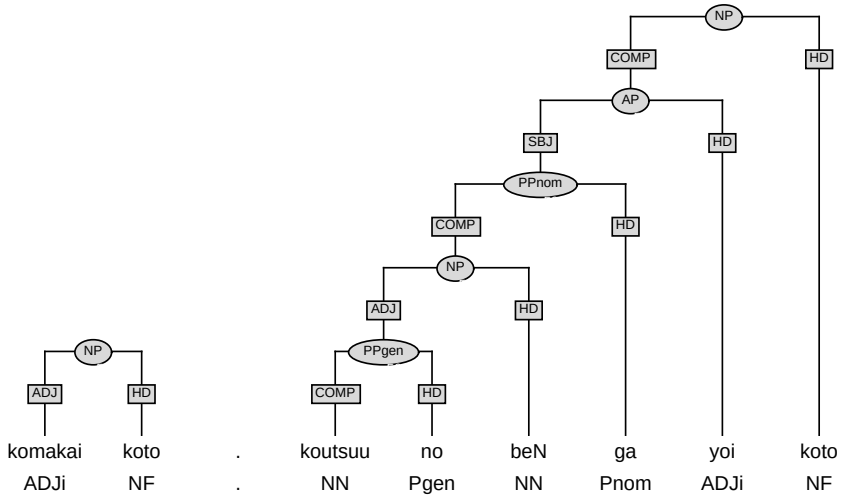

Figure 5.15: AP and formal noun

Formal nouns are normally complemented with one of the predicative phrases as we have seen. The complementation can also be satisfied with a demonstrative reference expression such as *kono, sono, ano*, or one of the genitive postpositional phrases (Figure 5.16).

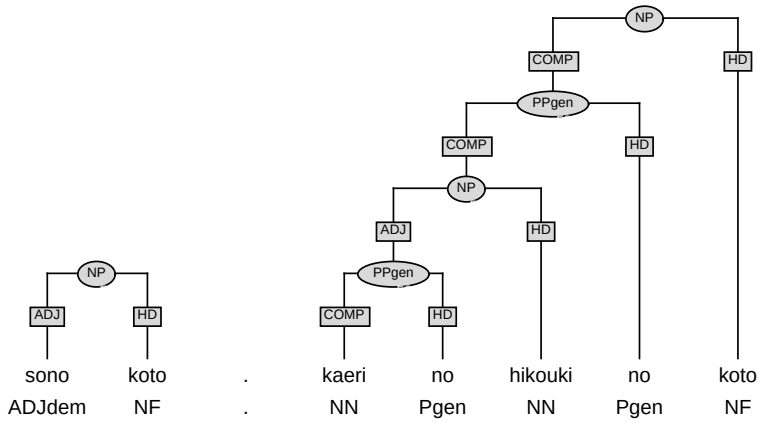

Figure 5.16: Reference expressions and formal noun

There is also a formal noun phrase headed by *nano*[2], which is complemented with a noun phrase, or an adjective phrase. The token *nano* constitutes an independent token in the Verbmobil transcription, however, the adjective phrases preceding *nano* often select, as its right continuation class, adjective suffix *na* (See AP in Figure 5.17 left). Therefore, one could also have considered this *nano* as an adjective suffix *na* followed by a formal noun *no* just as mentioned above. However, since *nano* is a token and the resultant phrase is nominal, the best possible description of *nano* is as a formal noun (Figure 5.17).

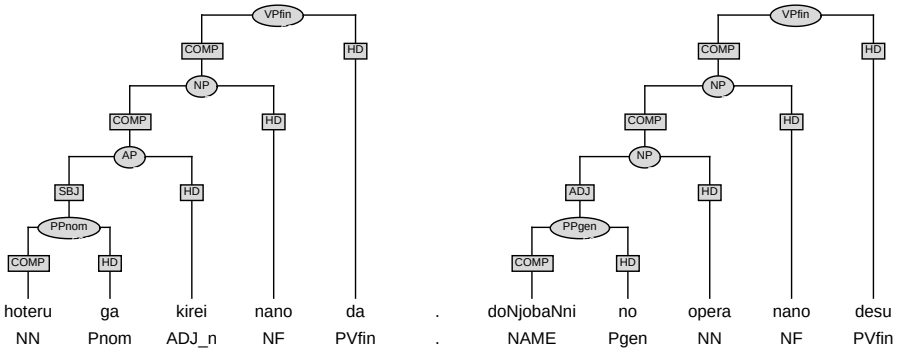

Figure 5.17: Formal noun *nano*.

[2]Also, there are many cases where it is transcribed as *naN* with the final vowel being dropped.

Special-purpose suffixes

The term "suffix" is a term in morphology. One might think suffix should not be an independent token, however, numerous bound morphemes are tokenized in the transcriptions due to the morphological oriented characteristics of the language mentioned in Chapter 2. When this is the case, a bound morpheme dependent on its preceding constituent may be described as a complement followed by a head, conforming to the one of the general schemata mentioned in Chapter 6.

The suffixes for the personal name, *saN* and *sama*, have been already exemplified above in Figure 5.3. Suffixes, such as *hatsu* ('departure'), *chaku* ('arrival'), *keiyu* ('via.'), and *iki* ('bound for') are particularly frequent in the travel arrangement conversation domain. Among others, *hatsu* and *chaku* are especially frequently preceded by a temporal expression (date, time), and/or a name of the location (Figure 5.18)[3]. In either of the cases, the suffix is bound to the immediately adjacent expression.

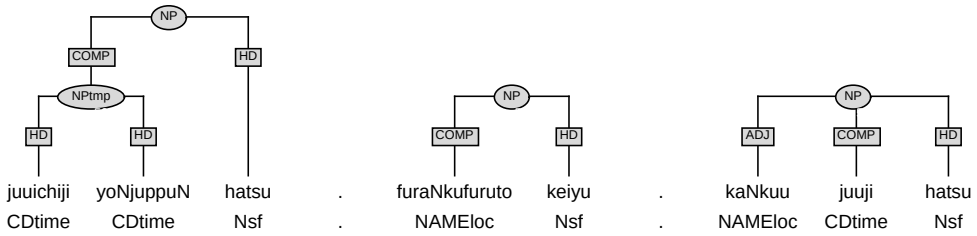

Figure 5.18: Noun and suffix

Headed by VN

There is a class of verbal nouns **VN** whose semantic content is a predicate. Verbal nouns may appear in the same contexts as common nouns, but they have specific characteristics perhaps because of their semantic nature. The left contexts of verbal nouns are arguments of that predicate[4]. The support verb, *suru* often follows verbal nouns (Figure 5.19 left).

The phrase headed by the verbal noun and accompanied by some of its argument phrases is annotated as a **VNP**, which is named ambiguously because

[3]The order of time and location expressions does not matter. Both time-location and location-time are, in fact, equally frequent.

[4]Some derived nominals are also similar in this respect, but they are not classified into **VN** because their right continuation class is not support verbs, for example *chikaku* ('neighboring'), is a derivation from adjective *chikai*, *go-zonji* ('knowing'-honorific) is a derivation from verb *zonjiru* affixed with honorific noun prefix *go*.

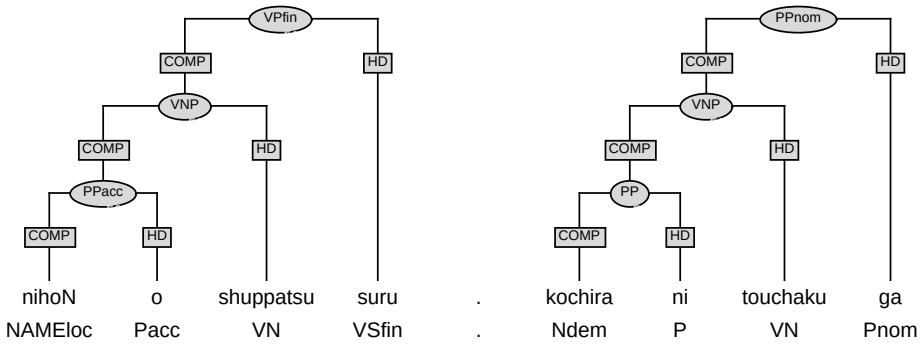

Figure 5.19: Verbal noun phrases

it may be either a phrase with verbal nature or a phrase with nominal nature depending on its following context (Figure 5.19).

5.2.5 Coordinated NP

There are coordinated structures with and without an explicit coordination marker. Let us first look at the cases without a coordination marker. Noun-noun sequences that are listed items are annotated so that the items may indicate the same edges (Figure 5.20). So are NP-NP sequences that are coordinated (Figure 5.21).

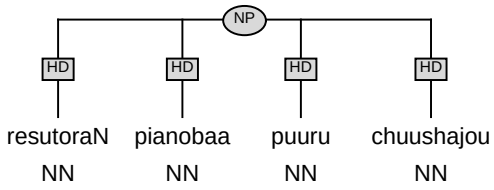

Figure 5.20: Listing

Among the coordination structures with a marker, two different types of coordination structures can be recognized (Figure 5.22). The figure on the left, as usual explanations of the coordination, a coordinating conjunction appears between the coordinated items in order to represent equal status of those items. On the other hand, the figure on the right, similarly to other common constructions in the Japanese language, a "postpositional morpheme" appears at the end of each coordinated item. Both types of coordination are equally frequent in utterances in the real world. Under the term "coordination" in the research, mostly the for-

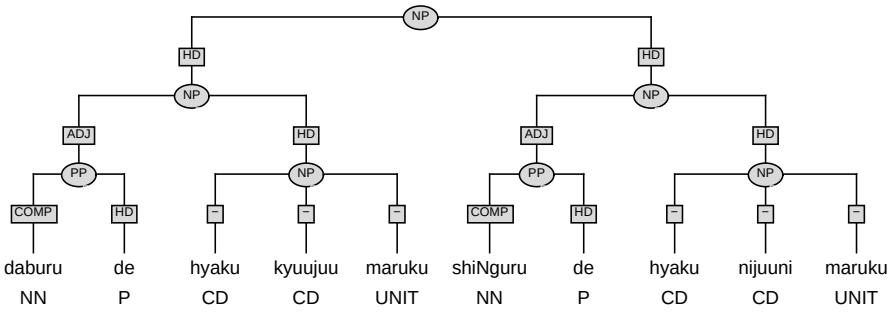

Figure 5.21: Coordination

mer construction is mentioned. We do not know yet, in the current status of our treebank, which is the basic or the derived one.

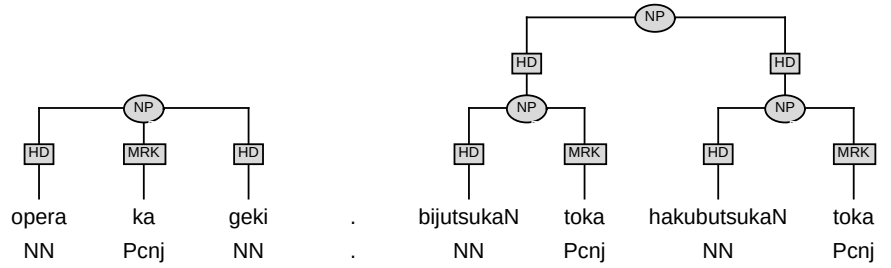

Figure 5.22: Two types of coordination

In theories, coordinated items belong to the same syntactic category, which is the case most of the time. In the real world, however, there are cases of coordinations among different major categories, for example, a finite verb phrase and a noun phrase, for example,

$[_{VP}$ *hikouki desu* $]$ *toka* $[_{NP}$ *hoteru* $]$ *wa toremashita ka ?.*

5.3 Postpositional Phrases

5.3.1 Case PP

Nominative - *ga* - **and Accusative** - *o* -

Postpositions significantly encode the grammatical functions of the phrase. The nominative and the accusative cases are most straightforwardly associated with

the predicate[5]. The nominative case marker *ga* indicates that the phrase is the subject. The accusative case marker *o* indicates that the phrase is the direct object. These are prominent complements of the predicate (Figure 5.23).

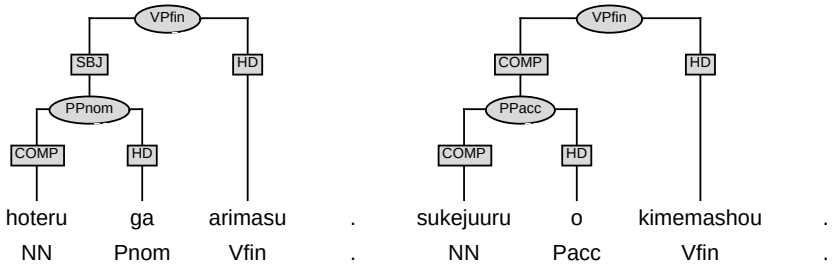

Figure 5.23: Case PP

Dative - *ni* -?

One might think that the dative case could also be straightforward, but this is not the case. There are di-transitive verbs, such as *okuru* ('send'), *ageru* ('give'), which normally select a *ga* marked subject phrase, an *o* marked direct object phrase, and a *ni* marked indirect object phrase. The postpositional *ni*, however, also indicates that the phrase has various semantic functions such as goal, direction, location, time and so on. We found that it is extremely hard to draw the line among the instances of postpositional *ni*, between the dative PPs and the PPs of temporal, locative, and directional senses. Therefore, the notion of the dative case has not been introduced in the Japanese treebank for the sake of consistency[6].

