ipadic version 2.7.0 User's Manual

Masayuki Asahara and Yuji Matsumoto

November 2003

IPADIC version 2.7.0 User's Manual

Masayuki Asahara and Yuji Matsumoto

This translation of the IPADIC user's manual was made with support from the non-profit organization GSK by Eric Nichols.

This edition is for "IPADIC for Japanese" version 2.7.0.

version 1.0b	25 May 1998
version 1.0	27 April 1999
version 2.0	15 December 1999
version 2.1	30 December 1999
version 2.4.0	6 December 2000
version 2.5.0	13 April 2001
version 2.6.0	19 June 2003
version 2.7.0	15 November 2003

目 次

Introduction

The ChaSen morphological analyzer was released by Nara Institute of Science and Technology as free software for natural language processing. This manual is for the Japanese dictionary, ipadic 2.7.0 used in ChaSen version 2.3.2 and above. This dictionary is based on the [[IPA Part of Speech Tagset]] (THiMCO97) established by the Information-technology Promotion Agency of Japan (IPA) with some modifications. This manual includes excerpts reproduced with permission and some modification from the [[IPA Part of Speech Tagset]] (THiMCO97) explanation which originally appeared in "The Text Database Report (1996 issue)" published by the Real-World Computing Partnership (RWCP).

Furthermore, the current IPA Japanese part of speech dictionary is ipadic 1.0b2 , as released in May of 1998, with large-scale modification and improvement made by the group members of the "Japanese Speech Dictation Software Development Group" (IPA research and development of original, advanced information technology), represented by Professor Kiyohiro Shikano of the Graduate School of Information Science at Nara Institute of Science and Technology.

We would like to give our heartfelt gratitude to all of the people who participated in the construction of this dictionary system.

Please send any inquiries regarding this manual to the following address.
Computational Linguistics Laboratory Graduate School of Information Science Nara Institute of Science and Technology 8916-5 Takayama, Ikoma, Nara 630-0192, Japan Tel: +81-743-72-5240, Fax: +81-0743-72-5249
E-mail: chasen@is.naist.jp

The latest information about ipadic can be found at the following URL. URL: http://chasen.aist-nara.ac.jp/

1 Installation

1.1 Installing the Dictionary in UNIX

This dictionary requires ChaSen version 2.3.2 or later.
Download and install ChaSen before installing ipadic.
Standard Installation Method

1. Run the `./configure` script

   ```
   %./configure
   ```

   The install directory is also needed by ChaSen, so it is set automatically. If you need to change the install directory, use the `--with-dicdir` flag.

   ```
   % ./configure --with-dicdir=/home/masayu-a
   ```

   Doing so will cause the dictionary to be created under /home/masayu-a/ipadic.

2. Run `make`.

   ```
   % make
   ```

   If compilation fails when using the OS-standard make, GNU make should be used instead.

3. Run `make install` with root permission.

   ```
   # make install
   ```

   By default ipadic is installed into `/usr/local/share/chasen/dic/ipadic` (this may vary from system to system). Root permission is not required to install into the user's home directory.

4. Editing `/usr/local/etc/chasenrc`

   If this is the first time installing ChaSen and Ipadic, the installer will automatically create `/usr/local/etc/chasenrc`. Otherwise, the user will have to create their own chasenrc file. Ipadic's package includes a copy to use as a guide.

1.2 Installing the Dictionary in Windows

The following instructions assume that WinCha is installed in the following location.

```
c:\Program Files\chasen21\dic
c:\Program Files\chasen21\dll
c:\Program Files\chasen21\doc
c:\Program Files\chasen21\mkchadic
c:\Program Files\chasen21\wincha
c:\Program Files\chasen21\wvshell
```

Ipadic is normally automatically installed with WinCha, but when it is installed manually, the user will need to prepare an SJIS-encoded dictionary. The SJIS dictionary package can be found at the following URL.

`http://chasen.aist-nara.ac.jp/stable/ipadic/win/`

Copy the expanded dictionary files (files with the `.dic` extension like `Noun.dic`), part of speech connection file (`cforms.cha`), conjugation type definition file (`ctypes.cha`), conjugation type definition file (`ctypes.cha`), and conjugation form definition file (`cforms.cha`) to the `c:\Program Files\chasen21\dic` inside of the WinCha installation.

Next, copy the `Makefile.bat` file inside the dictionary package to `c:\Program Files\chasen21` and run `Makefile.bat` at the command prompt.

```
C:\Program Files\chasen21> Makefile.bat
```

Under Windows XP/2000/NT and later, Administrator privileges are needed to install the dictionary.

2 The Various File Formats

2.1 Definitions in the Part of Speech Definition File

A list of parts of speech is described in the format file `grammar.cha`. The part of speech categories are organized into hierarchies with the most basic categories at the top and the most detailed categories as the bottom. Parts of speech that inflect have ⟨root categories⟩ marked with a `%`.

For inflectional parts of speech, the possible inflection types must be listed in `ctypes.cha`, and the possible inflected forms must be put in `cforms.cha`.

```
(接頭詞 ; prefix
    (名詞接続) ; nominal prefix
    (動詞接続) ; verbal prefix
    (形容詞接続) ; adjectival prefix
    (数接続)) ; numerical prefix

(動詞% ; verb
    (自立) ; main verb
    (非自立) ; auxiliary verb
    (接尾)) suffix verb
```

- ⟨POS definition⟩ ::= "(⟨top POS information⟩ (⟨lower POS information⟩)*)"
- ⟨top POS category⟩ ::= ⟨top POS definition ⟩|"⟨top POS name⟩%"
- ⟨lower POS definition⟩ ::= ⟨POS category name⟩ | "⟨POS category name⟩ (⟨lower POS information⟩)*"

2.2 Inflection Type Definition File Format

In the inflection types file `ctypes.cha`, the inflection types that each part of speech can take are described.

```
((形容詞 自立) ; main adjective
        (形容詞・アウオ段 ; a-o-u group
         形容詞・イ段 ; i group
         不変化型) ; non-inflectional
)
```

- ⟨inflection type definition⟩ ::= "((⟨POS name⟩) (⟨inflection type⟩*))"

2.3 Inflection Form Definition File Format

In the inflected forms file `cforms.cha`, the inflection types and inflectional suffixes that each part of speech can take are described. The inflectional suffixes can be given in kanji, kana, or pronunciation format.

```
(形容詞・イ段 ; i-adjective
    (   ; (語幹            *            ) ; stem
     (基本形          い      イ    ) ; base form
     (文語基本形      *       *    ) ; written language base form
     (未然ヌ接続      から    カラ  ) ;
     (未然ウ接続      かろ    カロ  ) ;
     (連用タ接続      かっ    カッ  ) ;
     (連用テ接続      く      ク    )
     (連用テ接続      くっ    クッ  )
     (連用ゴザイ接続  ゆう    ユウ  ユー)
     (連用ゴザイ接続  ゆぅ    ユゥ  ユー)
     (体言接続        き      キ    )
     (仮定形          けれ    ケレ  )
     (命令 e          かれ    カレ  )
     (仮定縮約 1      けりゃ  ケリャ)
     (仮定縮約 2      きゃ    キャ  )
     (ガル接続        *            ))
)
```

- ⟨inflection type definition⟩ ::= "(⟨inflected form name⟩ (⟨inflection type information⟩*))"
- ⟨inflection type information⟩ ::= "(⟨inflection type 名⟩ ⟨kanji inflectional suffix⟩ ⟨kana inflectional suffix⟩ ⟨pronunciation inflectional suffix⟩)" | "(⟨inflection type 名⟩ ⟨kanji inflectional suffix⟩ ⟨kana inflectional suffix⟩)" | "(⟨inflection type name⟩ ⟨kanji inflectional suffix⟩)"

2.4 Dictionary File Format

Below is an example dictionary file. The dictionary file is divided by part of speech

```
(品詞 (名詞 一般)) ((見出し語 (お正月 3641)) (読み オショウガツ) (発音 オショーガツ))
; general noun "otsuki"

(品詞 (動詞 自立)) ((見出し語 (あきらめる 2377)) (読み アキラメル) (活用型 一段))
; main verb "akirameru"

(品 詞 (名 詞 一 般)) ((見 出 し 語 (天 文 学 3556)) (読 み テ ン モ ン ガ ク)
; general noun "tenmongaku"
 (複合語 ; compound words
     ((品詞 (名詞 一般)) (見出し語 天文) (読み テンモン)) ; general noun "tenmon"
   ((品詞 (名詞 接尾 一般)) (見出し語 学) (読み ガク)) )) ; general suffix "gaku"
```

The definition of a morpheme in the dictionary is as follows.

