Lexicon Features for Japanese Syntactic Analysis in Mu-Project-JE

Yoshiyuki Sakamoto
Electrotechnical Laboratory
Sakura-mura, Niihari-gun, Ibaraki, Japan

Masayuki Satoh
The Japan Information Center of Science and Technology
Nagata-cho, Chiyoda-ku
Tokyo, Japan

Tetsuya Ishikawa
Univ. of Library & Information Science
Yatabe-machi, Tsukuba-gun, Ibaraki, Japan

0. Abstract

In this paper, we focus on the features of a lexicon for Japanese syntactic analysis in Japanese-to-English translation. Japanese word order is almost unrestricted and *Kakujo-shi* (postpositional case particle) is an important device which acts as the case label(case marker) in Japanese sentences. Therefore case grammar is the most effective grammar for Japanese syntactic analysis.

The case frame governed by *Yougen* and having surface case(*Kakujo-shi*), deep case(case label) and semantic markers for nouns is analyzed here to illustrate how we apply case grammar to Japanese syntactic analysis in our system.

The parts of speech are classified into 56 sub-categories.

We analyze semantic features for nouns and pronouns classified into sub-categories and we present a system for semantic markers. Lexicon formats for syntactic and semantic features are composed of different features classified by part of speech.

As this system uses LISP as the programming language, the lexicons are written as S-expression in LISP, punched onto tapes, and stored as files in the computer.

1. Introduction

The Mu-project is a national project supported by the STA(Science and Technology Agency), the full name of which is "Research on a Machine Translation System(Japanese - English) for Scientific and Technological Documents."*

We are currently restricting the domain of translation to abstract papers in scientific and technological fields. The system is based on a transfer approach and consist of three phases: analysis, transfer and generation.

In the first phase of machine translation, analysis, morphological analysis divides the sentence into lexical items and then proceeds with semantic analysis on the basis of case grammar in Japanese. In the second phase, transfer, lexical features are transferred and at the same time, the syntactic structures are also transferred by matching tree pattern from Japanese to English. In the final generation phase, we generate the syntactic structures and the morphological features in English.

2. Concept of a Dependency Structure based on Case Grammar in Japanese

In Japan, we have come to the conclusion that case grammar is most suitable grammar for Japanese syntactic analysis for machine translation systems. This type of grammar had been proposed and studied by Japanese linguists before Fillmore's presentation.

As word order is heavily restricted in English syntax, ATNG(Augmented Transition Network Grammar) based on CFG(Context Free Grammar) is adequate for syntactic analysis in English. On the other hand, Japanese word order is almost unrestricted and *Kakujo-shi* play an important role as case labels in Japanese sentences. Therefore case grammar is the most effective grammar for Japanese syntactic analysis.

In Japanese syntactic structure, the word order is free except for a predicate(verb or verb phrase) located at the end of a sentence. In case grammar, the verb plays a very important role during syntactic analysis, and the other parts of speech only perform in partnership with, and equally subordinate to, the verb.

That is, syntactic analysis proceeds by checking the semantic compatibility between verb and nouns. Consequently, the semantic structure of a sentence can be extracted at the same time as syntactic analysis.

3. Case Frame governed by *Yougen*

The case frame governed by *Yougen* and having *Kakujo-shi*, case label and semantic markers for nouns is analyzed here to illustrate how we apply case grammar to Japanese syntactic analysis in our system.

Yougen consists of verb, *Keiyou_shi*(adjective) and *Keiyoudou_shi*(adjectival noun). *Kakujo_shi* include inner case and outer case markers in Japanese syntax. But a single *Kakujo_shi* corresponds to several deep cases: for instance, 'NI' indicates more than ten case labels including SPAce, Space TO, TIMe, ROLe, MANner, GOAl, PARtner, COMponent, CONdition, RANge..... We analyze relations between *Kakujo_shi* and case labels and write them out manually according to the examples found out in sample texts.

* This project is being carried out with the aid of a special grant for the promotion of science and technology from the Science and Technology Agency of the Japanese Government.

As a result of categorizing deep cases, 33 Japanese case labels have been determined as shown in Table 1.