Genitive - *no* -

As an example shown in Figure 5.12 NP modified by genitive PP, the genitive case marker *no* indicates the phrase is possessive. Looking at their contents, however, one might find various different usages of the genitive phrases headed by *no*. Similarly to English possessive expression *'s* and *of*, it is possible to construct various sorts of modifications by using the Japanese postposition *no*. For examples:

- To refer to the smaller part of the larger part, "the large part *no* the small part" is a common expression (e.g., Figure 5.4 *"doitsu no hanoofaa"*).

[5]Mapping between syntactic and logical semantic levels is beyond the scope of the stylebook. For a discussion about exceptional case marking see, e.g., (Tsujimura 1996).

[6]This does not mean we are denying the notion of the dative case, but we wait until the other parts of the treebank develop for us to be able to see clearer pictures of postpositional *ni*.

- For nominalized predicates, an argument that would otherwise be the subject, the object, and so forth, may appear as a phrase marked with *no*. The classic example of the nominalized phrases of English "Enemy's destruction of the city" could be translated in a way into Japanese as follows:

teki	*no*	*sono*	*machi*	*no*	*hakai.*
(enemy	's	the	city	's	destruction)

- In relative clause, an argument that would otherwise be the *ga* marked subject often appears as a *no* marked phrase, probably because of present of the nominal form following the relative clause (Figure 5.24)[7].

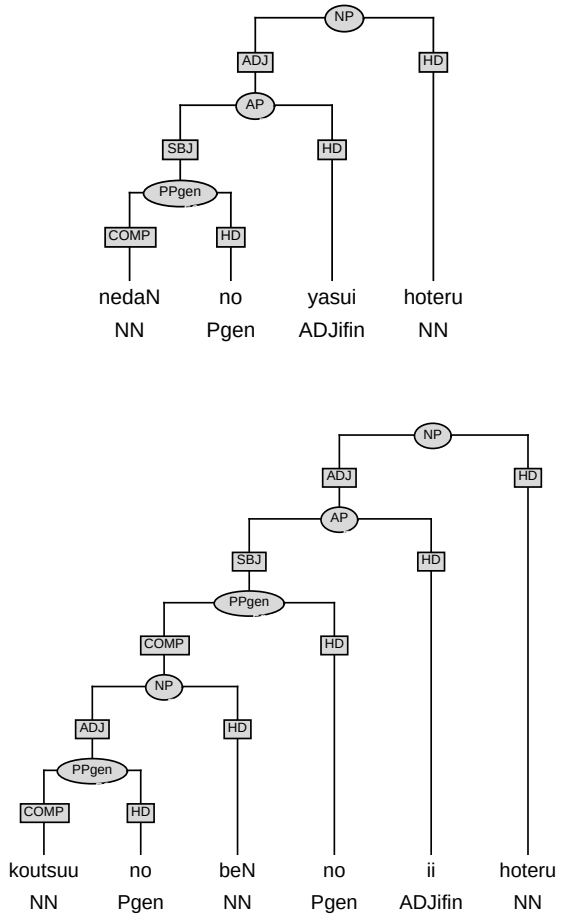

Figure 5.24: *no* marked subject

[7] *nedaN* **ga** *yasui hoteru* is equally grammatical. See also Figure 5.35.

5.3.2 Focus PP – *wa, mo, etc.* –

Focus[8] postpositions **Pfoc** mark phrases with some emphasis. The typical examples are *wa* and *mo*. Not as often as after NP or PP though, they appear after other categories.

NP.foc and PP.foc

A **Pfoc** often occurs after a **PP** headed by a semantic postposition, or a bare **NP** (Figure 5.25). In the latter case, none of the case markers or the semantic relation

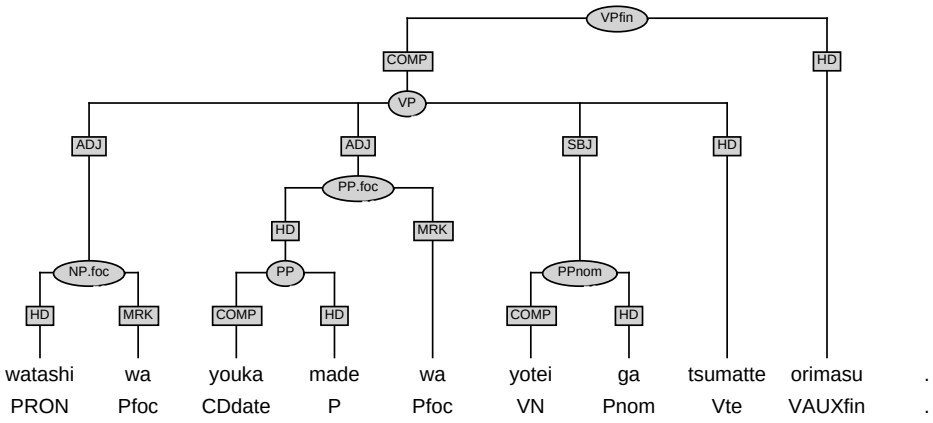

Figure 5.25: Foci on NP and PP

indicators is present, thus we call such a phrase **NP.foc**. More than one phrases can be focused as in the Figure 5.25 Foci on NP and PP.

VP.foc, AP.foc, and ADVP.foc

Verb phrases in participle, whose head verb ends in *-te* form, are often focused on with one of the focus postpositions (Figure 5.26).

Similarly, adverbial phrases are often focused on as **ADVP.foc**, and sometimes so are adjective phrases **AP.foc** (Figure 5.27).

5.3.3 Quotative PP

There is a class of postpositions **PQ**, which indicates that the preceding part of it is quoted. The quotative **PP** is usually a complement of a verb of saying *iu*, of

[8]The term "Topic" is also common in linguistic literatures.

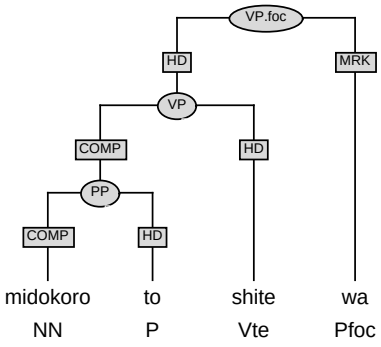

Figure 5.26: Focus on VP

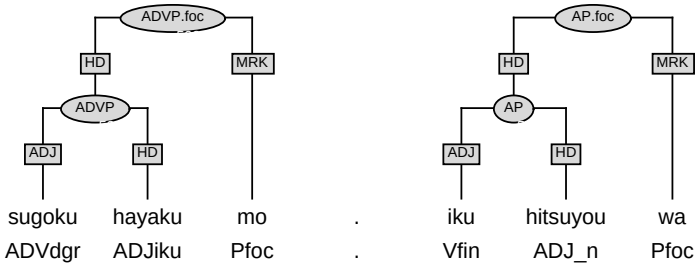

Figure 5.27: Focus on AP and ADVP

belief *omou*, of thought *kaNgaeru*, and so forth. Since an explicit left delimiter of the quoted part does not exist in Japanese, as one of the PP attachment problem in general, the scope of the quotation could be ambiguous in some contexts, and is to be determined by the hearer's prosodic and other extralinguistic knowledge, in that particular context and situation. For example,

sono	*hito*	*ga*	*isha*	*da*	*to*	*itta*
(the	person	**Pnom**	doctor	is	PQ	said)

can have different readings; (a) someone said "the person is the doctor", or (b) the person said "is a doctor", depending upon the different PP attachment.

Considering the nature of the kind of the utterances, the lefthand side constituent of the quotation could obviously be any possible linguistic object, from a complete well-formed sentence (Figure 5.28) to a piece of fragment of an utterance, (Figure 5.29)[9]. Note that this is one of the few patterns where S appears as

[9]Our statistics shows that adjective phrases, finite verb phrases, noun phrases, and sentences, are frequent ones. Varieties of sorts of single words are also found in the left of the **PQ**.

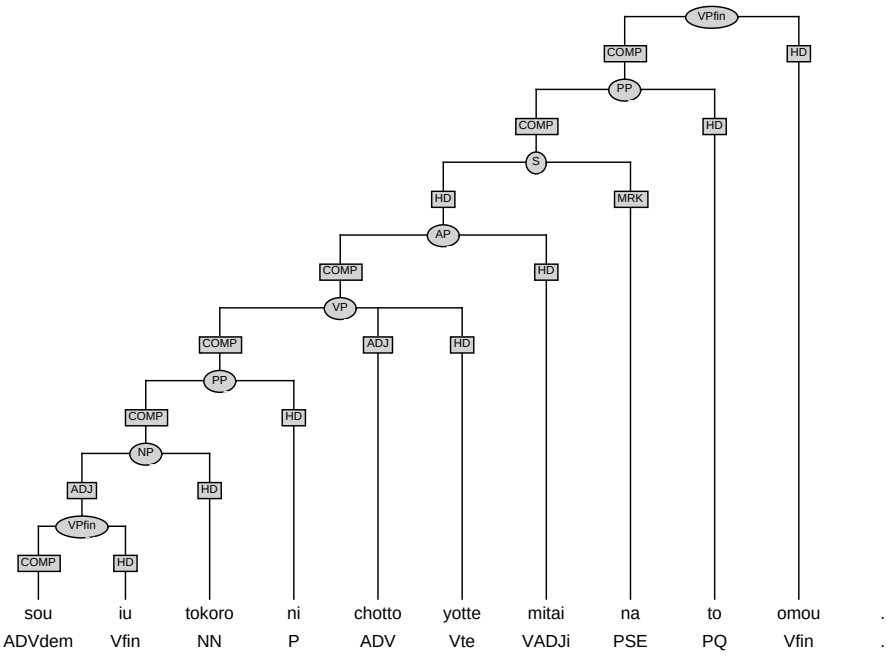

Figure 5.28: S quote

a non-root category.

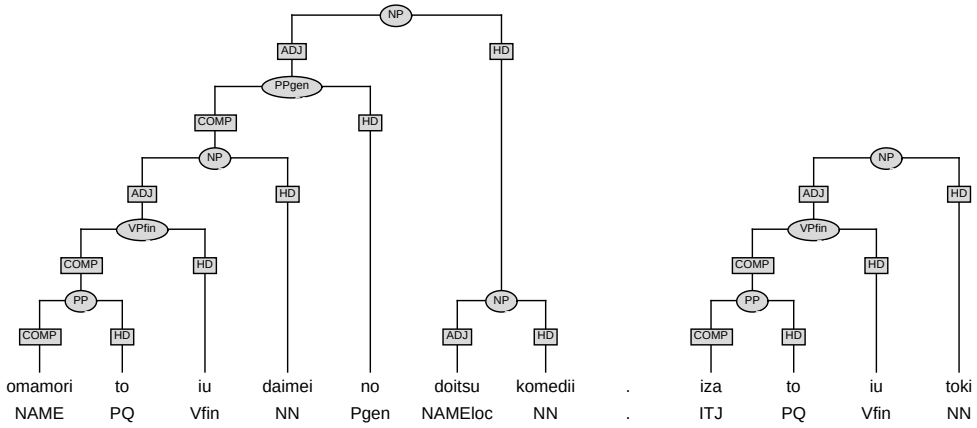

Figure 5.29: NP quote, ITJ quote

5.3.4 Other PP

There are **PP**s that are headed by one of many semantically significant postpositions **P** expressing time, location, reciprocal relation, direction, source, goal, instrument, and so on[10]. These **PP**s are either one of the complements of the predicate according to the subcategorization specifications of the relevant predicate, or an adjunct modifying some part of the tree. For examples, *youka made wa* is an adjunct of *tsumatte* in Figure 5.25, *midokoro to* is a complement of *shite* in Figure 5.26, *sou iu tokoro ni* is a complement of *yotte* in Figure 5.28. Those **PP**s that are **NP** adjuncts are already mentioned above in Section 5.2.3.

5.3.5 Remarks on PP

Just as those languages with prepositional phrases have potential PP attachment ambiguities rightward, there are also PP attachment ambiguities in Japanese but leftward, as already mentioned in Section 5.3.3. A canonical Japanese clause consists of a sequence of several PPs followed by a verb or an adjective. A preceding PP can potentially be a modifier of the immediately following NP in the following PP, or one of the constituents of the head predicate. Naturally, there are more possibilities in the more complex sentences. They should not be a problem for the human annotators most of the time, however, if one come across a "real" ambiguity with no particular preference, attachment to the higher possible node is recommended. See also Section 3.3 and 5.6.