- ⟨Morpheme entry⟩ ::= "(⟨POS information⟩) (⟨lexical entry information⟩ ⟨morpheme information⟩*)"
- ⟨POS information⟩ ::= "(品詞 (⟨POS name⟩))"
- ⟨lexical entry information⟩ ::= "(見出し語 (⟨lexical entry⟩ ⟨morpheme occurrence cost⟩))" | "(見出し語 ⟨lexical entry⟩)"
- ⟨morpheme information⟩ ::= ⟨reading information⟩ | ⟨pronunciation information⟩ | ⟨inflection type information⟩ | ⟨additional information⟩ | ⟨semantic information⟩ | ⟨compound word information⟩
- ⟨reading information⟩ ::= "(読み ⟨reading⟩)"
- ⟨pronunciation information⟩ ::= "(発音 ⟨pronunciation⟩)"
- ⟨inflection type information⟩ ::= "(活用型 ⟨inflection type⟩)"
- ⟨compound word information⟩ ::= "(複合語 ⟨compositional word entry⟩*)"
- ⟨compositional word entry⟩ ::= "(⟨POS information⟩ ⟨lexical entry information⟩ ⟨compositional word morpheme information⟩*)"
- ⟨compositional word morpheme information⟩ ::= ⟨reading information⟩ | ⟨pronunciation information⟩ | ⟨inflection type information⟩ | ⟨ additional information⟩ | ⟨semantic information⟩ | ⟨inflected form information⟩
- ⟨inflected form information⟩ ::= "(活用形 ⟨inflected form⟩)"

Furthermore, repetition of items is forbidden inside of "morpheme information" and "compositional word morpheme information" definitions.

- ⟨POS name⟩

  The POS name and each level in its hierarchical structure are separated by whitespace.

  Example:

  ```
  (品詞 (名詞 一般)) ; (POS (noun general))
  (品詞 (動詞 自立)) ; (POS (verb main))
  (品詞 (名詞 接尾 一般)) ; (POS (noun suffix general))
  ```

- ⟨lexical entry⟩

  A list of words that appear in text. Only the basic form of each word is registered.

  Example:

```
(見出し語 (お正月 3641)) ; (entry (otsuki 3641))
(見出し語 (あきらめる 2377)) ; (entry (akirameru 2377))
(見出し語 (天文学 3556)) ; (entry (tenmongaku 3556))
(見出し語 天文)  ; (entry tenmon)
(見出し語 学) ; (entry gaku)
```

- ⟨Morpheme occurrence cost⟩

  The number next to a lexical entry is called its "morpheme occurrence cost." Smaller numbers indicate words that are more likely to appear. The morpheme occurrence costs in Ipadic were calculated based on word occurrence probabilities trained from morphologically analyzed data.

  When users add their own entries, using the same morpheme occurrence cost as a morpheme with a close frequency should have no adverse effect on the morphological analysis results in most cases. If the results are adversely affected , users should try using a smaller morpheme occurrence cost.

  Example:

```
(見出し語 (お正月 3641)) ; (entry (otsuki 3641))
(見出し語 (あきらめる 2377)) ; (entry (akirameru 2377))
(見出し語 (天文学 3556)) ; (entry (tenmongaku 3556))
```

- ⟨Reading⟩

  A list of possible readings for an entry. Readings are given in katakana.

  Example:

```
(読み オショウガツ)  ; (reading oshougatsu)
(読み アキラメル) ; (reading akirameru)
(読み テンモンガク) ; (reading tenmongaku)
(読み テンモン) ; (reading tenmon)
(読み ガク) ; (reading gaku)
```

- ⟨Pronunciation⟩

  A list of possible pronunciations for an entry. Pronunciations are given in katakana.

  Example:

```
(発音 オショーガツ) ; (pronunciation osho-gatsu)
```

- ⟨Inflection type⟩

  Inflectional words require an inflection type. Only the inflection types defined in `ctypes.cha` are permitted.

Example:

```
(活用型 五段・サ行) ; (inflection type go-dan・sa-gyou)
```

- ⟨Inflected form⟩

  Used to given the decomposed entries for a compound word when its morphemes are inflectional and not in base form.

  Example:

```
(活用形 未然ウ接続) ; (inflected form imperfective\_u-connection)
```

- ⟨Additional information⟩

  Used for additional information about a lexical entry. The user may use it unrestricted. It can be used to record information about accent or the part of speech name in other part of speech tagsets.

  Example:

```
(付加情報 アクセント型=4)  ; (additional-information accent type 4)
```

- ⟨Semantic information⟩

  Semantic information for a lexical entry. The user may use it unrestricted. It can be used to record information from a thesaurus or dictionary entry.

  Example:

```
(意味情報 "思い切る。仕方がないと断念する。")  ; (semantic-information "to resign to fate. to give up as a lost cause.")
```

2.5 Connection File Format

Below is an example of the connectivity rules in the part of speech connection file `connect.cha`. A `*` indicates complete compatibility. Rules near end of the file overwrite rules defined earlier in the file. This makes it necessary to write general rules first and follow them with more specific ones.

```
(( (((名詞 固有名詞 人名 姓)   )) ; proper noun surname
   (((名詞 接尾 人名)   )) ) 842) ; noun suffix person

(((((動詞 自立) 五段・ラ行アル 連用形 ))  ; verb  main "go"-dan"ra"-gyou "aru"-modifier
  (((助動詞) 特殊・マス  ))) 604) ; auxiliary verb special "masu"

(((((助詞 接続助詞) * * て))  ; particle conjunctive "te"
  (((助詞 係助詞) * * も)) ; particle dependency "mo"
  (((形容詞 非自立) 形容詞・アウオ段 * よい))) 35) ; adjectives auxiliary "aou"-dan "yoi"
```

- ⟨connection rule entry⟩ ::= "(⟨connection information⟩ ⟨connectivity cost⟩)"
- ⟨connection information⟩ ::= ⟨POS definition⟩ ⟨POS defintion⟩+
- ⟨POS definition⟩+ ::= ⟨POS definition⟩ | ⟨POS definition⟩+
- ⟨POS definition⟩ ::= "(⟨POS information⟩ ⟨inflection type information⟩ ⟨inflected form information⟩ ⟨lexicalized POS rule⟩)" | "(⟨POS information⟩ ⟨inflection type information⟩ ⟨inflected form information⟩)" | "(⟨POS information⟩ ⟨inflection type information⟩)" | "(⟨POS information⟩)"
- ⟨POS information⟩ ::= "(⟨POS name⟩)"
- ⟨inflection type information⟩ ::= "⟨inflection type⟩" | "*"
- ⟨inflected form information⟩ ::= "⟨inflected form⟩" | "*"
- ⟨lexicalized POS rule⟩ ::= "⟨lexicalization POS definition⟩" | "*"

3 The chasenrc Resource File

The chasenrc resource file is used to define the various necessary options for running the ChaSen morphological analyzer.

These definitions are usually kept in `PREFIX/etc/chasenrc`, but they can also be stored in the file '`.chasenrc`' in the user's home directory.

The chasenrc file can also be specified by an option when chasen is initialized.

The following precendence order wil be used to determine which chasenrc file will be loaded when ChaSen is run.