Table 1. Case Labels for Verbal Case Frames

Japanese Label	English Label	Examples
(1) 主体	SUBject	～が
(2) 対象	OBJect	～を
(3) 受け手	RECipient	～に与える
(4) 与え手	ORIgin	～から受ける，奪う
(5) 相手 1	PARtner	～と協議する，異なる，～に関連する
(6) 相手 2	OPPonent	～から保護する，独立する
(7) 時	TIMe	1980年に
(8) 時・始点	Time-FRom	5月から
(9) 時・終点	Time-TO	来年まで
(10) 時間	DURation	5分間加熱する
(11) 場所	SPAce	～に位置する，～で発生する
(12) 場所・始点	Space-FRom	～から帰る
(13) 場所・終点	Space-TO	～へ送る，～に到達する
(14) 場所・経過	Space-THrough	～を通る，上空を飛ぶ
(15) 始状態	SOUrce	5.5%から6%へ引き上げる
(16) 終状態	GOAl	英語から日本語に翻訳する
(17) 属性	ATTribute	適応性に富む，欠ける，乏しい
(18) 原因・理由	CAUse	事故で死ぬ，～から分かる
(19) 手段・道具	TOOl	イオン法で，ドリルで
(20) 材料	MATerial	ペーストで作る
(21) 構成要素	COMponent	～から成る，～で構成する
(22) 方式	MANner	並列に，10m/secで
(23) 条件	CONdition	焦点深度で決まる
(24) 目的	PURpose	～に適する，備える，必要な
(25) 役割	ROLe	議長に選ぶ，～として用いる
(26) 内容規定	COnTent	～と呼ぶ，述べる，みなす
(27) 範囲規定	RANge	～について，～に関して
(28) 提題	TOPic	～は，～とは
(29) 観点	VIEwpoint	立場から，～の点で
(30) 比較	COmpaRison	～より大きい，～に劣る，～を上回る
(31) 随伴	ACCompaniment	～とともに，～に伴って
(32) 度合	DEGree	5%増加する，3キロやせる
(33) 陳述	PREdicative	～である

Note: The capitalized letters form the English acronym for that case label.

When semantic markers are recorded for nouns in the verbal case frames, each noun appearing in relation to *Yougen* and *Kakujo-shi* in the sample text is referred to the noun lexicon.

The process of describing these case frames for lexicon entry are given in Figure 1.

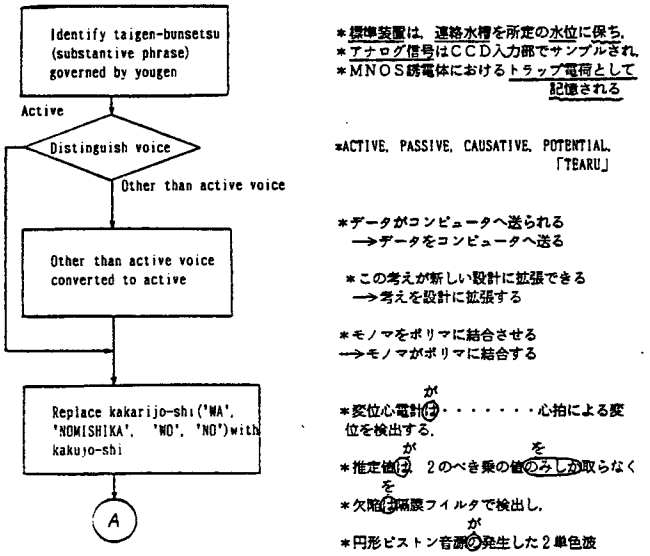
For each verb, *Kakujo-shi* and *Keiyoudou-shi*, *Kakujo-shi* and case labels able to accompany the verb are described, and the semantic marker for the noun which exist antecedent to that *Kakujo-shi* are described.

4. Sub-categories of Parts of Speech according to their Syntactic Features

The parts of speech are classified into 13 main categories:

nouns, pronouns, numerals, affixes, adverbs, verbs, *Keiyou-shi*, *Keiyoudou-shi*, *Rentai-shi* (adnoun), conjunctions, auxiliary verbs, markers and *Jo-shi* (postpositional particles). Each category is sub-classified and divided into 56 sub-categories (see Appendix A); those which are mainly based on syntactic features, and additionally on semantic features.