The case postpositions and the semantic postpositions are mostly subcategorized for a noun phrase complement. In some cases, **PP** may be a complement of another **P**, for example,

$[_{PP}$ *Tokyo kara*] *ni shimasu.*

Some **P**s and **Pfoc** are subcategorized for other phrases than a noun phrase complement, for example,

$[_{NP}$ *Tokyo*] *shika arimaseN.*

$[_{VP}$ *Tokyo e iku*] *shika nai desu.*

In a speech dialogue, postpositions are often absent either being dropped by the speaker, or having not been picked up by the hearer or the recording device (Figure 5.30). Also, case postpositions *ga*, *o*, on one hand, and focus postpositions *wa*, *mo*, and so on, on the other hand, distribute themselves complementary each

[10]See also Section ?? and Appendix A.

other. Therefore, focused noun phrases with one of the **Pfoc** is regularly missing the nominative or accusative case marking postposition regardless of their grammatical function (Figure 5.31). Yet, it is more natural as a Japanese sentence that one of the phrases in the main clause appears with one of the focus postpositions. Therefore, in many cases, the annotator must impose an appropriate PP structure on an **NP.foc** and a bare **NP** to annotate their grammatical functions. For example, the subject of the verb phrase is supposedly present with nominative case marker *ga* in principle, but in fact in many cases, the subject is a bare **NP** or **NP.foc**, in either case with no nominative case marker *ga* being present (Figure 5.31). After all, formal signs should prompt a particular analysis to the hearer, but in reality they are not guaranteed to be there.

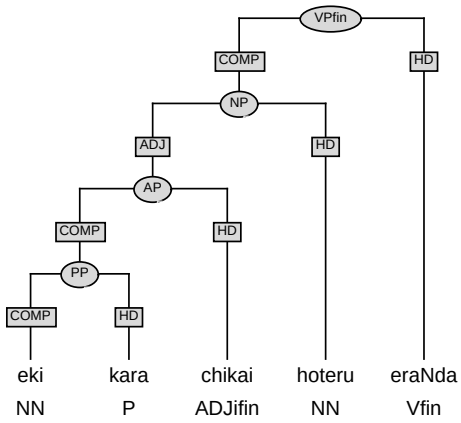

Figure 5.30: Bare NP object

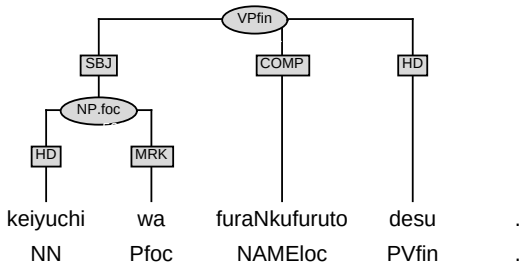

Figure 5.31: Focus NP subject

5.4 Adjective Phrases

Most adjectives have both in attributive and predicative uses. Few of them find only one of the usages.

5.4.1 Attributive

An adjective phrase in attributive use precedes an noun phrase. This is annotated as a modifying adjunct of the following head noun phrase as already shown in Section 5.2.3, and is repeated here as Figure 5.32.

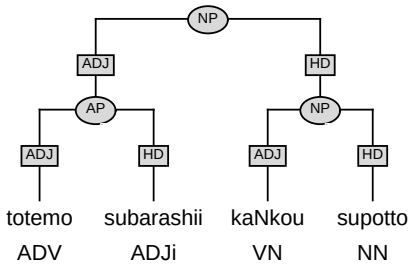

Figure 5.32: Attributive AP (*-i* ending)

Not only *-i* ending adjectives, but also adjectives concatenating to adjective suffix *na*, which also includes *-teki* ending adjectives (See Section 4.5), can also be attributive to their following nominals with the ending suffix *na* present (Figure 5.33).

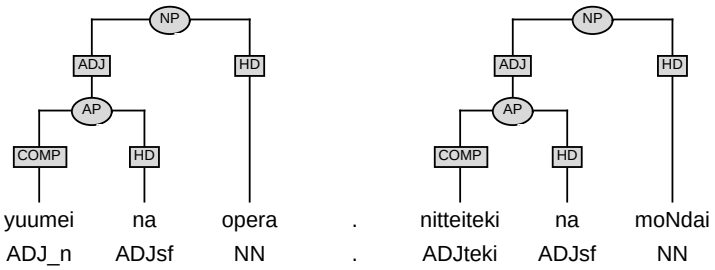

Figure 5.33: Attributive AP (*-na*)

5.4.2 Predicative

An adjective of predicative use appears in the final position in the AP in which the adjective is the head. The argument structure of the adjective is assumed to be represented in the lexicon of the adjective. The whole AP may constitute the head of the main clause (Figure 5.34), or may possibly be an attributive modifier of the following head NP in turn (Figure 5.35).

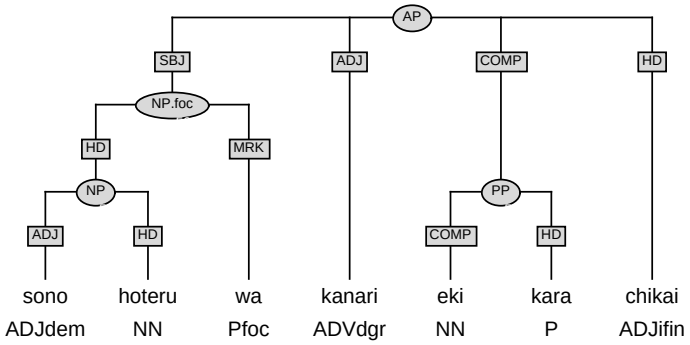

Figure 5.34: Predicative AP

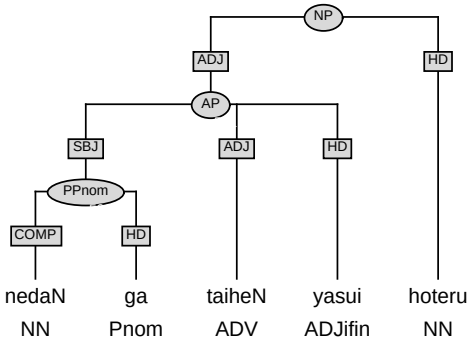

Figure 5.35: Predicative AP (modifier)

5.4.3 VADJ, PADJ – *-tai, -nai, rashii, etc.* –

The negative ending form of the verbs is *-(a)nai* ('not doing'), voluntative ending form is *-tai* ('wanting to do'), one of the modal forms ending *-sou* ('likely to do'), and so on. Verbs together with these endings are tokenized as single tokens for

morphological reasons. Their left adjacent contexts are identical to those of their leftmost matrix verb, and their right adjacent contexts are identical to those of i-adjectives or na-adjectives. See also Section 4.5. Thus, we call these **VADJ** (Figure 5.36).

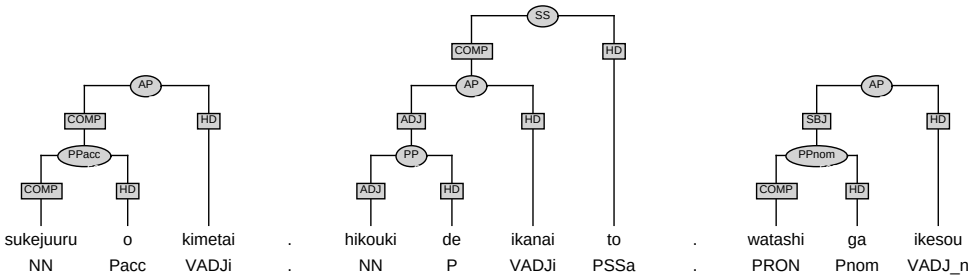

Figure 5.36: AP with VADJ

Modal particles such as *rashii, mitai* on the other hand, may appear after verbs in finite *-Ru/Ta* ending forms, adjectives ending in *-i*, and sometimes noun phrases. These modal particles are independently tokenized, and we call them particle adjectives **PADJ**. Their left adjacent contexts are discrete finite ending form possibly concluding the sentence. Their right adjacent contexts are corresponding to other adjective phrases. Phrases headed by particle adjectives, **PADJ**, are also **AP**s (Figure 5.37).

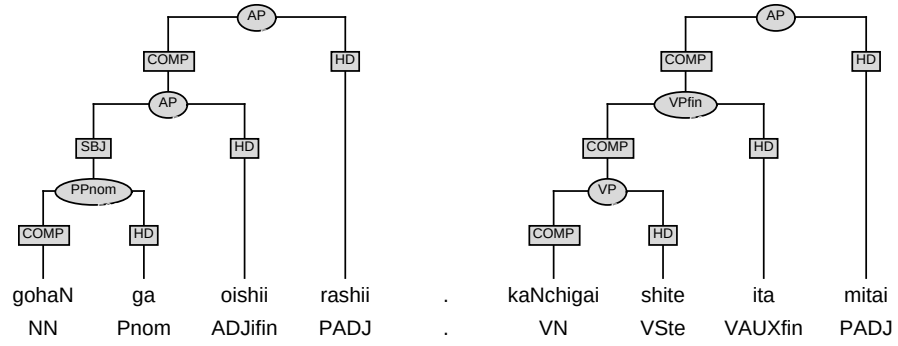

Figure 5.37: AP with PADJ

5.5 Adverbial Phrases

An adverbial phrase **ADVP** is, in principle, a phrase headed by an adverb and modifying a verbal element. In practice, adverbials modify verbs, adjectives, other adverbials,and some numerical expressions. Figure 5.38 ADVP shows a canonical example in which an adverb modifies another adverb, and that **ADVP** phrase modifies the following adjective.

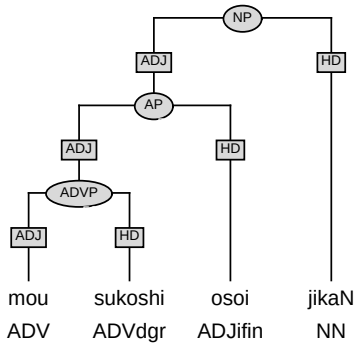

Figure 5.38: ADVP

5.5.1 Derived adverb – ADJiku –

The *-ku* ending form of the i-adjective (See Section 4.5) is a derivational adverb, and typically modifies verbals. Therefore, the *-ku* ending form of i-adjectives being modified by other adverbs are also a canonical adverbial phrase (Figure 5.39).

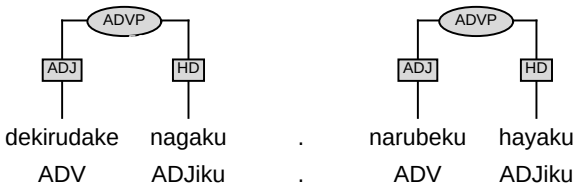

Figure 5.39: ADVP

Since **ADJiku**'s are derivation of i-adjectives, they have common argument structures with their adjective counterparts. Their left adjacent contexts can be those of the corresponding adjectives.

5.5.2 With postpositional – *ni, to* –

Adjectives are often accompanied by a following postpositional *ni*. Similarly to the *-ku* ending form of i-adjectives just mentioned above, the postpositional *ni* sometimes plays a role to alter an adjective into an adverbial. If this is the case, an adjective followed by postpositional *ni* is annotated as **ADVP** as a complement-head structure (Figure 5.40).

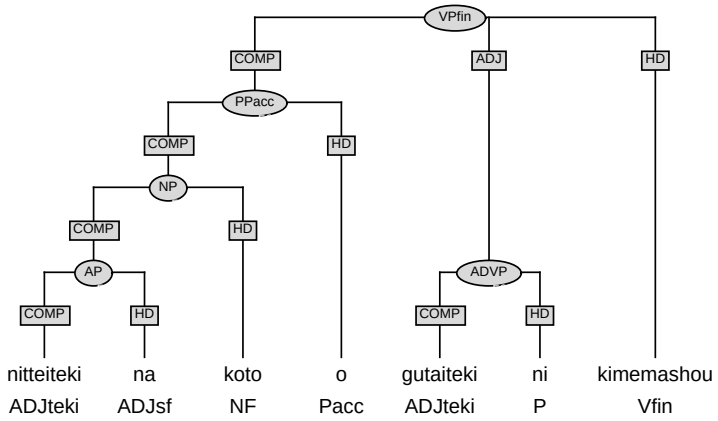

Figure 5.40: ADVP from adjectives

Also, an adverb is often accompanied by a postpositional *ni* or *to*. Since the resultant phrases are also adverbials, they are annotated as a head marker construction. One of the diagnostic features of this ADVP construction, contrasting with the one just mentioned above, is the presence or absence of the prepositional marker does not affect grammaticality at all. Typical examples are 5.41.

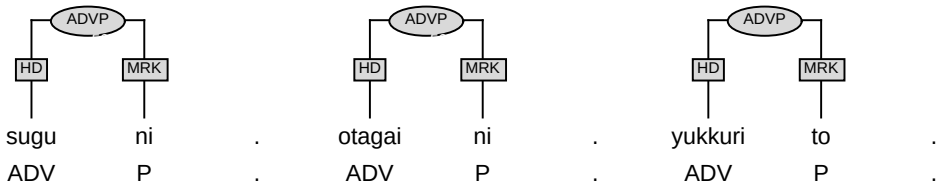

Figure 5.41: ADVP with marker

5.5.3 Particle ADV – *fuuni, shidai, nagara* –

Verb phrases ending in one of certain strings behave like adverbs. They modify the following verb or adjective phrases adding some sense of the mode of the event or the temporal relations between events. For example, *fuuni* ('as if'), *youni* ('as if'), *shidai*('as soon as'). They are annotated as ADVP with those particle adverbs **PADV** being the syntactic head of the phrase (Figure 5.42).