1. (Unix, Windows) the file specified by the `-r` option at initialization time
2. (Unix, Windows) the file set in the `CHASENRC` environment variable
3. (Windows) The chasenrc set in the registry key `chasenrc` in `HKEY_CURRENT_USER\Software\NAIST\ChaSen`
4. (Unix) the `.chasen2rc` file in the user's home directory
5. (Unix) the file `.chasenrc` in the user's home directory
6. (Unix) `PREFIX/etc/chasenrc` (not installed by default)

A list of settings is given below.
Of these settings, "`DADIC`", "UNKNOWN_POS", and "POS_COST" absolutely must be defined.

1. **The grammar file directory setting**

   This setting specifies the directory where the grammar files (`grammar.cha, ctypes.cha, cforms.cha, connect.cha`) reside.

   ```
   (GRAMMAR /usr/local/lib/chasen/ipadic/dic)
   ```

   This setting can be omitted, in which case it is assumed to be the same as the directory that the chasenrc file resides in.

   In the chasenrc file distributed with version 1.01 or later of chasen's dictionary, ipadic, "GRAMMAR" is omitted.

2. **System dictionaries**

   This setting is used to specify double array dictionaries (`chadic.{da,lex,dat}`) omitting the extensions of their file names.

   Multiple dictionary sets may also be specified.

   Relative paths, i.e. paths not starting with "/", are assumed to start in the same directory as the grammar files. Here is an example.

```
(DADIC chadic
       /home/rikyu/mydic/chadic)
```

   In the example below, two sets of dictionaries are read in.

   (a) `chadic.{da,lex,dat}` in the grammar file directory

   (b) `chadic.{da,lex,dat}` in `/home/rikyu/mydic/`

   When dictionary lookups are done, both of the above dictionary sets will be used.

   [1] .

   The setting `DADIC` is used to specify a double array dictionary for Darts.

```
(DADIC chadic)
```

   In the above example, `chadic.da`, `chadic.lex`, and `chadic.dat` in the same directory as the grammar files will be read.

   The maximum number of usable dictionaries is set to 32.

3. **Unknown word part of speech**

   When an unknown word is detected, this setting indicates what part of speech to treat it as while applying ChaSen's connection rules. If multiple parts of speech are given, then the connection rules for each part of speech are applied.

```
(UNKNOWN_POS (名詞 サ変接続))              ; one part of speech
(UNKNOWN_POS (名詞 サ変接続) (名詞 一般))  ; multiple parts of speech
```

4. **Part of speech cost**

   The morphological analyzer calculates analysis precidences as costs. When there is ambiguity while analyzing, the result with the lowest total cost is given precidence.

   The part of speech cost setting is used to define the magnitude of cost associated with each part of speech as well as set the cost of unknown words. Costs must be integer values.

[1] The same morpheme cannot be registered in a single dictionary set multiple times, but a given morpheme may appear in multiple dictionary sets. In this case, there will be duplicates of a morpheme.

```
(POS_COST
        ((*)                1) ; any part of speech -- default cost 1x
        ((未知語)           500) ; unknown words -- cost 500x
        ((名詞)             2) ; nouns -- cost 2x
        ((名詞 固有名詞)    3) ; proper nouns -- cost 3x
)
```

When multiple costs are defined for a part of speech, the last cost is given precedence. In the above example, the cost of nouns (名詞) is 2, but the morpheme cost of proper nouns (名詞-固有名詞) increases to 3. The '`(*)`' setting at the top indicates that the morpheme cost for parts of speech not explicitly defined should be set to 1 (i.e. no change in the total cost of the path). The cost of unknown words is set to 500.

5. **Relative weights of connectivity and morpheme costs**

   The cost in morphological analysis is calculated as the sum of morpheme cost and connectivity cost. This setting lets users assign weights to these two kinds of costs. The cost of an analysis result will be calculated as the sum of each cost multiplied by its weight. If this setting is omitted, it defaults to 1.

```
(CONN_WEIGHT 1)      ; connectivity cost of 1
(MORPH_WEIGHT 1)    ; morpheme cost of 1
```

6. **Cost threshold**

   In the process of morphological analysis, there may be situations where users want to allow all analyses within a beam search cost width. This setting is used to specify a cost width. To ouput all solutions within the cost width, use the `-m` and `-p` options.

```
(COST_WIDTH 0)   ; cost width -- default value
```

   The cost width can also be specified with the `-w` option, overriding the value set in the chasenrc file.

7. **Undefined connectivity cost**

   This setting specifies the connectivity cost for morpheme sequences not defined in the connection rule file. If an undefined connectivity cost is not given, or it is set to 0, then morpheme sequences not in the connection rule file will never be permitted. The default value is 0.

```
(DEF_CONN_COST 500) ; undefined connectivity cost of 500
```

8. **Output format**

   This settings lets users change the output format of ChaSen's results.

```
(OUTPUT_FORMAT "%m\t%y\t%P-\n")
```

   The output format can also be specified using the `-F` flag, overriding any value set in chasenrc. For more information on formatting, see Section **??**.

9. **BOS string**

   The setting specifies the string to display at the beginning of the results for a sentence. Using "%S" will display the entire input sentence. The default is the empty string.

```
(BOS_STRING "Input sentence: [%S]\n") ; BOS string is "Input sentence: [%S]"
```

10. **EOS string**

    The setting specifies the string to display at the end of the results for a sentence. Using "%S" will display the entire input sentence. The default is "EOS\n".

```
(EOS_STRING "END\n") ; EOS string is "END"
```

11. **Whitespace part of speech**

    ChaSen treats the halfspace whitespace character (ASCII code 32) and tab (ASCII 9) as whitespace and ignores them during analysis. Normally whitespace information is not included in ChaSen's output, but this can be changed by using the "SPACE_POS" setting. For example, the setting given below will output "punct-whitespace" for whitespace.

```
(SPACE_POS (punct-whitespace)) ; whitespace part of speech is "punct-whitespace"
```

    Furthermore, by setting the output format to "%m" and specifying a whitespace part of speech, uesrs can get output that is corresponds exactly to the input sentence, whitespace included.

12. **Annotations**

    This setting allows strings that begin and end with a certain sequence to be treated as an annotation and ignored during morphological analysis. In the results, the annotation string will be output as a single morpheme.

    Each annotation definition consists of a list of a start string and stop string followed by optional part of speech information or a formatting string. The stop string can also be omitted, in which case the start string itself will be treated as the annotation. If the part of speech information and format string are omitted, then absolutely no information about the annotation's morpheme will be output.

```
(ANNOTATION (("<" ">") "%m\n") ; output as is
      (("「") (記号 一般)) ; punctuation
      (("」") (記号 一般)) ; punctuation
      (("\"" "\"") (名詞 引用文字列)) ; noun quotation sting
      (("[" "]")) ; nothing will be output
      )
```

    For example, when using the above annotation definition, ChaSen will output its results in the following format.

    - text starting with "¡" and ending with "¿", such as <img src="cha.gif">, will be output as is
    - 記号-一般 will be output for "「" and "」"

- 名詞 引用文字列 will be output for strings in double quotes like "hello (again)"
- strings enclosed in square brackets like [ChaSen] will be ignored in morphological analysis and no information will be included in its output

13. **Part of speech concatenation**

This setting is used to concatonate together morphemes of certain parts of speech that appear in succession and output them as a single morpheme.

```
(COMPOSIT_POS ((複合名詞) (名詞) (接頭詞 名詞接続) (接頭詞 数接続))
  ((記号)))
```

For example, with the above declaration of COMPOSIT_POS, parts of speech are concatonated together in the following manner.

(a) Consecutive nouns (名詞), noun prefixes (接頭詞-名詞接続), numeric prefixes (接頭詞-数接続) are concatenated together and displayed as "compound noun (複合名詞)." However, this part of speech must be defined in the part of speech definition file `grammar.cha`.

(b) Consecutive punctuation (記号) is concatenated together, and displayed as "punctuation (記号)."

14. **Compound word output**

ChaSen can be configured to treat compound words defined in the morphological dictionary file in (`.dic`) two different ways.