For example, nouns are divided into 11 sub-categories; proper nouns, common nouns, action nouns 1 (*Sahen-meishi*), action nouns 2 (others), adverbial nouns, *Kakujo-shi-teki-meishi* (noun with case feature), *Setsuzokujo-shi-teki-meishi* (noun with conjunction feature), unknown nouns, mathematical expressions, special symbols and complementizers. Action nouns are classified into *Sahen-meishi* (a noun that can be a noun-plus-*SURU* (doing) composite verb) and other verbal nouns, because action noun 1 is also used as the word stem of a verb.



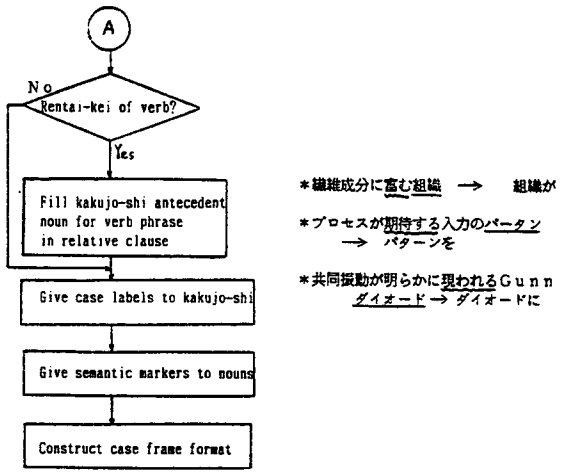


Figure 1. Block Diagram of Process of Describing Verbal Case Frames

Adverbs are divided into 4 sub-categories for modality , aspect and tense. In Japanese, the adverb agrees with the auxiliary verb.

Chinjutsu-fuku-shi agrees with aspect, tense and mood features of specific auxiliary verb,

Joukyou-fuku-shi agrees with aspect and tense,

Teido-fuku-shi agrees with gradability.

Auxiliary verbs are divided into 5 sub-categories based on modality, aspect, voice, cleft sentence and others.

Verbs may be classified according to their case frames and therefore it is not necessary to sub-classify their sub-categories.

5. Semantic Marking of Nouns

We analyze semantic features, and assign semantic markers to Japanese words classified as nouns and pronouns. Each word can give five possible semantic markers.

The system of semantic markers for nouns is made up of 10 conceptual facets based on 44 semantic slots, and 38 plural filial slots at the end(see Figure 2).

5.1 Concept of semantic markers

The 10 conceptual facets are listed below.

1) Thing or Object

This conceptual facet contains things and objects; that is, actual concrete matter. This facet consists of such semantic slots as Nation/Organization, Animate object, Inanimate object, etc.

2) Commodity or Ware

This conceptual facet contains commodity and wares; that is, artificial matter useful to humans. This facet consists of such semantic slots as Material, Means/Equipment, Product, etc.

3) Idea or Abstraction

This conceptual facet contains ideas and abstractions; that is, non-matter as the result of intellectual activity in the human brain. This facet contsists of such semantic slots as Theory, Conceptual object, Sign/Symbol, etc.

4) Part

This conceptual facet contains parts; that is, structural parts, elements and contents of things and matter.