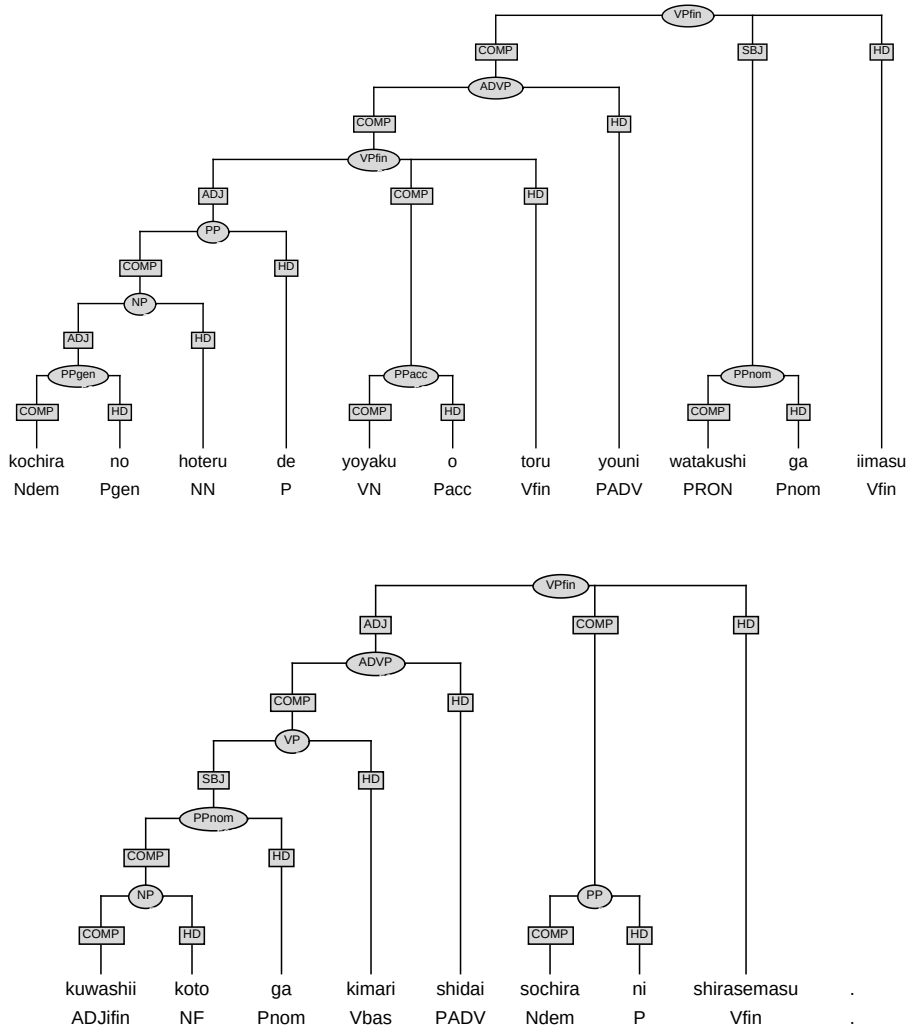

Figure 5.42: ADVP with PADV

5.5.4 – *mou ichido* –

A POS tag string pattern, **ADV** followed by **CDU**, that is, an adverb followed by a cardinal number with a unit, is often a noun phrase, for example, *choudo nanokakaN*, *daitai saNjuppuN*. The string *mou ichido* is, however, annotated as an **ADVP** of an adjunct-head schema. See also Chapter 6. This may be seen as analogous to *futatabi*[11] being tagged as an **ADV** instead of **CDU** because there does not seem to be similar expressions such as *mou nido*, *mou saNdo*, and *N*-times, and also because no nominal usage of *mou ichido* has been found in the dialogs so far.

5.6 Verb Phrases

Verbs are, in principle, subcategorized for their complements according to the lexicon. Modifiers are annotated as adjuncts that are not specifically selected by the verb. Therefore, canonical verb phrases are the head verb phrase preceded by a few PPs. If the annotator find a real ambiguity with no preference in PP attachment, then he or she chooses the structure with PP attachment to the higher possible node[12].

5.6.1 Complement and adjunct

Since we assume that the logical structure of the semantic contents is generally reflected on the syntax, full verbs are responsible for determining the syntactic structure according to their predicate-argument structure. The syntactic category of a complement depends on the specifications in the lexicon of each head category. Complements of verbs are canonically specific **PP**s and few other categories, but they are often materialized as bare **NP**s in the real world. Complements are generally annotated as sister nodes of the head verb as we have already seen (Figure 5.14, Figure 5.23, Figure 5.40, etc.). An example of an exceptional case is shown in Figure 5.45.

5.6.2 VP with auxiliary verb

Tokenizing the verbal sequence may sometimes be problematic due to the morpho-syntactic nature of the language. The auxiliary verbs are bound forms and may be considered to be agglutinating to the preceding full verbs. However, an auxiliary

[11] *Mou ichido* is literally 'once more', and *futatabi*, 'twice'.
[12] See also the "high attachment" strategy in Section 3.3.

verb[13] follows the full verb to attribute a certain function, such as temporal aspect of the whole event that the verb phrase refers to. Therefore, in most cases, auxiliaries are annotated as a head subcategorizing for a non-finite **VP** complement, which reflect a predicate-argument structure (Figure 5.43).

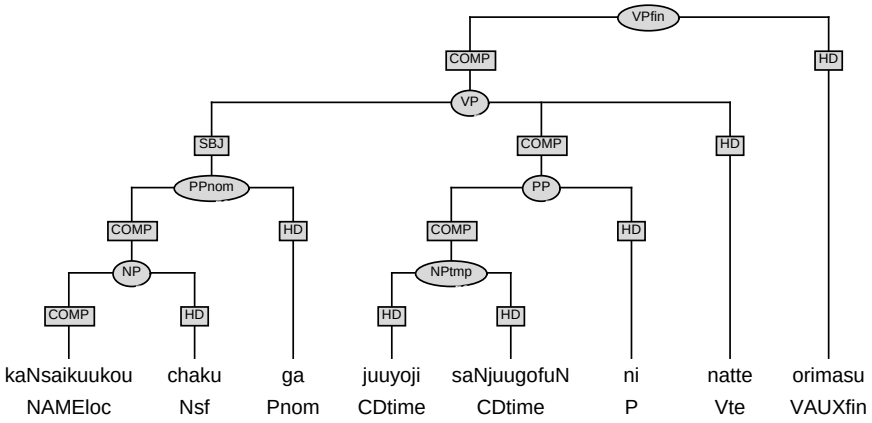

Figure 5.43: Full and Auxiliary Verbs

5.6.3 Complex predicate

There are several morpho-syntactic constructions in Japanese predicate expressions as we mentioned in Chapter 2. The causative *-(s)ase* and the passive *-(r)are* are classic examples having been talked about often in the linguistic literatures. The "morpheme" *-(r)are* also appears for expressing honorific and potential. They are not independently tokenized, but are a part of the verbal token in our Japanese treebank[14]. For example, *kime-sase-te* ('decide'-causative-participle), and *mi-rare-ru* ('see'-passive-present) are single tokens.

In polite speech, as opposed to plain speech, verbs end in a form of the verbal suffix -masu (e.g., *-masu, -mashita(ra), -mase(N), -mashou*). In our treebank, a matrix verb followed by *-masu* is tokenized as a single token, for example, *kimemasu* ('decide'-polite), *mimaseN* ('look'-polite-negation) are single tokens.

Combinations of more than one cases mentioned above are also single tokens. For example, *kaNgae-rare-masu* ('think'-passive-polite) is a single token

[13] A sequence of more than one auxiliaries possibly follows a full verb.

[14] In (Shibatani 1976), for example, the matrix verb and the causative *"sase"* are not only discrete but also the matrix verb belong to the subordinate clause, which, in turn, is a constituent of the causative main clause.

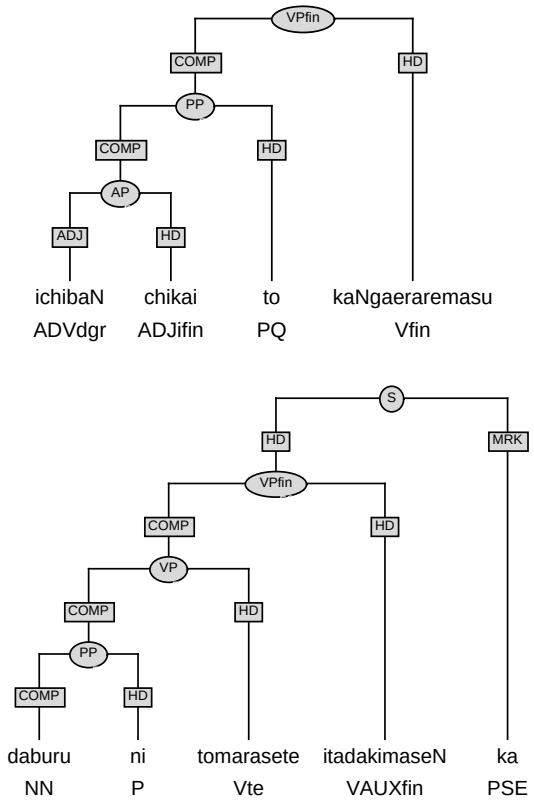

Figure 5.44: Complex verb (token)

(Figure 5.44)[15].

There are cases where a verb and one of its arguments are tightly bound to each other to form an idiomatic complex verb phrase, and that complex phrase functions as a single predicate[16]. In this case, the complex predicate is annotated to be a lower level constituent in the tree (Figure 5.45).

5.6.4 Support verb – *suru* –

It is recognized that a verbal noun and *suru* make a verbal constituent. The semantic content, or the predicate-argument structure, is inherited from the verbal noun **VN**, and the syntactic verbal features inherited from the support verb *suru*. Since we attempt to represent the predicate-argument structure in terms of sister

[15]See also Section 5.4 regarding adjectives and *-tai, -nai*.

[16]For example, *te ni ireru* (hand-in put) means 'acquire', *kyoumi ga aru* (interest-nom. exist) means 'be interested in', *tsugou ga ii* (circumstance-nom. good) means 'convenient'.

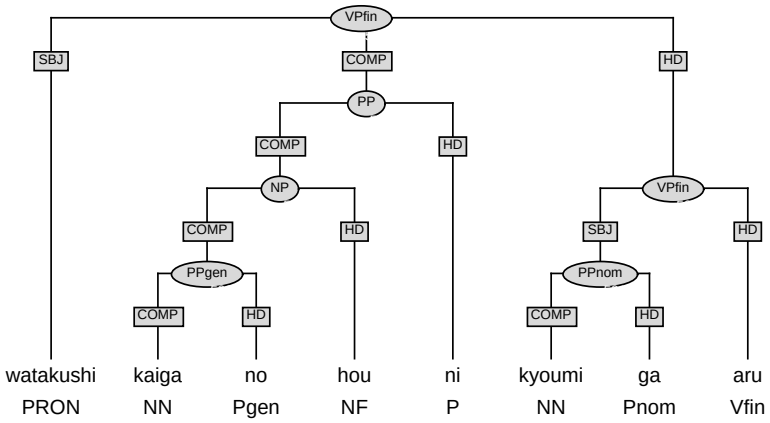

Figure 5.45: Idiomatic complex VP

relations on the tree configuration, a verbal noun followed by a support verb is annotated similarly to the case where a full verb followed by an auxiliary verb (Figure 5.46).

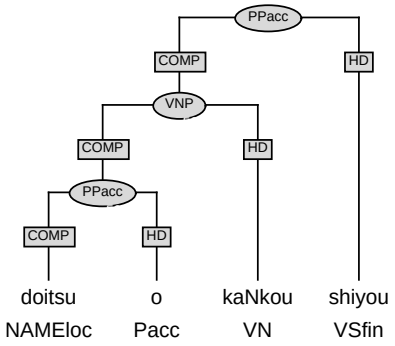

Figure 5.46: VNP *suru*

5.6.5 Particle verb – *desu, da* –

Particle verbs (e.g., *desu, deshita, deshita, da, datta, darou*) are, similarly to the verbs of existence in other languages, some kind of verbalizer, and at the same time, can be regarded as a copula verb.

A particle verb **PV** may be regarded as similar to an auxiliary, in the way that it conveys limited functions, such as temporal aspect and politeness, if it appears at the end of an expression that could stand without that particle verb.

For example (Figure 5.47), *"watakushi no hou wa kono bin de daijoubu"* is a grammatical utterance and makes sense. The following *desu* only adds a finite present verbal ending and increases politeness.

Alternatively, a **PV** may be regarded as a copula connecting two nominal expressions (Figure 5.48 left). In this case, the particle verb *desu* is assumed to be specified, in the lexicon, as subcategorizing for two arguments, a subject and a complement.