(a) compound (複合語): the morphological information for the entire compound word is output

(b) compositional (構成語): the compound word is decomposed into individual words, and the morphological information for eachword is output

The default setting is "compound (複合語)."

```
(OUTPUT_COMPOUND "複合語") ; output compound morphological information
```

Compound word output can also be controlled by the `-Oc` and `-Os` options.

15. **Delimiters**

This setting allows users to define the characters that are used as sentence delimiters when the `-j` option is set (see **??**). Both half-width and full-width characters can be used as delimiters. For example, the following definition treats the full-width characters "。．、，!?" , the half-width characters "`.,!?␣`", and whitespace as sentence delimiters.

```
(DELIMITER "。．、，!?.,!? ")
```

16. **Encodings**

The character encoding that ChaSen supports can be changed by reencoding the morphological file and recompiling ChaSen. The ENCODE setting is used to indicate the encoding that ChaSen will use. For example, the following definition denotes Unicode.

```
(ENCODE "u")
```

The supported encodings are e: EUC-JP, s:Shift_JIS, w:UTF-8, u:UTF-8, a:ISO-8859-1.

4 Adding Morphological Entries

4.1 Editing the Various Files

Download and unzip either `ipadic-X.X.X.tar.gz` or `ipadic-sjis-X.X.X.zip`.
These files can be found at the following location.

- http://chasen.aist-nara.ac.jp/stable/ipadic/
- http://chasen.aist-nara.ac.jp/stable/ipadic/win/

Add new entries following the aforementioned formats.

- `*.dic`
  morpheme dictionaries
- `connect.cha`
  part of speech connections
- `grammar.cha`
  part of speech definitions
- `ctypes.cha`
  inflection type definitions
- `cforms.cha`
  inflected form definitions

4.2 Recompiling System Dictionaries under UNIX

Whenever a change is made to the part of speech tagset or the morpheme dictionary is edited, the dictionaries need to be recompiled.

1. Run `./configure`.

   To change the default install location, run `./configure` in the following manner.

   ```
   % ./configure --with-dicdir=/home/masayu-a
   ```

2. Run `make`.

   ```
   % make
   ```

   If compilation fails when using the OS-standard make, GNU make should be used instead.

3. Run `make install` with root permission.

```
# make install
```

By default ipadic is installed into `/usr/local/share/chasen/dic/ipadic` (this may vary from system to system). Root permission is not required to install into the user's home directory.

4.3 Recompiling User Dictionaries under UNIX

A user dictionary can be used for simple vocabulary additions that do not involve changes to the part of speech tagset.

First, create a directory for the user dictionary.

After adding a file morpheme dictionary that has a file name with extension `.dic`, run the following command.

```
% mkdir ~/mydic
% cd ~/mydic
% emacs Noun2.dic (形態素情報を記述)
$ ‘chasen-config --mkchadic‘/makeda -i e chadic *.dic
```

The `-i` option set on `makeda` indicates the dictionary's character encoding. The following 4 encoding are supported: e:EUC-JP, s:Shift_JIS, w:UTF-8, a:ISO-8859-1.

```
% chasen-config --mkchadic
```

Next make a copy of `chasenrc` in your home directory named `.chasenrc`.

```
% cd
% cp /usr/local/share/chasen/dic/ipadic/chasenrc .chasenrc
```

Edit `.chasenrc` setting "GRAMMAR" and adding the user dictionary to "DADIC."

```
(GRAMMAR /usr/local/share/chasen/dic/ipadic)
(DADIC chadic
        /home/masayu-a/mydic/chadic)
```

4.4 Recompiling Dictionaries under Windows

Copy the unzipped files to the `dic` directory in WinCha's install directory.

After adding new entries, run `Makefile.bat` from the command prompt.

```
C:\Program Files\chasen21> Makefile.bat
```

5 The IPA Part of Speech Tagset

Format Explanation

Part of Speech Names

We will refer to the names of parts of speech as "tags" through the rest of this document. In the various part of speech explanations, the following symbols are used for annotation.

\# Part of speech explanation

例: Example words

1 Notes on the part of speech explanation

& Notes on reading or inflection

Areas of Caution Regarding Part of Speech Names

Ipadic is based on the IPA part of speech tagset (THiMCO97), but we had to make some changes to use it in ChaSen. The characteristics and changes made to Ipadic's part of speech tagset are summarized below.

- Parts of speech are organized into a stratified hierarchy. For example, 「名詞 固有名詞 人名 姓 (noun common person-name surname)」 is the name of a fourth level part of speech. In the rest of this explanation, we will join together hierarchical part of speech names with hyphens: 「名詞-固有名詞-人名-姓」. ChaSen versions 2.0 and later support definition of part of speech hierarchies with an arbitrary number of levels. These definitions can be added directly to the grammar file (`grammar.cha`).
- In THiMCO97, part of speech categories and inflection types and forms were mixed together in definitions:「動詞 一段 連用形 自立 (verb 1-dan stem-form main)」. In ChaSen, the definitions for part of speech categories and inflections are separate, so we divide definitions into the the items (part of speech name, inflection type, inflected form) like so: 「動詞-自立 一段 連用形 (verb-main 1-dan stem-form) 」.
- We changed the category names used to define parts of speech following the below criteria.
  1. We deleted all parentheses from part of speech names: 「(助動詞語幹)」→ 「助動詞語幹」.
  2. We eliminated redundant 「(助動詞) (verb-aux)」: 「動詞 接尾 (助動詞)」「形容詞 接尾 (助動詞)」→「動詞-接尾」「形容詞-接尾」.
  3. In THiMCO97 verbs are roughly divided into the categories「動詞 (verb)」「動詞 非自立 (auxiliary verb)」 and 「動詞 接尾 (suffix verb)」, but in ChaSen's part of speech hierarchy,「動詞」 always indicates a verb, so we add the classification 「自立 (main)」: 「動詞-自立 (verb-main)」「動詞-非自立 (verb-auxiliary)」「動詞-接尾 (verb-suffix)」

     Likewise, we renamed category names for non-inflectional words like 「名詞 (noun)」「名詞 固有名詞 (noun proper)」「名詞 固有名詞 人名 (noun proper person-name)」「名詞 固有名詞 人名 姓 (noun proper person-name surname)」 to remove category overlap by adding the sub-category 「一般 (general)」: 「名詞-一般」「名詞-固有名詞-一般」「名詞-固有名詞-人名-一般」「名詞-固有名詞-人名-姓」.

4. In THiMCO97, inflected forms are defined in detail by the categorizing the auxiliary verb that follows the inflected word like so: 「未然ナイ接続 (imperfective nai-connection)」「未然レル接続 (imperfective reru-connection)」「未然ウ接続 (imperfective u-connection)」「連用タ接続 (conjunctive ta-connection)」「連用マス接続 (conjunctive masu-connection」「連用タイ接続 (conjunctive tai-connection)」···, but in the individual inflection types few words have an ending other than imperfective or conjunctive. So we use「未然形 (imperfective form)」「連用形 (conjunctive form)」「基本形 (basic form)」「仮定形 (subjunctive form)」「命令 (imperative form)」 as the basic categories of inflected forms, and only follow THiMCO97's naming conventions for forms with exceptions. Furthermore, since ChaSen's dictionary is set up to use basic form for its entries, we renamed THiMCO97's 「見出し形 (entry form)」 inflected form name to 「基本形 (basic form)」.
5. The「未然ウ接続 (imperfective u-connection)」 inflected form is defined so that the auxiliary verb 「う 'u'」attaches to 5-dan verbs but 「よう 'you'」 attaches to all other verbs. Here only 「う 'u'」 is recognized as a word; the 「よ 'yo'」 in 「来よ (う)」 and 「食べよ (う)」 is treated as the part of the inflected word.