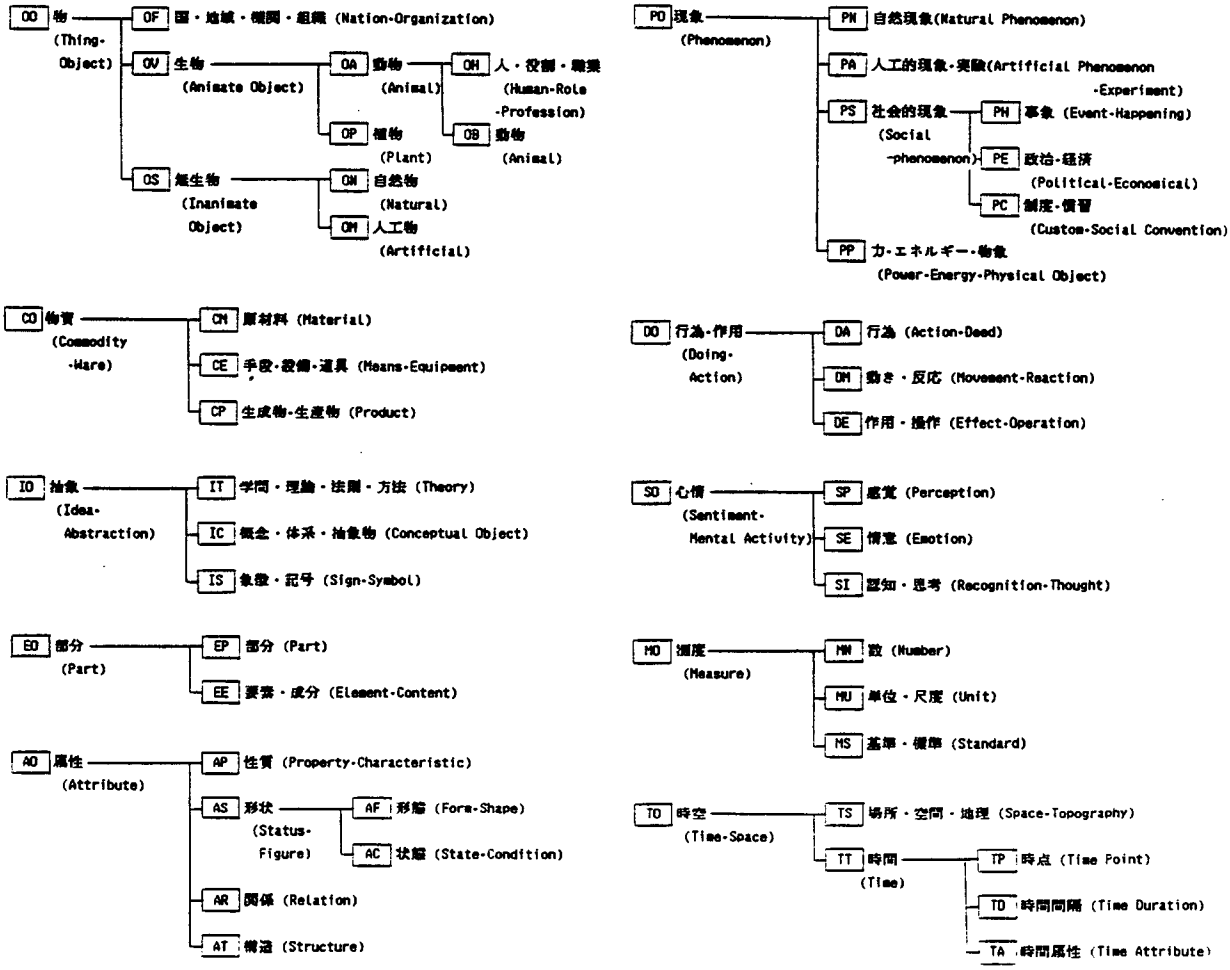


Figure 2. System of Semantic Markers for Nouns

5 Attribute

This conceptual facet contains attributes; that is, properties, qualities or features representative of things. This facet consists of semantic slots such as Property Characteristic, Status Figure, Relation, Structure, etc.

6 Phenomenon

This conceptual facet contains phenomena; that is, physical, chemical and social actions without human activity. This facet consists of semantic slots such as Natural phenomenon, Artificial phenomenon Experiment, Social phenomenon, Power Energy, etc.

7 Doing or Action

This conceptual facet contains human doing and actions. This facet consists of such semantic slots as Action Deed, Movement Reaction, Effect Operation, etc.

8 Mental activity

This conceptual facet contains operations of the mind and mental process. This facet consists of semantic slots such as Perception, Emotion, Recognition Thought, etc.

9) Measure

This conceptual facet contains measure; that is, the extent, quantity, amount or degree of a thing. This facet consists of semantic slots such as Number, Unit, Standard, etc.

10) Time and Space

This conceptual facet contains space, topography and time.

5.2 Process of semantic marking

The semantic marker for each word is determined by the following steps.

1) Determine the definition and features of a word. 2) Extract semantic elements from the word. 3) Judge the agreement between a semantical slot concept and extracted semantical element word by word, and attach the corresponding semantic markers. 4) As a result, one word may have many semantic markers. However, the number of semantic markers for one word is restricted to five. If there are plural filial slots at the end, the higher family slot is used for semantic featurization of the word.

It is easy to decide semantic markers for technical and specific words. But, it is not easy to mark common words, because one word has many meanings.

6. Lexicon Format for Syntactic Analysis

Lexicon formats for syntactic and semantic features are composed of different features classified by part of speech.

1) Features of verb:

Subject code: verb used in specific field, only electrical in our experiment

Part of speech in syntax: verb

Verb pattern: classifing the verbal case frame, a categorized marker like Hornby's case pattern is planned to be used.