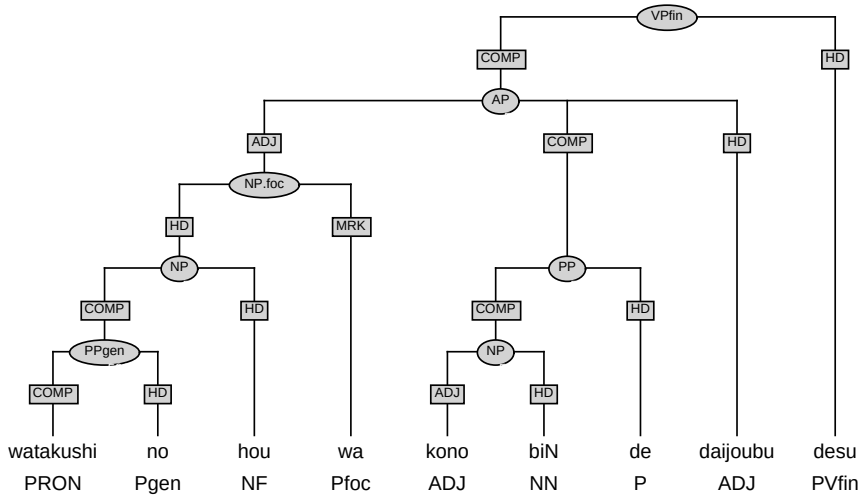

Figure 5.47: PV (particle verb)

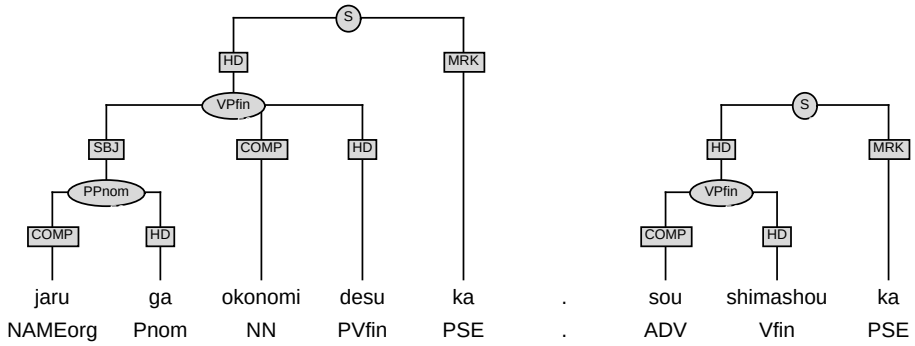

Figure 5.48: PV (copula) and sentence end marker

5.7 Sentences

What is an S? In many theories, the sentence is defined as a specific verb phrase or predicative adjective phrase with every complement and the subject being present[17]. In Japanese, however, as mentioned above in Chapter 2, finite verb phrases are well formed utterances, regardless of the absence of the subject or one or more of the complements that the head subcategorizes for in the lexicon.

In reality, missing arguments are so common in the Japanese sentences as mentioned also in Chapter 2, that completely saturated verb phrases are rarely found. If we took the position for the saturated verb phrase as **S**, almost no **S** would appear, and then the symbol **S** would not carry much information in the annotation. Therefore, the label **S** is used to indicate the syntactic unit described below.

What is a sign of S? There is a class of postpositions (**PSE**) which appears at the end of the utterance. Typical examples are, *ne* (tag question marker), *ka* (question marker), and *yo* (emphasis). These sentence end postpositions normally appear at the end of the finite verb phrase or the predicative adjective phrases. Since a sentence end postposition delimits the utterance in a way, we annotated as an **S** when the finite verb phrases (**VPfin**) or the predicative adjective phrases (**AP**) with some predicate-argument structure in it followed by a sentence end postposition (Figure 5.48, Figure 5.44).

Being preceded by a sentence initial conjunction **CNJ** is also a sign of **S** if the string ends with a proper ending form of a finite verb phrase or a predicative adjective phrase.

5.8 Combination of sentences

A subordinate clause is a finite verb phrase or an adjective phrase which is followed by one of the sentence-conjunctive postpositions. In principle, sentence-conjunctive postpositions are responsible for expressing logical relations between the subordinate clause on the left of it and the following part of the sentence, such as the causal, temporal, adversative, or embedded interrogative relations. Sentence-conjunctive postpositions are subclassified into three in terms of the POS tags [18].

[17] For example, "**S** denotes the SYNSEM value of a saturated verbal sign" (HPSG (Pollard and Sag 1994) p.28)

[18] Regarding POS tags, see Chapter 4, and Appendix A

The dependency of the subordinate clause is not necessarily upon an immediately following part, nor necessarily upon the rightmost predicate in the complex sentence. The annotator must combine more than two clauses in accordance with logical relations among those sentences convey. If two subordinate clauses precede a main clause, for example, the semantic content of the former subordinate clause could either be the premise of the semantic content of the main clause, or the premise of the semantic content of the second subordinate clause.

In any case, the combination follows one of the head final schemata, the adjunct-head. See also Chapter 6.

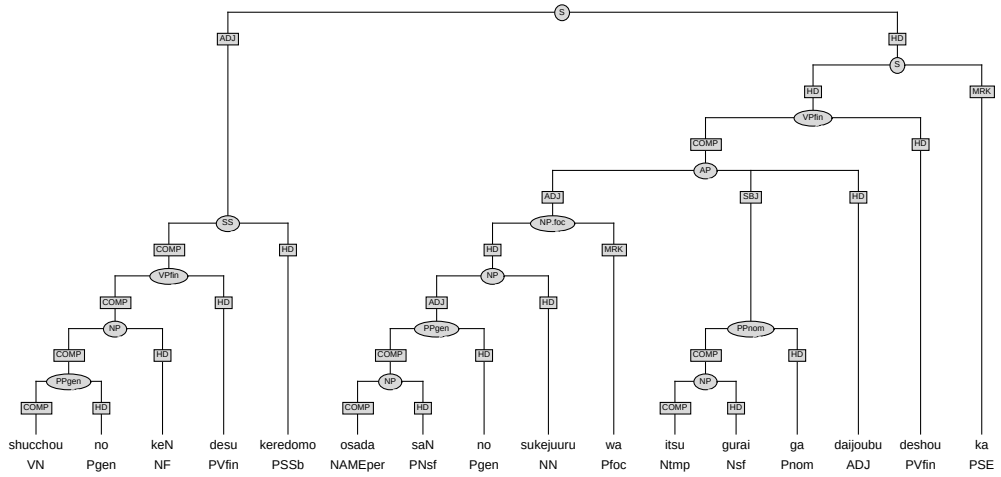

Figure 5.49: Combination of Sentences.

Chapter 6

Edge labels

In order to describe grammatical functions of constituents, we employ HPSG terminology[1]: **HD** (head), **SBJ** (subject), **COMP** (complement), **ADJ** (adjunct), and **MRK** (marker). The following six ID schemata describe the concatenation of constituents on various levels. Since Japanese is a head final language, the head usually constitutes the right hand side constituent except HD-MRK schema in which the markers appear after the head:

COMP-HD the left hand side constituent is subcategorized for by the right hand side constituent.

SBJ-HD the subject is a special form of complement, which the predicate selects.

ADJ-HD the left hand side constituent is not subcategorized for by the right hand constituent.

HD-MRK postpositions are markers when they carry pragmatic-rhetoric (in contrast to syntactic-semantic) function such as focus postpositions and sentence end postpositions.

HD-HD describes phenomena such as coordination and compositional expressions.

Dash-Dash describes the cases that do not fall into any of the above.

[1]See (Pollard and Sag 1994), and also Appendix C. Recent developments of unification based grammars in Japanese generally do not make a distinction among the subject, complements, and adjuncts ((Gunji and Hasida 1998)).

6.1 Complement

Complements are syntactically and/or semantically obligatory elements with respect to the head constituent of the phrase. Bound forms are also described in the complement-head schema. The predicate-argument structures are described in the same schema.

6.1.1 Bound Forms

Bound forms usually immediately follow the constituent they subcategorize for. Bound forms which appear in the VERBMOBIL-II transcription data are nominal suffixes, postpositions, formal nouns, auxiliary verbs, and so on. The relation between the bound form mentioned above and their precedent is normally described as complementation. There are some exceptions for this in which the bound form is classified as a marker, for example, sentence end postpositions, focus postpositions (see Section 6.4).

Figure 6.1 shows an example of suffixation, Figure 6.2 depicts a case of nominalization with the formal noun *no*, and Figure 6.3 contains five complement-head relations between postpositions (*kara, no, ga, ni, node*) and the phrases they follow.

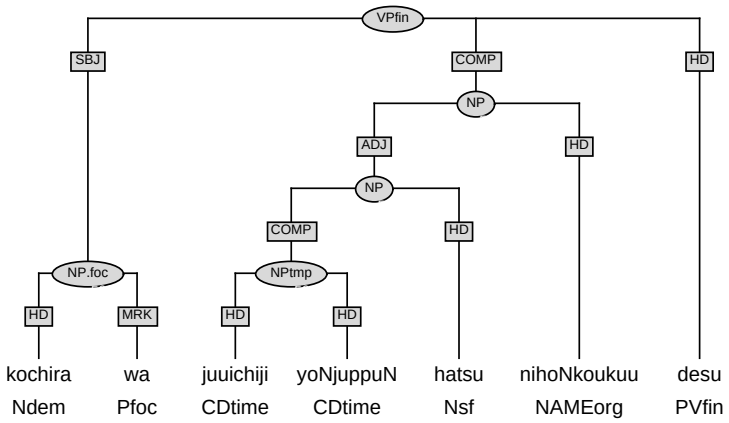

Figure 6.1: Suffixation (COMP-HD)

6.1.2 Predicate-argument Structure

Edge labels are used to describe the predicate-argument structure. As stated above, if a constituent is subcategorized for by the head constituent, it is called

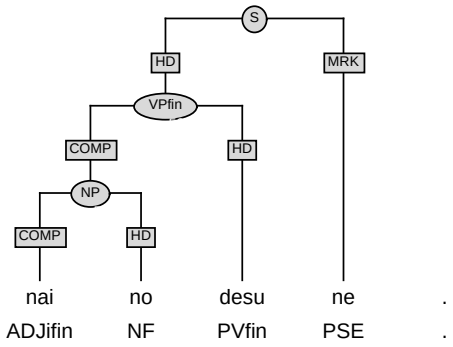

Figure 6.2: Formal noun (COMP-HD)

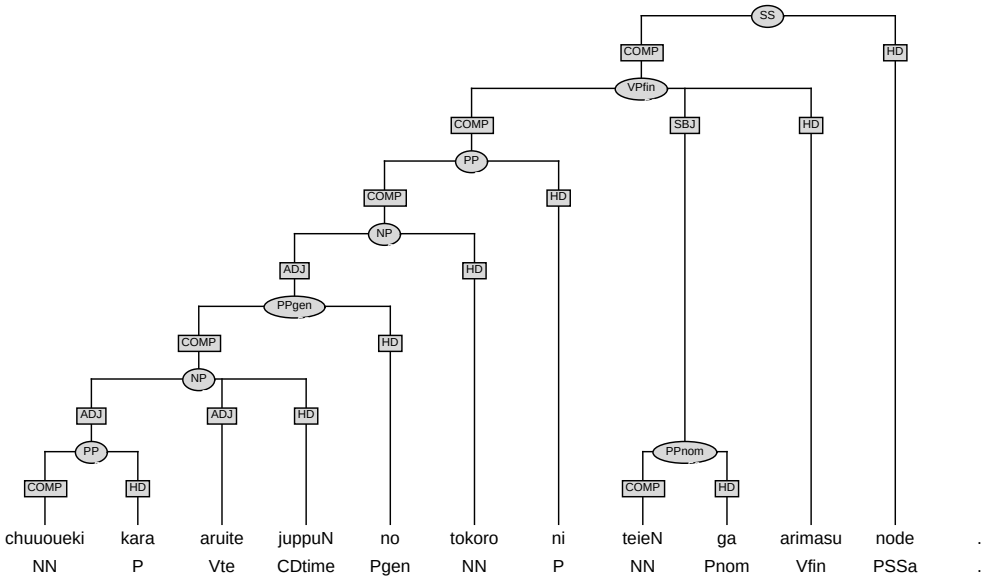

Figure 6.3: Postpositions (COMP-HD)

complement (or subject, being a special kind of complement, see Section 6.2).

The distinction between adjunct and complement is not always as straightforward in Japanese because complements can be dropped, and also because case postpositions are not always present in conversations. The decision whether a constituent is complement or adjunct has to be made upon considering not only morphology and syntax but also the semantic aspects of the strings to be annotated. Therefore, it is necessary to assume a lexicon that carries every predicative

item with all its possible subcat feature lists[2]. See Figure 6.4 for instance. The adjective *suki* ('to like') that subcategorizes for an experiencer, the one who likes, as subject. Its second complement, a theme, the object that is liked, is dropped probably because it is known to both dialog participants.

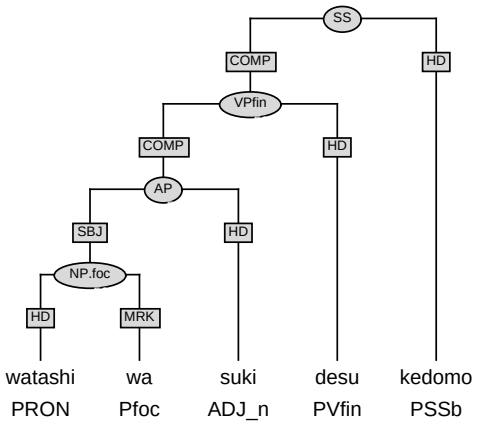

Figure 6.4: Predicate-argument structure

6.2 Subject

The notion of subject in Japanese is a controversial issue in the research literature. One can find such statements that there may be double subject constructions in Japanese, or that the notion of subject is not applicable to the Japanese language at all (Rickmeyer 1995). Unlike European languages, there is no morphological agreement between the supposed subject and the predicate in Japanese. Word order does not help to identify the subject either.