In Ipadic version 2.0, a pronunciation field was added to words in the dictionary. This information was added thanks to the efforts of the Japanese Speech Dictation Software Development Group." For example, the dependency particle 「は」 has the reading "wa," and「常識」 has the reading "jo-shiki" with the long vowel represented by "-". Also, for words where the orthography and part of speech are the same and only the readings differ, like in the case of「私 (ワタシ/ワタクシ) (watashi/watakushi)」, all of the possible readings are collected into { ワタシ/ワタクシ } and registered as one entry.

5.1 名詞 (Nouns)

5.1.1 名詞-一般 (noun-common)

Common nouns or nouns where the sub-classification is undefined.

5.1.2 名詞-固有名詞-一般 (noun-proper-misc)

miscellaneous proper nouns or proper nouns where the sub-classification is undefined.

5.1.3 名詞-固有名詞-人名-一般 (noun-proper-person-misc)

names that cannot be divided into surname and given name; foreign names; names where the surname or given name is unknown

例: 「お市の方」

5.1.4 名詞-固有名詞-人名-姓 (noun-proper-person-surname)

Mainly Japanese surnames.

例: 「山田」…

5.1.5 名詞-固有名詞-人名-名 (noun-proper-person-given_name)

Mainly Japanese given names.

例: 「太郎」…

5.1.6 名詞-固有名詞-組織 (noun-proper-organization)

Names representing organizations.

例: 「通産省」「NHK」…

5.1.7 名詞-固有名詞-地域-一般 (noun-proper-place-misc)

Place names excluding countries.

例: 「アジア」「バルセロナ」「京都」

5.1.8 名詞-固有名詞-地域-国 (noun-proper-place-country)

Country names.

例: 「日本」「オーストラリア」…

5.1.9 名詞-代名詞-一般 (noun-pronoun-misc)

Pronouns.

例: 「それ」「ここ」「あいつ」「あなた」「あちこち」「いくつ」「どこか」「なに」「みなさん」「みんな」「わたくし」「われわれ」…

5.1.10 名詞-代名詞-縮約 (noun-pronoun-contraction)

Spoken language contraction made by combining a pronoun and the particle 'wa.'

例: 「ありゃ」「こりゃ」「こりゃあ」「そりゃ」「そりゃあ」

5.1.11 名詞-副詞可能 (noun-adverbial)

Temporal nouns such as names of days or months that behave like adverbs. Nouns that represent amount or ratios and can be used adverbially.

例: 「金曜」「一月」「午後」「少量」…

1 In the original IPA part of speech tagset, the distinction was made between whether a word was used adverbially in an actual usage (「名詞 副詞可能 副詞的」) or not (「名詞 副詞可能」) and it was classified accordingly, but for ChaSen, we classify all nouns that can be used adverbially into a single category.

5.1.12 名詞-サ変接続 (noun-verbal)

Nouns that take arguments with case and can appear followed by 'suru' and related verbs (「する」「できる」「なさる」「くださる」)

例: 「インプット」「愛着」「悪化」「悪戦苦闘」「一安心」「下取り」…

1 Onomatopoeia(+suru) is classified as 「副詞-助詞類接続 (adverb-particle_conjunction)」.

1 When a word is considered to have usages as both 「名詞-一般 (noun-common)」 and 「名詞-サ変接続 (noun-verbal)」, this category is given precedence.

5.1.13 名詞-形容動詞語幹 (noun-adjective-base)

The base form of adjectives: words that appear before 「な ('na')」.

例: 「健康」「安易」「駄目」「だめ」…

1 In the original IPA part of speech tagset, these were called 「名詞 (形容動詞語幹)」. We removed the parentheses on the second level part of speech category name.

1 When a word is considered to have usages as both 「名詞-一般 (noun-common)」and 「名詞-形容動詞語幹 (noun-adjective-base)」, this category is given precedence. However, in the case of 「自然」 and 「自然な」, which roughly have the meaning "nature," the meanings and grammatical forms differ, so 「自然」 is registered as 「名詞-一般 (noun-common)」 and 「自然な」 as 「名詞-形容動詞語幹 (noun-adjective-base)」.

5.1.14 名詞-ナイ形容詞語幹 (noun-nai_adjective)

Words that appear before the auxiliary verb 「ない ('nai')」 and behave like an adjective.

例: 「申し訳」「仕方」「とんでも」「違い」…

1 In the original IPA part of speech tagset these were treated as adjectives, but since they are derivational in nature like in the case of 「申し訳-ない」「申し訳-ありません」「申し訳-ございません」, we group all variations under the base form. However, not every word classified as 「ナイ形容詞語幹 (noun-nai_adjective)」 has all possible forms.

5.1.15 名詞-数

Arabic numbers, Chinese numerals, and counters like 「何 (回)」「数 (

例: 「0」「1」「2」「何」「数」「幾」…

5.1.16 名詞-非自立-一般 (noun-affix-misc)

Of adnominalizers, the case-marker 「の ("no")」, and words that attach to the base form of inflectional words, words that cannot be clisified into any of the other categories below. This category includes indefinite nouns.

1 If it can be used as a common noun, even if it takes a restrictive modifier it is not an 「非自立 (affix)」.

例: 「あかつき」「暁」「かい」「甲斐」「気」「きらい」「嫌い」「くせ」「癖」「こと」「事」「ごと」「毎」「しだい」「次第」「順」「せい」「所為」「ついで」「序で」「つもり」「積もり」「点」「どころ」「の」「はず」「筈」「はずみ」「弾み」「拍子」「ふう」「ふり」「振り」「ほう」「方」「旨」「もの」「物」「者」「ゆえ」「故」「ゆえん」「所以」「わけ」「訳」「わり」「割り」「割」「ん-口語/」「もん-口語/」…

5.1.17 名詞-非自立-副詞可能 (noun-affix-adverbial)

Of adnominalizers, the case-marker "no" and words that attach to the base form of inflectional words, words that can behave as adverbs.

1 In the original IPA part of speech tagset, words that were actually used as adverbs in a sentence were tagged 「名詞-非自立-副詞可能-副詞的」 , however, we omit the final tag.

例: 「あいだ」「間」「あげく」「挙げ句」「あと」「後」「余り」「以外」「以降」「以後」「以上」「以前」「一方」「うえ」「上」「うち」「内」「おり」「折り」「かぎり」「限り」「きり」「っきり」「結果」「ころ」「頃」「さい」「際」「最中」「さなか」「最中」「じたい」「自体」「たび」「度」「ため」「為」「つど」「都度」「とおり」「通り」「とき」「時」「ところ」「所」「とたん」「途端」「なか」「中」「のち」「後」「ばあい」「場合」「日」「ぶん」「分」「ほか」「他」「まえ」「前」「まま」「儘」「侭」「みぎり」「矢先」…

5.1.18 名詞-非自立-助動詞語幹 (noun-affix-aux)

1 Of adnominalizers, the case-marker "no" and words that attach to the base form of inflectional words, words treated as「助動詞 ("auxiliary verb")」in school grammars with the stem「よう (だ) ("you(da)"」.

例: 「よう」「やう」「様 (よう)」

1 In the original IPA part of speech tagset, this category was written as 「名詞-非自立-(助動詞語幹)」.

5.1.19 名詞-非自立-形容動詞語幹 (noun-affix-adjective-base)

1 Of adnominalizers, the case-marker "no" and words that attach to the base form of inflectional words, words that can connect to the indeclinable connection form, 「な (aux "da")」.

例: 「みたい」「ふう」

1 In the original IPA part of speech tagset, this category was written as 「名詞 非自立 (形容動詞語幹)」.

5.1.20 名詞-特殊-助動詞語幹 (noun-special-aux)

The 「そうだ ("souda")」stem form that is used for reporting news, is treated as 「助動詞 ("auxiliary verb")」 in school grammars, and attach to the base form of inflectional words.

例: 「そう」

1 In the original IPA part of speech tagset, this category was written as 「名詞 特殊 (助動詞語幹)」.

5.1.21 名詞-接尾-一般 (noun-suffix-misc)

Of the nouns or stem forms of other parts of speech that connect to 「ガル」 or 「タイ」 and can combine into compound nouns, words that cannot be clisified into any of the other categories below. In general, this category is more inclusive than 「接尾語 (”suffix”)」 and is usually the last element in a compound noun.