Entry to lexical unit of transfer lexicon

Aspect: stative, semi-stative, continuative, resultative, momentary or progressive/transitive

Voice: passive, potential, causative or *'TEARU'* (perfective/stative)

Volition: volitive, semi-volitive or volitionless

Case frame: surface case, deep case, semantic marker for noun and inner-outer case classification

Idiomatic usage: to accompany the verb(ex. catch a cold) syntax, verb pattern.

2) Features of *Keiyou-shi* and *Keiyoudou-shi*:

both syntactic features are described in almost the same format.

Sub-category of part of speech: emotional, property, stative or relative

Gradability: measurability and polarity

Nounness grade: nounness grade for *Keiyou-shi*(++, +, -, --)

3) Features of noun: sub-category of noun(proper, common, action, adverbial, etc), lexical unit for transfer lexicon, semantic markers, thesaurus code, and usage.

4) Features of adverb: sub-category of adverb(*Joukyou*, *Teido*, *Chinjutsu*, *Suuryou*) considering modality, aspect, tense and gradability

5) Features of other taigen: sub-category of *Rentai-shi*(demonstrative, interrogative, definitive, or adjectival) and conjunction(phrase or sentence)

6) Features of *Jodou-shi*(auxiliary verb):

Jodou-shi are sub classified by sub-category on semantic feature:

Modality(negation, necessity, suggestion, prohibition.....)

Aspect(past, perfect, perfective stative, progressive, continuative, finishing, experiential,...)

Voice(passive or causative)

Cleft sentence(purpose and reason)

etc(*'TEMIRU'*, *'TEMISERU'*, *'TEOKU'*, *'SOKONAU'* and *'TEIKERU'*)

7) Features of *Jo-shi*:

Sub-category of *Jo-shi*: case, conjunctive, adverbial, collateral final or *Juntai*

Case: features of surface case(ex. *'GA'* *'WO'* *'NI'* *'TO'*.....), modified relation(*Rentai* or *Renyou* modification)

Conjunctive: sub-category of semantic feature(cause/reason, conditional/provisional, accompanyment, time/place, purpose, collateral, positive or negative conjunction, etc)

7. Data Base Structure of the Lexicon

As this system uses LISP as the programming language, the lexicons are punched up as

S-expressions and input to computer files (see Figure 3).

For the lexicon data base used for syntax analysis, only the lexical items are hold in main storage; syntactic and semantic features are stored in VSAM random acess files on disk(see Figure 4).

```
((S見出し番号 "V0001500-01")
 (S見出し情報
    (S見出し語 "合わせる")
    (S語尾字数 2)
    (S漢字部 1 1)
    (S読み "あわせる")
    (S異形語 "合せる" "併せる" "あわせる"))
 (S形態素情報
    (S形態品詞 動)
    (S動詞活用型 下一)
    (S活用行 ヤ)
    (S前接情報 2)
    (S後接情報 64))
 (S構文ー意味情報
    (S分野コード 電気)
    (S構文品詞 動詞)
    (S格パターン
       V1
       (Sアスペクト 継続)
       (S態 可能 'てある')
       (S意志 有)
       (S備考 "混合させる")
       (S格支配情報
          ((S表層格 が) (S深層格 SUB) (S名詞意味コード OF OH) (S必須性 1))
          ((S表層格 を) (S深層格 OBJ) (S名詞意味コード AS CE) (S必須性 1))
          ((S表層格 に) (S深層格 REC) (S名詞意味コード XX) (S必須性 1))))
    (S格パターン
       V2
       (Sアスペクト 結果)
       (S態 受身 可能)
       (S意志 有)
       (S備考 "合併する")
       (S格支配情報
          ((S表層格 が) (S深層格 SUB) (S名詞意味コード OF OH) (S必須性 1))
          ((S表層格 を)
           (S深層格 OBJ)
           (S名詞意味コード IT IC CO)
           (S必須性 1))
          ((S表層格 と に)
           (S深層格 PAR)
           (S名詞意味コード IT IC CO)
           (S必須性 0))))
    (S格パターン
       V3
       (Sアスペクト 継続)
       (S態 受身 'てある')
       (S意志 有)
       (S格支配情報
          ((S表層格 が) (S深層格 SUB) (S名詞意味コード OF OH) (S必須性 1))
          ((S表層格 に) (S深層格 REC) (S名詞意味コード XX) (S必須性 1)))
       (S共起情報 (S慣用 "気成を")))))
```