Yet, we think that it is worth trying to distinguish between subject and other complements, and describe each predicative lexical item as subcategorizing for a subject, the most salient argument in the subcat list.

A subject is usually a nominative phrase marked with postpositional *ga*. There are several exceptions for this generalization. Some verbs of sense (e.g., *wakaru* 'understand') and verbs of potential (e.g., *dekiru* 'can') allow a dative subject "experiencer" and a nominative complement "object". Adjectives of feeling (e.g., *suki* 'like') allow a nominative complement "object" (Figure 6.5). There are also other cases where the nominative *ga* is not marking the subject but the object

[2]Along with the treebank annotation, a reference valence list for verbs and adjectives has been created according to their semantic-syntactic structure.

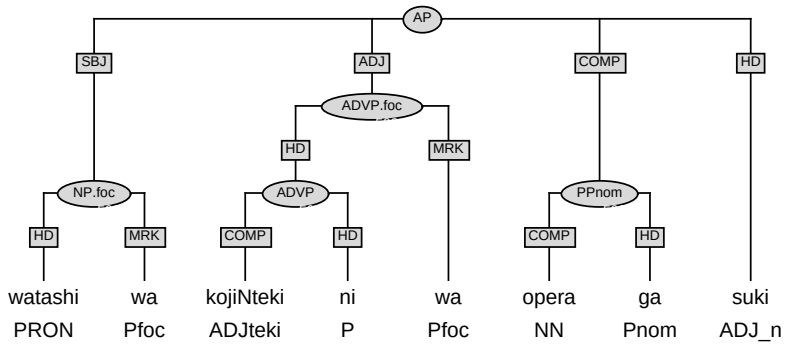

Figure 6.5: Nominative complement

of the reference of comparison, thus more than one nominative *ga* phrases are present in a local tree (Figure 6.6), in which one of the nominative *ga* phrases is an adjunct. According to (Narrog 1995), a sentence might theoretically contain

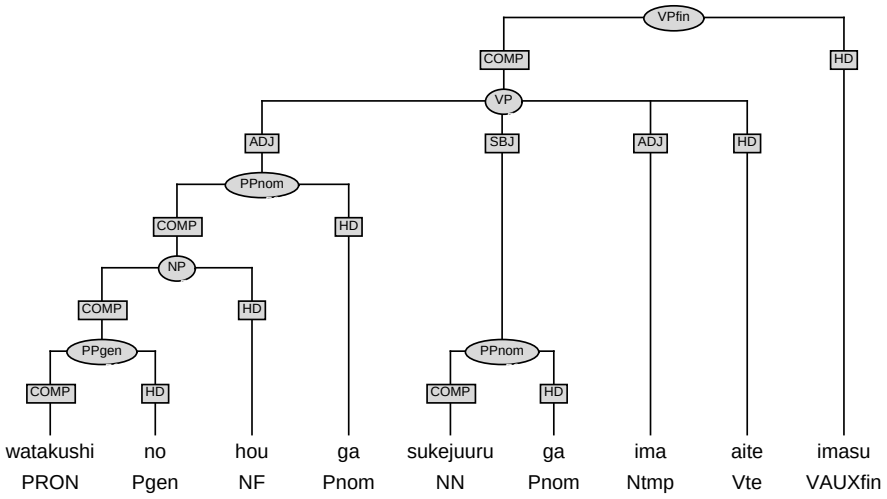

Figure 6.6: Double *ga*

up to three *ga*-phrases. The nominative *ga* might also be dropped in the real conversation, or replaced with one of the focus postpositions depending on the topic structure[3].

[3]It is problematic to distinguish focus (or topic) phrases from subjects because they may be but are not necessarily the same.

6.3 Adjunct

All the left hand side constituents of heads that are not complements nor subjects are adjuncts. Therefore, the left hand side part of a compound noun is an adjunct (Figure 6.7). Modifying adverbs and attributive adjectives are typical adjuncts

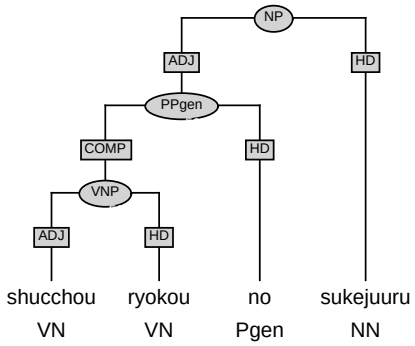

Figure 6.7: Compound noun (ADJ-HD)

(Figure 6.8). Sentential conjunctions are adjuncts (Figure 6.9). PPs and NPs that

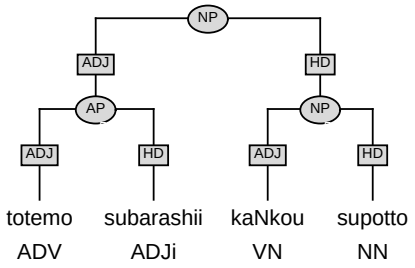

Figure 6.8: Adverb and adjective (ADJ-HD)

are not subcategorized for by the head are also adjuncts (Figure 6.10, Figure 6.11). A subordinate clause in combination with the main clause is described as adjuncts (Figure 6.12).

6.4 Markers

The following postpositions are treated as markers:

- sentence end postpositions (**PSE**), for example, *ka, ne,*

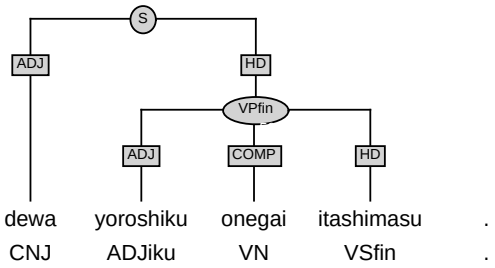

Figure 6.9: Conjunction (ADJ-HD)

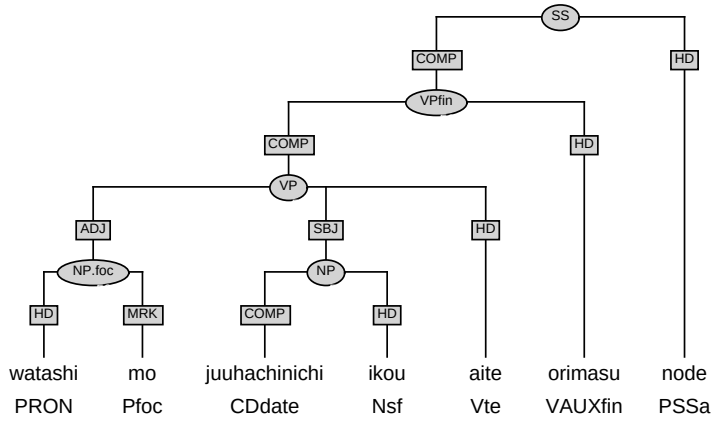

Figure 6.10: PP (adjunct)

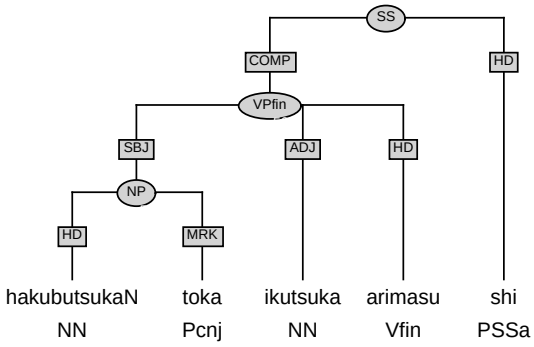

Figure 6.11: NP (adjunct)

- focus postpositions (**Pfoc**), for example, *wa, mo*,

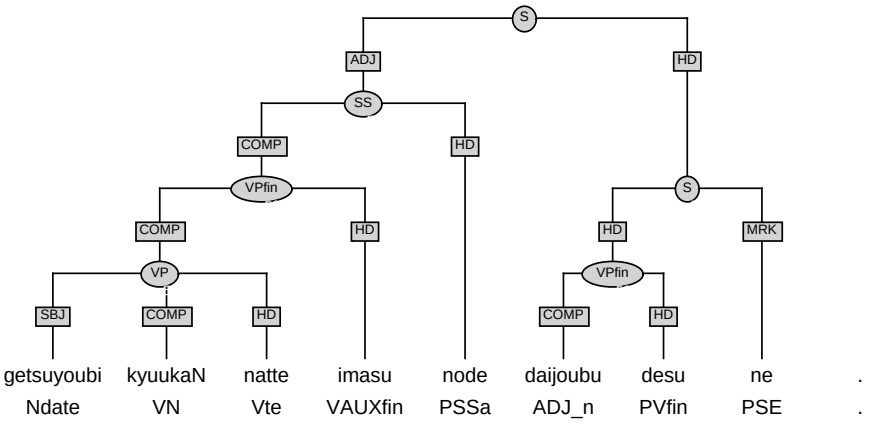

Figure 6.12: Subordinate clause (adjunct)

- postpositions which follow adverbs, namely *ni* and *to*,
- coordination markers, either particles (**Pcnj**), for example, *to, toka*, (Figure 6.19) or conjunctions (**CNJ**), for example, *aruiwa, ato, sorekara* (Figure 6.20).

Sentence end postpositions

Sentence end postpositions such as *ka, kana, kke, mono, kashira, ne, yo*, typically appear after a finite verb phrases or adjective phrases, and they are then described as a sentence. Sentence end postpositions are bound to their preceding part, and are the markers of the left hand side constituent. Figure 6.13 shows an example of a finite verb phrase headed by the interjectional sentence final postposition *ne*, which is described as a sentence.

PSE appearing after other phrases are rare in polite speech, nevertheless one can find examples like in Figure 6.14.

Focus postpositions

Focus postpositions organize the topic and information structure. They often appear in many fixed expressions such as *"shite wa ikenai/dame"* ('must not do'), *"shite mo ii"* ('may do'). They are bound to their left hand side constituent, which are the head of the phrases. A noun phrase followed by a focus postposition will be a focused noun phrase (**NP.foc**) as in Figure 6.15.

Similarly, there are focused postpositional phrases (**PP.foc**). Focus postpositions normally cooccur neither with a nominative *ga* nor with an accusative

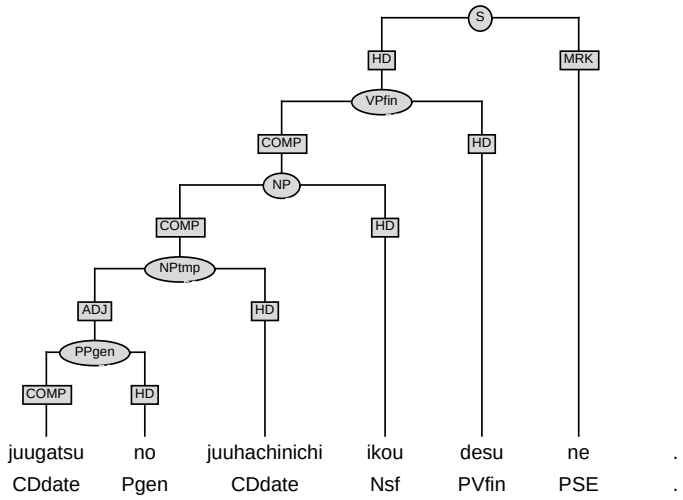

Figure 6.13: VP and PSE (marker)

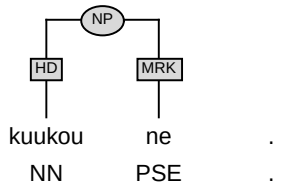

Figure 6.14: NP and PSE (marker)

o[4], however, they may appear after other postpositions as in Figure 6.16. Focus postpositions may occur after a non-finite verb phrases, which ends in the participle (*-Te*) form, and form focused verb phrases (**VP.foc**). An adjective phrase ending in the participle (*-Te*) form may be followed by one of the focus postpositions to form focused adjective phrases (**AP.foc**). An adverbial phrase may also be followed by one of focus postpositions to form a focused adverbial phrase (**ADVP.foc**).

Postpositions after adverbs

One of the two postpositions *ni* and *to*, often follows an adverb. They are described as markers because their function is not syntactic but rhetorical. In Figure 6.17,

[4]In a certain style of writing, the combination of accusative *o* and focus postposition *mo* is possible.