例: 「おき」「かた」「方」「甲斐 (がい)」「がかり」「ぎみ」「気味」「ぐるみ」「(～した) さ」「次第」「済 (ず) み」「よう」「(でき)っこ」「感」「観」「性」「学」「類」「面」「用」…

5.1.22 名詞-接尾-人名 (noun-suffix-person)

Suffixes that form nouns and attach to person names more often than other nouns.

例: 「君」「様」「著」など.

5.1.23 名詞-接尾-地域 (noun-suffix-place)

Suffixes that form nouns and attach to place names more often than other nouns.

例: 「町」「市」「県」など.

5.1.24 名詞-接尾-サ変接続 (noun-suffix-verbal)

Of the suffixes that attach to nouns and form nouns, those that can appear before 「スル (”suru”)」.

例: 「化」「視」「分け」「入り」「落ち」「買い」

5.1.25 名詞-接尾-助動詞語幹 (noun-suffix-aux)

The stem form of 「そうだ (様態)」 that is used to indicate conditions, is treated as 「助動詞 (”auxiliary verb”)」 in school grammars, and attach to the conjunctive form of inflectional words.

例: 「そう」

1 In the original IPA part of speech tagset, this category was written as 「名詞 接尾 (助動詞語幹)」.

5.1.26 名詞-接尾-形容動詞語幹 (noun-suffix-adjective-base)

Suffixes that attach to other nouns or the conjunctive form of inflectional words and appear before the copula 「だ (”da”)」.

例: 「的」「げ」「がち」

1 In the original IPA part of speech tagset, this category was written as 「名詞 接尾 (形容動詞語幹)」.

5.1.27 名詞-接尾-副詞可能 (noun-suffix-adverbial)

Suffixes that attach to other nouns and can behave as adverbs.

1 In the original IPA part of speech tagset, the distinction was made between whether a word was used adverbially in an actual usage or not and it was classified accordingly, but we classify all noun suffixes that can be used adverbially into this category.

例: 「後 (ご)」「以後」「以降」「以前」「前後」「中」「末」「上」「時 (じ)」

5.1.28 名詞-接尾-助数詞 (noun-suffix-classifier)

Suffixes that attach to numbers and form nouns. This category is more inclusive than 「助数詞 ("classifier")」 and includes common nouns that attach to numbers.

例: 「個」「つ」「本」「冊」「パーセント」「cm」「kg」「カ月」「か国」「区画」「時間」「時半」…

1 In the IPA part of speech tagset, words used adverbially were tagged indicating that usage, but this tagset does not include that information.

5.1.29 名詞-接尾-特殊 (noun-suffix-special)

1 A new category defined for special suffixes that mainly attach to inflecting words.

例: 「(楽し) さ」「(考え) 方」

2 In the original IPA part of speech tagset, this was classified as 「名詞 接尾 ("noun suffix")」.

5.1.30 名詞-接続詞的 (noun-suffix-conjunctive)

Nouns that behave like conjunctions and join two words together.

例: 「(日本) 対 (アメリカ)」「対 (アメリカ)」「(3) 対 (5)」「(女優) 兼 (主婦)」

5.1.31 名詞-動詞非自立的 (noun-verbal_aux)

Nouns that attach to the conjunctive particle 「て ("te")」 and are semantically verb-like.

例: 「ごらん」「ご覧」「御覧」「頂戴」

Caution In the IPA part of speech tagset, 「名詞 引用文字列 ("noun quoatation")」 is used to represent text that cannot be segmented into words, proverbs, Chinese poetry, dialects, English, etc. The tag 「名詞 数式 ("noun mathematical formula"」 is used for mathematical formulae. These tags are hard to think of as parts of speech, and we take the position of not formally supporting them in out tagset. Currently, the only entry for 「名詞 引用文字列 ("noun quoatation")」 is 「いわく ("iwaku")」.

5.2 接頭詞 (prefix)

5.2.1 接頭詞-名詞接続 (prefix-nominal)

Prefixes that attach to nouns (including adjective stem forms) excluding numerical expressions.

例: 「お (水)」「某 (氏)」「同 (社)」「故 (〜氏)「高 (品質)」「お (見事)」「ご (立派)」

5.2.2 接頭詞-数接続 (prefix-numerical)

Prefixes that attach to numerical expressions.

例: 「約」「およそ」「毎時」など

5.2.3 接頭詞-動詞接続 (prefix-verbal)

Prefixes that attach to the imperative form of a verb or a verb in conjunctive form followed by 「なる/なさる/くださる」.

例: 「お (読みなさい)」「お (座り)」

5.2.4 接頭詞-形容詞接続 (prefix-adjectival)

Prefixes that attach to adjectives.

例: 「お (寒いですねえ)」「バカ (でかい)」

5.3 動詞 (verb)

Words of caution regarding inflected forms

未然形 (imperfective form) In THiMCO97, this form was divided into the subcategories listed below, but we unite them into 「未然形 (imperfective form)」 whenever there is no change in the inflection itself.

- Imperfective reru-connection form

  # Forms that attach to (ラ) レル, (サ) セル

  例: 「読ま」「さ」…

- Imperfective nai-connection form

  # Forms that attach to ナイ.

  例: 「読ま」「し」…

- Imperfective nu-connection form

  # Forms that attach to ヌ, (サ) シメル.

  例: 「読ま」「せ」「来」…

- Imperfective u-connection form

  # Forms that attach to (ヨ) ウ.

  例: 「読も」「し」…

  & In ipadic1.0 and later defined as those verb forms that attach to the auxiliary verb "u." For example, "shiyo" is the imperfective u-connector form for the verb "suru."

連用形 (conjunctive form) All conjunctive forms are united under this name except for those with irregular suffixes.

- Conjunctive masu-connection form

  # Forms that attach to マス.

  例: 「読み」「し」「なさい」…

- Conjunctive tai-connection form

  # Forms that attach to タイ, ソウ, ヅライ, 方 (かた), 読点など.

  例: 「読み」「し」「なさり」「向かひ」「習ひ」…

- Conjunctive ta-connection form

  # Forms that attach to タ, テ.

  例: 「読ん」「書い」「行っ」「問う」…

基本形 (basic form) Known as 「見出し形 (dictionary form)」 in THiMCO97.

Forms that attach to punctuation, uninflected words, マイ, etc.

例: 「読む」「なさる」「問う」…

仮定形 (conditional) Known as 「仮定バ接続 (conditional ba-connection form)」 in THiMCO97.

Forms that attach to バ, ドモ.

例: 「読め」「すれ」…

命令 i (imperative "i") # The imperative form of irregular "kuru" verbs and the spoken form of the imperative form of "suru."

例: 「来い」「なさい」「せい」…

命令 e (imperative "e") # The imperative form of group 5 verbs and the stem of the group 1 verb imperative 止め "stop" ("kure" only).

例: 「読め」「(とは) いえ」「(程度の差こそ) あれ」「(やめて) くれ」…

1 「(やめて) くれ」 is the result of dropping 「ろ」 from 「(やめて) くれろ」. 「くれる」 has special inflected forms for a group 1 verb and needs to be treated specially. In addition, the 「くれ」 in 「(やめて)(お) くれ (なさい)」 is classified as 「動詞-非自立 一段連用タイ接続 (verb-aux 1-dan conjunctive-tai-connection-form)」.

命令 yo (imperative "yo") # Imperative form for 一段・サ変・文語 (カ変) that ends in "yo."

例: 「せよ」「みよ」「来よ」…

命令 ro (imperative "ro") # Imperative form for 一段・サ変 that ends in "ro."

例: 「しろ」「みろ」…

ベキ接続 (beki-connection) # The form that is followed by "beki." Only for サ変.

例: 「す」…

仮定縮約 1 (conditional contracted form 1) 3 The shortened form produced by combining "ba" and the conditional ba-connection form (spoken language).

例: 「分かれりゃ」

体言接続 (Uninflected word connection form)

Only for written words that have an irregular dictionary form.