Figure 3. Lexicon File Format in LISP S-expression

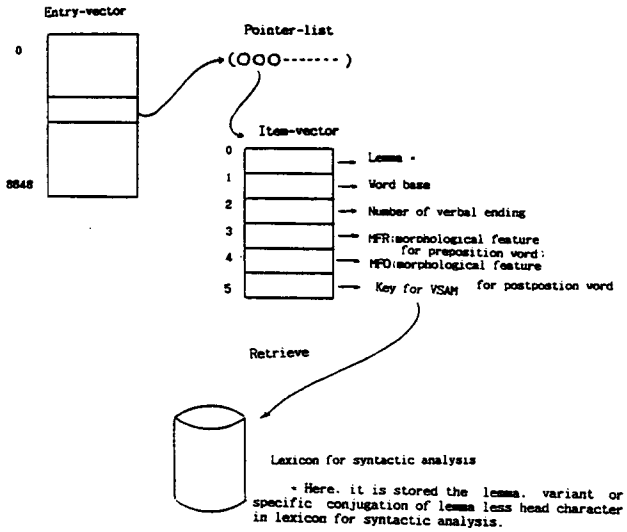


Figure 4. Lexicon Data Base Structure for Analysis

The head character of the lexical unit is used as the record key for the hashing algorithm to generate the addresses in the VSAM files.

8. Conclusion

We have reached the opinion that it is necessary to develop a way of allocating semantic markers automatically to overcome the ambiguities in word meaning confronting the human attempting this task.

In the same thing, there are problems how to find an English term corresponding to the Japanese technical terms not stored in dictionary, how to collect a large number of technical terms effectively and to decide the length of compound words, and how to edit this lexicon data base easily, accurately, safely and speedily.

In lexicon development for a huge volume of *Yougen*, it is quite important that we have a way of collecting automatically many usages of verbal case frames, and we suppose it exist different case frames in different domains.

Acknowledgment

We would like to thank Mrs. Mutsuko Kimura(IBS), Toyo information Systems Co. Ltd., Japan Convention Service Co. Ltd., and the other members of the Mu-project working group for the useful discussions which led to many of the ideas presented in this paper.

References

(1) Nagao, M., Nishida, T. and Tsujii, J.: Dealing with Incompleteness of Linguistic Knowledge on Language Translation, COLING84, Stanford, 1984.

(2) Tsujii, J., Nakamura, J. and Nagao, M.: Analysis Grammar of Japanese for Mu-project, COLING84.

(3) Nakamura, J., Tsujii, J. and Nagao, M.: Grammar Writing System (GRADE) of Mu-Machine Translation Project, COLING84.

In Japanese

(4) Nakai, H. and Satoh, M.: A Dictionary with Taigen as its Core, Working Group Report of Natural Language Processing in Information Processing Society of Japan, WGNL 38-7, July, 1983.

(5) Nagao, M.: Introduction to Mu Project, WGNL 38-2, 1983.

(6) Sakamoto, Y.: Yougen and Fuzoku-go Lexicon in Verbial Case Frame, WGNL 38-8, 1983.

(7) Sakamoto, Y.: Japanese Syntactic Lexicon in Mu project, Proc. of 28th Conference of IPSJ, 1984.

(8) Ishikawa, T., Satoh, M. and Takai, S.: Semantical Function on Natural Language Processing, Proc. of 28th CIPSJ, 1984.