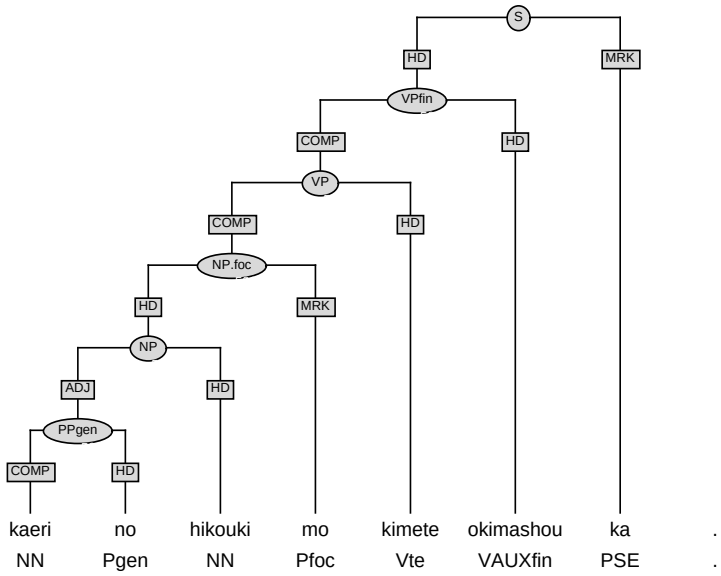

Figure 6.15: NP focus (marker)

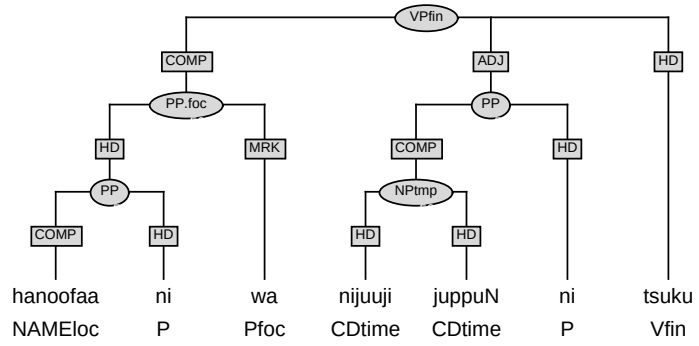

Figure 6.16: PP focus (marker)

for example, the postposition *to* is optional.

Coordination markers

Conjunctive postpositions mark either the coordinated phrase or the phrases to be coordinated (Figure 6.18). Coordination may also be marked by conjunctions as in Figure 6.19.

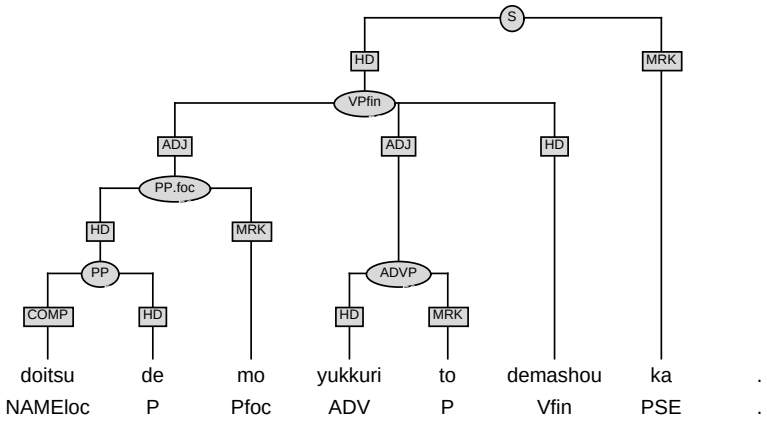

Figure 6.17: Adverb and postposition (marker)

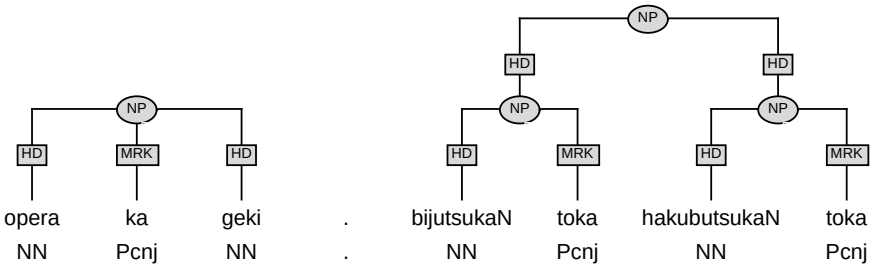

Figure 6.18: Two types of coordination

6.5 HD-HD

The head-head construction apply to the listing of items, coordinations, and compositional expressions. Listing and coordinations are described as each item being head. Examples of coordination are shown in Figure 6.18, 6.19, and 6.20. See also Section 5.2.5.

Compositional expressions are mostly as follows:

Date and time expressions are often composition of several temporal nouns and/or NPs, for example,

nijuurokunichi getsuyoubi,
kuji saNjuugofuN (Figure 6.21).

Full name is composition of a family name and a following first name, for example,

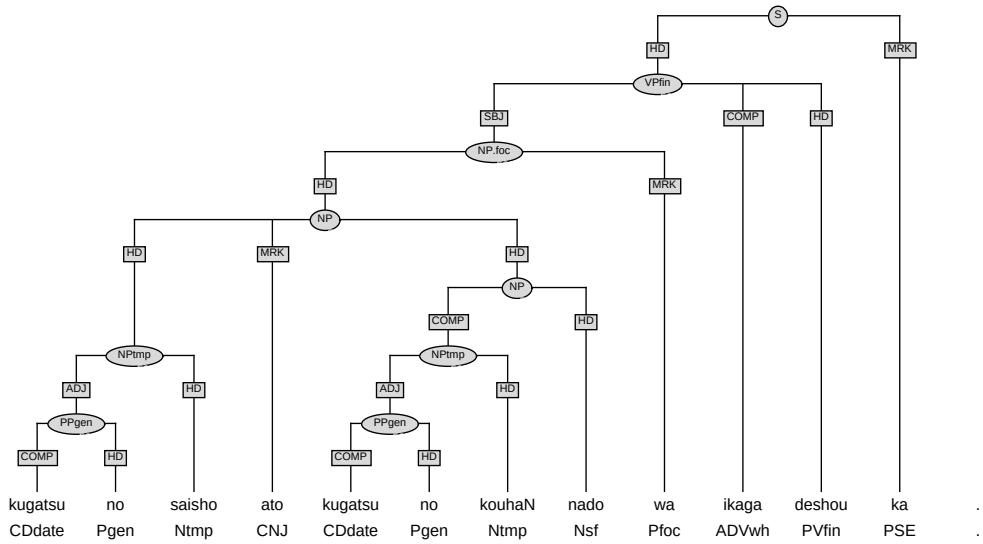

Figure 6.19: Conjunction (coordination marker)

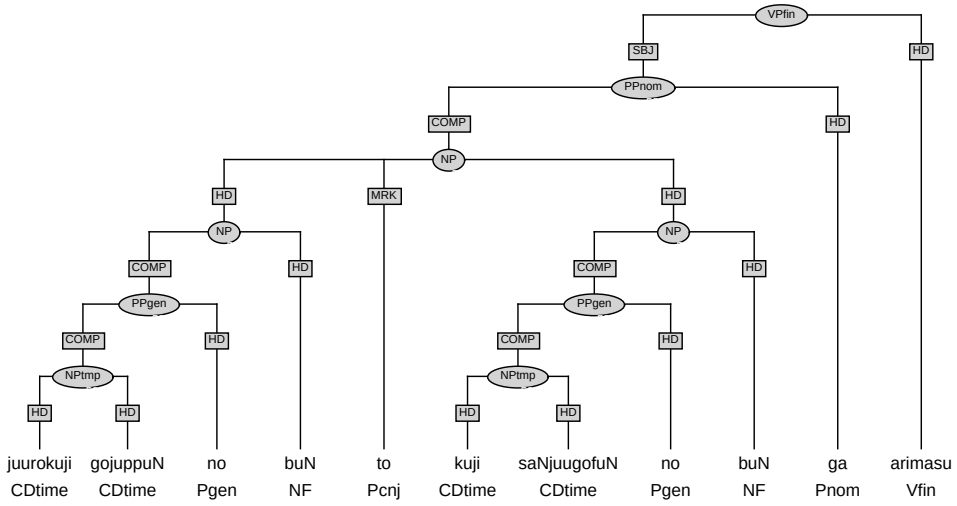

Figure 6.20: Coordination of NPs

saitou shirou,
kasahara arisa (Figure 6.22).

From-to expressions are two consecutive PPs that often behave like nouns syntactically and semantically. We assume that this is common also in many other languages. For example,

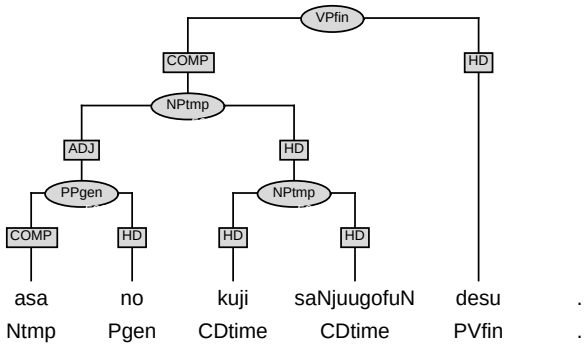

Figure 6.21: Time expression

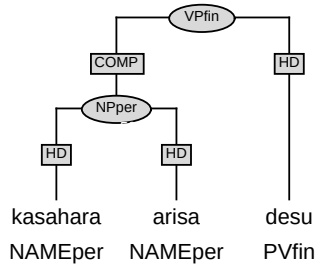

Figure 6.22: Personal name

saNgatsu no itsuka kara juuninichi made ga aite imasu. roNdoN kara kaNkuu made ga jaru ni narimasu (Figure 6.23).

6.6 "-" (Dash)

The edge label **Dash** ("-", standing for "unspecified") is used to deal with everything that does not fall into the above schemata.

Numericals A sequence of tokenized several cardinal numbers with or without a unit[5], for example,

hyaku hachijuu maruku, 'one hundred eighty marks' (Figure 6.24).

[5]Tokenizing some numerical expressions depends on pragmatic reason in the project goal, namely the translation into German.

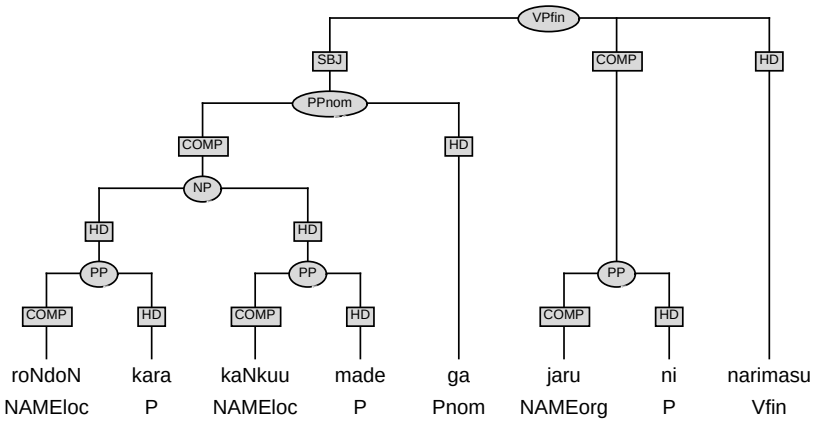

Figure 6.23: From-to expression

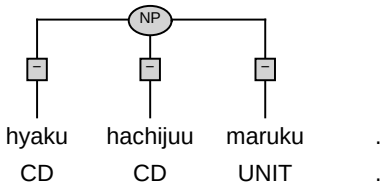

Figure 6.24: Cardinal number with unit

Speech errors Errors, hesitations, false starts, and the like — strings which are regarded not to be the intended speech and transcribed so — are structured as far as possible, but are not included in the tree structure, and marked as errors with the node label "**err**" and the edge label "**-**". Figure 6.25 gives an example of a false start[6].

Interjections interjectional expressions. Interjectional expressions which contain more than one word, are also comprised and labeled as "**ITJ**" and each member is given the edge label "**-**". Figure 6.26 gives an example of the most frequent interjectional expression in our data, *eetto desu ne*.

[6]Note that this is one of a few cases where the unary branching tree is not avoided.

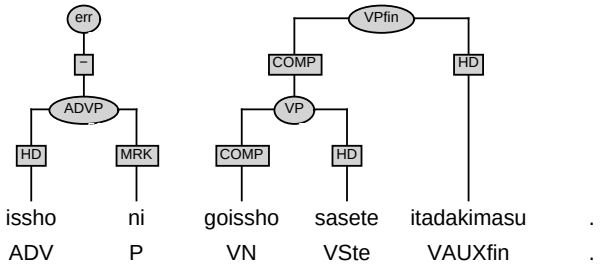

Figure 6.25: Error: False start

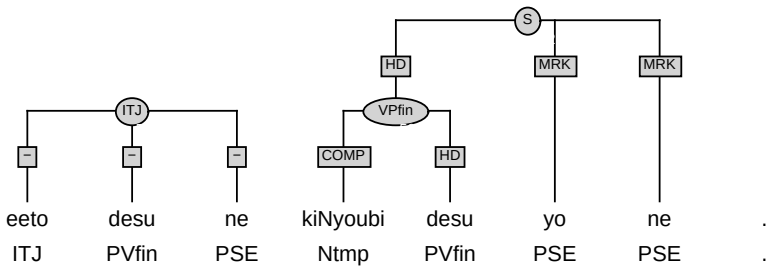

Figure 6.26: Interjectional expression

Chapter 7

Conclusion and Open Research Issues

The annotation scheme for the Japanese treebank described in this stylebook takes into account both the language specific nature of Japanese and the peculiarity of spontaneous spoken dialogs. Nevertheless, we tried to take as neutral linguistic assumptions as possible in order to ensure reusability of the data for further future research. The VERBMOBIL-II Japanese Treebank consists of about 18,000 trees as of September 2000. This is probably the first large scale Japanese treebank with its string level transcribed into Roman characters. We hope the treebank will serve for a wide group of users in the NLP community. As concluding remarks, we would like to mention some of the remaining problems and challenges for the future. We notice that many other problems might also remain unsolved in our treebank. However, we hope that the treebank provides useful data and contributes to further researches by users.