例: 「助くる」(cf.「助く」)

体言接続特殊 (Uninflected word connection special form) # For words that end in "ru" and undergo euphonic change when connecting to "no" (spoken language).

例: 「(何) すん (の?)」

体言接続特殊 2 (Uninflected word connection special form 2) # For verbs like ”kuru,” ”suru,” and ”toru” where the final ”n” is dropped from the uninflected word connection special form (spoken language).

Verb Inflection Types (Modern Language)

【infl.】 indicates a category that is an inflected form.

5.3.1 動詞-自立 カ変 (verb-main kuru) 【infl.】

例: 「くる」「来る」「やってくる」「やって来る」

5.3.2 動詞-非自立 カ変 (verb-aux kuru) 【infl.】

例: 「(て) くる」「(て) 来る」

5.3.3 動詞-自立 サ変・スル (verb-main suru) 【infl.】

The ”suru” that connects to verbal nouns.

例: 「する」

5.3.4 動詞-自立 サ変・ スル (verb-main suru) 【infl.】

和語系のサ変動詞.

例: 「接する」…

1 「 し+ない」「 せ+られる」「 せ+ぬ」「 し+よう」「 する」「 すれ+ば」「 せよ」「 しろ」are the only forms classified as 「動詞-自立 サ変 (verb-main suru)」. Other conjunctive forms like 「 し+,」「 し+た」「 し+たい」 are classified as group 5 consonant-s verbs.

5.3.5 動詞-自立 サ変・ ズル (verb-main suru) 【infl.】

和語系のザ変動詞.

例: 「信ずる」…

1 「 ぜ+られる」「 ぜ+ぬ」「 ずる」「 ずれ+ば」「 ぜよ」「 ず+べし」are the only forms classified as 「動詞-自立 サ変 (verb-main zuru)」. Other conjunctive forms like 「 じ+ない」「 じ+よう」の未然形および「 じ+,」「 じ+た」「 じ+たい」and the imperative form「 じろ」 are classified as group 1 verbs.

5.3.6 動詞-自立 一段 (verb-main group-1)【infl.】

Verbs that have only one inflection type.

例: 「着る」

1 「病める」 only has the base form.

5.3.7 動詞-非自立 一段 (verb-aux group-1) 【infl.】

例: 「あげる」「うる」「える」「得る」「おえる」「終える」「おおせる」「かねる」「兼ねる」「かける」「きれる」「切れる」「すぎる」「過ぎる」「そこねる」「損ねる」「そびれる」「そめる」「初める」「つける」「つづける」「続ける」「(お読み) できる」「(お読み) 出来る」「はじめる」「始める」「(て) いる」「(～しては) いけ (ない)」「(て) くれる」「(て) 差し上げる」「(て) のける」「(て) みる」「(て) みせる」[(て) もらえる」「(て) る-口語/」

1 The base form of 「(～しては) いけ (ない)」 is 「いける」.

1 「(勉強) できる」is not classified as an auxiliary verb.

1 「うる」 only has base and conditional forms. It is classified as 「動詞 文語 基本形 (verb written basic-form)」.

5.3.8 動詞-接尾 一段 (verb-suffix group-1)【infl.】

In school grammar, this is classified as an auxiliary verb.

例: 「させる」「せる」「しめる」「しむる」「られる」「れる」

5.3.9 動詞-自立 五段・カ行イ音便 (verb-main group-5 consonant-k i-onbin)【infl.】

Group 5 consonant-k verbs that undergo ki→i euphonic change when attaching to "te."

例: 「解く」「聞く」…

5.3.10 動詞-非自立 五段・カ行イ音便 (verb-aux group-5 consonant-k i-onbin) 【infl.】

例: 「つづく」「続く」「ぬく」「抜く」「(て) いただく」「(て) 頂く」「(て) おく」「とく-口語/」「どく-口語/」

5.3.11 動詞-非自立 五段・カ行促音便 (verb-aux group-5 consonant-k consonant-onbin) 【infl.】

Group 5 consonant-k verbs that undergo consonant-assimilation euphonic change when attaching to "te."

例: 「いく」「行く」「ゆく」

1 "yuku" has no corresponding form "yut(te)," but we classify it in this group anyway. "yuki(te)" is classified as 「動詞 文語 連用タ接続 (verb written conjunctive-ta-connection-form)」.

5.3.12 動詞-非自立 五段・カ行促音便 (verb-aux group-5 consonant-k consonant-onbin) 【infl.】

例: 「いく」「行く」「ゆく」「く-口語/」

1 "yuku" has no corresponding form "yut(te)," but we classify it in this group anyway. "yuki(te)" is classified as 「動詞 文語 連用タ接続 (verb written conjunctive-ta-connection-form)」.

5.3.13 動詞-自立 五段・ガ行 (verb-main group-5 consonant-g) 【infl.】

Group 5 consonant-g verbs that undergo gi→i euphonic change when attaching to "te."

例: 「継ぐ」「急ぐ」…

5.3.14 動詞-自立 五段・サ行 (verb-main group-5 consonant-s) 【infl.】

Group 5 consonant-s verbs that do not undergo euphonic change when attaching to "te."

例: 「話す」…

5.3.15 動詞-非自立 五段・サ行 (verb-aux group-5 consonant-s) 【infl.】

例: 「いたす」「致す」「だす」「出す」「つくす」「尽くす」「直す」

5.3.16 動詞-自立 五段・タ行 (verb-main group-5 consonant-t) 【infl.】

五段タ行で,[助詞 接続助詞] の「て」に接続するときに促音便になるもの.

Group 5 consonant-s verbs that undergo euphonic change when attaching to "te."

例: 「持つ」…

5.3.17 動詞-自立 五段・ナ行 (verb-main group-5 consonant-n) 【infl.】

五段ナ行で,[助詞 接続助詞] の「て」に接続するときにハツ音便になるもの.

Group 5 consonant-n verbs that undergo nasalization when attaching to "te."

例: 「死ぬ」

5.3.18 動詞-自立 五段・バ行 (verb-main group-5 consonant-b) 【infl.】

五段バ行で,[助詞 接続助詞] の「て」に接続するときにハツ音便になるもの.

Group 5 consonant-b verbs that undergo nasalization when attaching to "te."

例: 「呼ぶ」…

5.3.19 動詞-自立 五段・マ行 (verb-main group-5 consonant-m) 【infl.】

五段マ行で,[助詞 接続助詞] の「て」に接続するときにハツ音便になるもの.

Group 5 consonant-m verbs that undergo nasalization when attaching to "te."

例: 「進む」…

5.3.20 動詞-非自立 五段・マ行 (verb-aux group-5 consonant-m) 【infl.】

例: 「こむ」「込む」

5.3.21 動詞-自立 五段・ラ行 (verb-main group-5 consonant-r)【infl.】

Group 5 consonant-r verbs that undergo consonant-assimilation euphonic change when attaching to "te."

例: 「切る」「なる」…

5.3.22 動詞-非自立 五段・ラ行 (verb-aux group-5 consonant-r) 【infl.】

例: 「おわる」「終る」「終わる」「かかる」「きる」「切る」「しぶる」「渋る」「まいる」「まわる」「回る」「やがる」「(せねば/しては) なら (ない)」「(て) ある」「(て) おる」「(て) まわる」「(て) 回る」「(て) やる」「ちゃる-口語/」「じゃる-口語/」「ぢゃる-口語/」

1 「なら (ない)」の基本形は「なる」

5.3.23 動詞-接尾 五段・ラ行 (verb-suffix group-5 consonant-r) 【infl.】

例: 「がる」

5.3.24 動詞-自立 五段・ラ行特殊 (verb-main group-5 consonant-r special)【infl.】

Group 5 consonant-r verbs whose masu-connection form or imperative form is " i."