APPENDIX A　Parts of Speech in Japanese Syntax

P.O.S.	Sub-category		Usage
	Japanese	English	
名詞 M	固有名詞	Proper noun	(組織名、人名、地名等)
	普通名詞	Common noun	(例)自動車、山、構造、深層、緑
	動作名詞(サ変)	Action noun 1(Suhen meishi)	「名詞+する」(例)概説、利用、運動
	動作名詞(その他)	Action noun 2(others)	「連用形動名詞」(例)ずれ、ゆれ、ふれ
	副詞的名詞	Adverbial noun	(例)将来、昨年、従来
	格助詞的名詞	Kakujo shi teki-meishi	(例)中、内、前、間、側
	接続助詞的名詞	Setsuzokujo-shi-teki-meishi	(例)ため、ところ、とき、場合、際
	補文標識	Complementizer	(例)こと、もの
	未知語	Unknown noun	
	式	Mathematical expression	(例) f(x),exp
	特殊記号	Special symbols	(例) +,−
代名詞 P	疑問代名詞	Interrogative pronoun	(人、物、場所) (例)だれ、どれ、どこ
	人称代名詞	Personal pronoun	(人) 彼、彼女、わたし
	指示代名詞	Demonstrative pronoun	(物、場所) (例)これ、あれ、それ
数詞 N	数	Number	(例)1、2、三、四、百、千、万
	数量詞	Numeral plural	(例)5cm,10Kg,10cc,100箇所
	冠数詞	Kansuu shi	(例)第、約、昭和
	助数詞	Counter	(例)回、件、階、語、車、部
接辞 Z	接頭語	Prefix 1	(例)諸、各、全、高、低
	接頭辞	Prefix 2	(例)不、非、反
	接尾語	Suffix 1	(例)別、上、中、前
	接尾辞	Suffix 2	(例)的、性、化
記号	括弧	Open parenthesis	(例) (
	括弧	Close parenthesis	(例))
	区切り	Punctration mark	(例) 、。
	つなぎ	Infix	(例) ・、／
副詞 F	情況副詞	Joukyou fuku shi	(文修飾、動詞修飾)
			(例)結局、極力、ここまで
	程度副詞	Teido fuku shi	(尺度、極) (例)非常に、たいへん
	陳述副詞	Chinjutsu fuku shi	(仮定、疑問、打消、願望、比況)
			(例)もし、いつ、必ずしも
	数量副詞	Suuryou fuku shi	(例)たくさん、少し、おのおの
接続詞 S	文接続詞	Sentence connecting conj.	(例)したがって、しかし、ただし
	句接続詞	Phrase connecting conj	(例)または、もしくは、そして

P.O.S.	Sub-category		Usage
	Japanese	English	
連体詞 R	指示連体詞	Demonstrative Rentai shi	(例)この、その、あの
	疑問連体詞	Interrogative Rentai shi	(例)どの、どのような
	限定連体詞	Determinative Rentai shi	(例)ある、さる、あらゆる
	形容詞的連体詞	Adjectival Rentai shi	(例)大きな、小さな、少しの
動詞 D	動詞	Verb	(全ての動詞)
助動詞 A	法助動詞	Modal auxiliary verb	(否定、必要性、勧誘、許可…)
	相助動詞	Aspect auxiliary verb	(相動詞を含む)
	態助動詞	Voice auxiliary verb	(受身、使役)
	分裂構文	Cleft sentence	(目的、理由)
	その他の助動詞	Others	(終助詞的、補助用言的助動詞を含む)
助詞 B	格助詞	Kakujo shi	(例)が、に、と、で、から、より
	接続助詞	Setsuzokujo shi	(例)ば、と、が、のに、ので、から
	副助詞	Fukujo shi	(例)は、も、こそ、さえ、しか、のみ
	並列助詞	Heiretsujo shi	(例)と、や、か
	終助詞	Syuujo shi	(例)ね、さ、か、よ、わ
	準体助詞	Juntaijo shi	(例)の、か、かどうか、か否か
形容詞 K	情意形容詞	Emotional Keiyou shi	(「語幹+がる」で動詞になる)
			(例)うれしい,悲しい
	性質形容詞	Property Keiyou shi	(「もの」の性質を形容する傾向の強い語)(例)固い、柔らかい、新しい、細い
	状態形容詞	Descriptive Keiyou shi	(「属性」や「動作」を形容する傾向の強い語)(例)著しい、早い、強い、高い
	関係形容詞	Relational Keiyou shi	(「こと」、「もの」の間の関係を示す)
			(例) 違い、近い
形容動詞 L	情意形容動詞	Emotional Keiyou dou shi	(「語幹+がる」で動詞になる)
			(例) 愉快だ、残念だ
	性質形容動詞	Property Keiyou-dou shi	(「もの」の性質を形容する傾向の強い語)(例) きれいだ、四角だ、失礼だ
	状態形容動詞	Descriptive Keiyou dou shi	(「属性」や「動作」を形容する傾向の強い語)(例)十分だ、困難だ、正確だ
	関係形容動詞	Relational Keiyou dou shi	(「こと」、「もの」の間の関係を示す)
			(例) 同じ、別