7.1 Tokenization

Tokenization of Japanese presents one of the most difficult problems for syntactic analysis. There are two major conventions for romanizing Japanese characters. However, when it comes to the question of setting the boundaries between words, it is difficult to find the authoritative convention since the Roman character set is foreign to the Japanese language.[1] Because of the nature of the Japanese lan-

[1] On the other hand, there are some advantages in describing a language in the foreign character set that does not belong to the language. Each Japanese character represents a syllable. This causes the perceptual difficulty for the native speakers to see a segment boundary between an onset consonant and the following vowel no matter how it is linguistically motivated. Transcription of the Japanese language into Roman characters introduces more possibilities in

guage and its writing system mentioned in Chapter 2, there are more than one possibilities in tokenizing a given string. Decisions about segmentation will in cases affect and anticipate a specific way of syntactic analysis, which excludes other interpretations. This means also that two sets of data with different segmenting conventions will be difficult to compare. The emerge of the theory about tokenization is expected.

7.2 The Classification of the Postpositions

The treatment of the postpositions is another difficulty in dealing with Japanese syntax because a certain postposition may encode the variety of different grammatical functions. It is not at all clear which case postpositions are marking the major grammatical functions. The seemingly straightforward ones might have various uses. The nominative case postposition *ga* usually marks a subject, but in some cases, an complement (object) or an adjunct, for instances,

subject: *watakushi ga yoyaku itashimasu*
('I am making the reservation'),

object: *oNgaku ga suki*
('(I) like music'),

adjunct: *shiNbo saN no hou ga juunigatsu ga ii*
('December is better for you — Mr. Shinbo')

We tried the best to solve this problem, but we are aware that there are controversial opinions especially about the question whether postpositions are heads or not ((Pollard and Sag 1994), (Siegel 1998)). Also, further development of the theory of mapping between the morpho-syntactic level of description to the semantic and more abstract conceptual levels of descriptions is expected.

7.3 Notion of Subject

As stated in Chapter 6 the notion of subject is not straightforward in Japanese. There are controversial opinions about what is the subject logically and grammatically. It is not obvious by only looking at the surface structure because of frequent dropping of the subject and the complements, and because of the topic marking which override case marking. The case particles are dropped frequently in spoken language, grammar rules do not always capture facts in the real world, and the

linguistic description in a sense.

speaker might make mistakes. Within these situations, the identification of the subject are further complicated. In polite speech it frequently occurs that the subject (which would be marked with *ga*) is expressed indirectly or euphemistically by locative expressions that are generally seen as adjuncts and not an argument of a predicate, for example, *watakushi no hou de yoyaku irete okimasu node .* ('I – lit. on my side – will make the reservation').

7.4 Topics

The flat clustering strategy is employed in the treebank in order to present an account for the argument structure of each predicate. It is an advantage where scrambling is a common phenomena. However, sometimes this representation does not exactly account for the syntax of a language in which also the topic is prominent (besides the subject being prominent as well). Topics may or may not coincide with grammatical complements, and there may be more than one topics in one sentence. Therefore, topic phrases give us difficulties not only in identifying the complement of a predicate, but also in interpreting the remaining adjunct topics.

Appendix A

POS tags

POS	Description	Example
,	Comma	,
.	Sentence final punctuation	.
?	Question mark	*?*
ADJ	Atributive adjective	*iroNna, taishita*
ADJdem	Demonstarative adjective	*sono, kono, koNna, soNna, ano*
ADJicnd	i-adjective (conditional)	*yokereba, yasukereba*
ADJifin	i-adjective (finite)	*yoroshii, ii, nai, chikai*
ADJiku	i-adjective (*-ku* ending)	*hayaku, yoku, osoku, nagaku*
ADJite	i-adjective (*-te* ending)	*chikakute, yasukute, yokute*
ADJsf	Adjective suffix	*na*
ADJteki	na-adjective (*-teki* ending)	*jikaNteki, kojiNteki, gutaiteki*
ADJwh	Wh Adjective	*dono, doNna*
ADJ_n	na-adjective	*daijoubu, kekkou, beNri, kirei*
ADV	ADVerbials in general	*mou, mata, dekireba, daitai*
ADVdem	Demonstrative adverb	*sou, kou*
ADVdgr	Degree adverb	*ichibaN, sukoshi, chotto, amari*
ADVtmp	Temporal adverb	*mazu, sassoku, sakihodo*
ADVwh	Wh adverb	*dou, ikaga, doushite*
CD	Cardinal number	*hyaku, nihyaku, saN, ni*
CDdate	Cardinal⊕date unit	*juugatsu, tooka, nigatsu*
CDtime	Cardinal⊕time unit	*gojuppuN, juuichiji, juuji*
CDU	Cardinal⊕unit	*itsukakaN, futaheya*
CNJ	Conjunction	*dewa, soredewa*
GR	Greeting	*koNnichiwa, hajimemashite*
ITJ	Interjection	*eeto, maa*
NAME	Other proper noun	*doNjobaNni*
NAMEloc	Proper noun; location	*hanoofaa, doitsu, nihoN*
NAMEorg	Proper noun; organizaion	*rufutohaNza, jaru*
NAMEper	Proper noun; person	*matsumoto, yoshikawa*
Ndem	Demonstarative noun	*sore, kochira, sochira*

POS	Description	Example
NF	Formal noun	*hou, no, koto*
NN	Common noun	*hoteru, hikouki*
Nsf	Noun sufix	*hatsu, chaku, keiyu*
Ntmp	Noune (temporal)	*getsuyou, gozeN*
Nwh	Wh noun	*dochira, naNji, dore*
P	Postposition	*ni, de, kara, made, to*
Pacc	Accusative case	*o*
PADJ	Particle adjective	*youna, rashii*
PADV	Particle adverb	*youni, fuuni, shidai*
Pcnj	Conjunctive particle	*to, toka, ka, ya*
Pfoc	Focus	*wa, mo, demo, nara, koso*
Pgen	Genitive case	*no*
Pnom	Nominative case	*ga*
PNsf	Personal name suffix	*san, sama*
PQ	Quotative postposition	*to, tte*
PreN	Noun prefix	*yaku, dai*
PRON	Pronoun	*watashi, watakushi*
PSE	Sentence end postposition	*ka, ne, yo, na, kana*
PSSa	Subordinate S postposition (and)	*node, to, kara, shi*
PSSb	Subordinate S postposition (but)	*ga, keredomo, kedo*
PSSq	Subordinate S postposition (question)	*ka*
PV	Particle verb	*dattari*
PVcnd	Particle verb (conditional)	*dattara, deshitara*
PVfin	Particle verb (finite)	*da, datta, desu, deshita*
PVte	Particle verb (*-te* ending)	*deshite, de*
UNIT	Unit	*maruku, biN, meetoru*
V	Verb (othre forms)	*mitari, kaNgaenagara*
VADJi	Verb⊕i-adjctive	*shitai, mitai, ikitai, kimetai*
VADJicnd	Verb⊕i-adjctive(conditional)	*noranakya, ikanakereba*
VADJ_n	Verb⊕na-adjctive	*ikesou, arisou, toresou*
VAUX	Auxiliary verb	*shimattari*
VAUXbas	Auxiliary verb (base)	*itadaki*
VAUXcnd	Auxiliary verb (conditional)	*ireba*
VAUXfin	Auxiliary verb (finite)	*iru, ita*
VAUXte	Auxiliary verb (*-te* ending)	*itadaite*

POS	**Description**	**Example**
Vbas	Verb (base)	*mi, tore, kimari*
Vcnd	Verb (conditional)	*areba, attara*
Vfin	Verb (finite)	*iu, aru, omoimasu*
Vimp	Verb (imperative)	*kudasai, nome*
VN	Verbal noun	*kaNkou, shuppatsu, yoyaku*
VS	Support verb	*shitari*
VSbas	Support verb (base)	*shi*
VScnd	Support verb (conditional)	*shitara, sureba*
VSfin	Support verb (finite)	*suru, shita*
VSimp	Support verb (imperative)	*shiro*
VSte	Support verb (*-te* ending)	*shite*
Vte	Verb (*-te/de* ending)	*aite, shite, tsuite*
xxx	Tokenizing problem (temporary)	

Appendix B

Node labels

Node label	Description	Example
NP	Noun phrase	*shukuhaku ryoukin*
NPloc	Noun phrase (location)	*doitsu no hanoofaa*
NPper	Noun phrase (person)	*nagata youko, satou saN*
NPtmp	Noun phrase (temporal)	*juugatsu no juuhachinichi, gogo kuji saNjujgofuN*
NP.foc	Noun phrase (focus)	*watashi wa*
VNP	Verbal noun phrase (NP or VP)	
VP	Verb phrase	
VPcnd	Verb phrase (conditional)	
VPfin	Verb phrase (finite)	
VP.foc	Verb phrase (focus)	
PPacc	Postpositional phrase (accusative)	*hanoofaa o, yoyaku o*
PPgen	Postpositional phrase (genitive)	*kochira no, hoteru no*
PPnom	Postpositional phrase (nominative)	*yotei ga, jikaN ga*
PP	Postpositional phrase	*eki kara, kaNkuu ni*
PP.foc	Postpositional phrase (focus)	*eki kara wa*
AP	Adjective phrase	*chotto yayakoshii*
APcnd	Adjective phrase (conditional)	*moshi hoteru ga eki kara chikakereba*
AP.foc	Adjective Phrase (focus)	*chotto yayakoshii*
ADVP	Adverb phrase	*sugoku hayaku*
ADVP.foc	Adverb phrase (focus)	*gutaiteki ni wa*
S	Sentence	
SS	Subordinated sentence	
GR	Greeting	*ohayou gozaimasu*
ITJ	Interjective expression	*eeto desu ne*
err	False start, speech error	

Appendix C

Edge labels

Edge label	Description
HD	Head
COMP	Complement
ADJ	Adjunct
SBJ	Subject
MRK	Marker
-	Unspecified

References

Bos, J., and J. Heine. 2000. Discourse and dialog semantics for translation. In *Verbmobil: Foundations of Speech-to-speech Translation*, 336–347, Berlin. Springer.

Brants, T., and W. Skut. 1998. Automation of treebank annotation. In *Proceedings of the Conference on New Methods in Language Processing, Sydney, Australia*, 49–57.

Brill, E. 1992. A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing, Trento, Italy.*

Burger, S. 1997. Transliteration spontansprachlicher daten - lexikon der transliterationskonventionen - VERBMOBIL II. Technical Report 56, Verbmobil.

Gunji, T. 1987. *Japanese Phrase Structure Grammar.* Dordrecht: D. Reidel Publishing Company.

Gunji, T., and K. Hasida. 1998. *Topics in Constraint-Based Grammar of Japanese.* Dordrecht: Kluwer Academic Publishers.

Kordoni, V. 2000. Stylebook for the English Treebank in VERBMOBIL. Technical Report 241, Verbmobil.

Narrog, H. 1995. Typology of the japanese clause structure. In *Bochumer Jahrbuch für Ostasienforschung, Bd. 19, 1995. München: Iudicum*, 147–164.

Nerbonne, J., K. Netter, and C. Pollard. 1994. *German in Head-Driven Phrase Structure Grammar*. CA: CSLI publishers.

Plaehn, O. 1998. ANNOTATE: Bedienungsanleitung. NEGRA Project. Technical report, Saarbrücken, Germany: Universität des Saarlandes, Computerlinguistik.

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

Reape, M. 1994. Domain union and word order variation in german. In *German in Head-Driven Phrase Structure Grammar*, 151–197, CA. CSLI Publications.

Rickmeyer, J. 1995. *Japanische Morphosyntax*. Heidelberg: Julius Groos Verlag.

Shibatani, M. 1976. *Syntax and Semantics vol.5*. New York: Academic Press.

Siegel, M. 1996. Die japanische syntax im verbmobil-forschungsprototypen. Technical Report 133, Verbmobil.

Siegel, M. 1998. Japanese particles in an hpsg grammar. Technical Report 220, Verbmobil.

Spencer, A. 1991. *Morphological Theory*. Cambridge, USA.: Blackwell.

Stegmann, R., H. Telljohann, and E. W. Hinrichs. 2000. Stylebook for the german treebank in VERBMOBIL. Technical Report 239, Verbmobil.

Tsujimura, N. 1996. *An introduction to Japanese Linguistics*. Cambridge, USA.: Blackwell.

Wahlster, W. 2000. *Verbmobil: Foundations of Speech-to-speech Translation*. Berlin: Springer.