例: 「いらっしゃる」「おっしゃる」「仰言る」「くださる」「下さる」「なさる」「ござる」

5.3.25 動詞-非自立 五段・ラ行特殊 (verb-aux group-5 consonant-r special)【infl.】

例: 「(お読み) なさる」「(お読み) くださる」「(お読み) 下さる」「(て) くださる」「(て) 下さる」「(て) いらっしゃる」「(て) らっしゃる-口語/」

5.3.26 動詞-自立 五段・ワ行ウ音便 (verb-main group-5 consonant-w u-onbin)【infl.】

Group 5 consonant-w verbs that undergo [[ウ音便]] euphonic change when attaching to "te."

例: 「問う」「乞う」「沿う (て)」「ゆう (て)」「食う (て)」「すう (て)」「負う (て)」

1 This tag is reserved for only group 5 w-consonant verbs whose inflectional ending is "u." We tag all other group 5 w-consonant verbs as 「動詞-自立 五段・ワ行促音便 (verb-main group-5 consonant-w consonant-onbin)」 (in our manual training data these are 「ゆう」「食う」「すう」「負う」).

5.3.27 動詞-非自立 五段・ワ行ウ音便 (verb-aux group-5 consonant-w u-onbin)【infl.】

例: 「たまう」「給う」

5.3.28 動詞-自立 五段・ワ行促音便 (verb-main group-5 consonant-w consonant-onbin)【infl.】

Group 5 consonant-w verbs that undergo consonant-assimilation euphonic change when attaching to "te."

例: 「言う」「ゆう」「「食う」「負う」「憂う」‥

1 「憂う」 does not have a corresponding 「憂って」 form, but we tag it with this category anyway (our manual training data contained only the form 「憂い (,)」).

1 This tag is used unless the inflectional ending of a group-5 consonant-w verb is "u."

5.3.29 動詞-非自立 五段・ワ行促音便 (verb-aux group-5 consonant-w consonant-onbin)【infl.】

例: 「あう」「合う」「そこなう」「損なう」「(て) しまう」「(て) もらう」「じゃう-口語/」「じまう-口語/」「ちまう-口語/」「ちゃう-口語/」

Verb Inflection Types (Classical Language)

In the IPA part of speech tagset, the inflected forms of classical language are not classified in detail. In Ipadic 2.4, we added definitions for group 4 and upper and lower group 2 inflection types, but we have not added real examples yet. The inflection hierarchy includes examples from remaining classical language and historical kana usage, even if they are of spoken form.

5.3.30 動詞-自立 四段・ハ行 (verb-main group-4 consonant-h) 【infl.】

例: 「いふ」「云ふ」「向かふ」「習ふ」「思ふ」「能ふ」など.

1 Group 4 also includes consonant-k, consonant-g, consonant-s, consonant-t, consonant-b, consonant-m, and consonant-r.

5.3.31 動詞-自立 ラ変 (verb-main group-4 consonant-r-irregular) 【infl.】

例: 「あり」「なり」「しかり」

5.3.32 動詞-自立 上二・ハ行 (verb-main upper-group-2 consonant-h) 【infl.】

1 This group also includes consonant-d verbs.

5.3.33 動詞-自立 下二・ア行 (verb-main upper-group-2 vowel) 【infl.】

1 This group also includes consonant-k, consonant-g, consonant-s, consonant-z, consonant-t, consonant-d consonant-n, consonant-h, consonant-b, consonant-m, consonant-y, consonant-r, consonant-w, and "eru" verbs.

5.3.34 動詞-自立 一段・得ル (verb-main upper-group-1 eru) 【infl.】

The inflection types of the classical word "eru." Only base form and conditional form.

5.4 Adjectives

Other than 「見出し形 (dictionary form)」, 「仮定バ接続 (conditional ba-connection form)」, and 「文語見出し形 (classical dictionary form)」, which are renamed to 「基本形 (base form)」, 「仮定形 (conditional form)」, and 「文語基本形 (classical base form)」 respectively, we use THiMCO97's inflected form names as is. Also, we subdivide the inflection types into 「形容詞・アウオ段 (adjective auo-group)」, 「形容詞・イ段 (adjective i-group)」 and 「形容詞・文語 (adjective classical)」,

Imperfective nu-connection form)

Forms that attach to ヌ.

例: 「寒から」…

Imperfective u-connection form

Forms that attach to ウ.

例: 「寒かろ」…

Conjunctive ta-connection form

Forms that attach to タ.

例: 「寒かっ」…

Conjunctive te-connection form

Forms that attach to テ, ナイ, ナル, スル, and punctuation.

例: 「寒く」…

Conjunctive gozai-connection form

Forms that attach to ゴザイマス.

例: 「寒う」「大きゅう」「のう」…

Base form

Forms that attach to punctuation , uninflected words, etc.

例: 「寒い」「大きい」「ない」…

Uninflected classical word connection form

Forms that attach to uninflected classical words.

例: 「寒き」「なき」…

& The base form is registered as ” i.”

Conditional form

Forms that attach to バ.

例: 「寒けれ」「なけれ」…

& This was called 「仮定バ接続 (conditional ba-connetion form)」 in THiMCO97.

Classical imperative form

文語活用で命令形のもの.

例: 「よかれ」「美しかれ」…

& The final form is registered as ” i.”

Classical base form

シで終わるもの.

例: 「良し」「遠し」「やむなし」…

Conditional contraction 1

The first shortened form produced by combining ”ba” and the conditional ba-connection form (spoken language).

例: 「欲しけりゃ」「(それが) なけりゃ(困る)」

Conditional contraction 2

The second shortened form produced by combining ”ba” and the conditional ba-connection form (spoken language).

例: 「(それが) なきゃ(困る)」

Garu-connection form

Forms that attach to ガル, ゲ, ソウ.

例: 「寒」「悲し」…

【infl.】 indicates a category that is an inflected form.

5.4.1 形容詞-自立 形容詞・アウオ段 (adjective-main auo-group)【infl.】

Adjectives where the final vowel of the stem form is 'a,' 'u,' or 'o.'

例: 「青い」「赤い」「厚い」「暑い」「熱い」…

1 In the IPA part of speech tagset, "nashi," the classical dictionary form of "nai" is defined as the classical inflection dictionary form, however, in this tagset, we define "nashi" as the classical base form of 「形容詞-自立 形容詞・アウオ段 (adjective-main auo-group)」. Likewise, forms like 「悪しき」 that are treated as the classical uninflected word connection form are defined the same as other adjectives as just the uninflected word connection form.

5.4.2 形容詞-自立 形容詞・イ段 (adjective-main i-group)【infl.】

形容詞の活用型のうち, 語幹の最後の母音がイで終わるもの.

Adjectives where the final vowel of the stem form is 'i.'

例: 「哀しい」「楽しい」「頼もしい」…

5.4.3 形容詞-自立 形容詞・イイ (adjective-main ii-group)【infl.】

例: 「いい」「ええ」…

5.4.4 形容詞-自立 形容詞・不変化型 (adjective-main non-inflecting)【infl.】

Adjectives that only have a base form.

例: 「かっこいい」

5.4.5 形容詞-非自立 形容詞・アウオ段 (adjective-sub auo-group)【infl.】

Auo-group adjectives that attach to a verb's conjunctive tai-connection form or conjunctive ta-connection form.

例: 「がたい」「難い」「づらい」「にくい」「やすい」「(て) よい」「(て) 良い」

5.4.6 形容詞-非自立 形容詞・イ段 (adjective-sub i-group) 【infl.】

I-group adjectives that attach to a verb's conjunctive tai-connection form or conjunctive ta-connection form.

例: 「らしい」「(て) ほしい」「(て) 欲しい」

5.4.7 形容詞-非自立 形容詞・イイ (adjective-sub ii-group)【infl.】

例: 「いい」

5.4.8 形容詞-非自立 形容詞・不変化型 (adjective-sub non-inflecting)【infl.】

Adjectives that attach to a verb's conjunctive tai-connection form or conjunctive ta-connection form and onlly have base form.

5.4.9 形容詞-接尾 形容詞・アウオ段 (adjective-suffix auo-group)【infl.】

Auo-group adjectives classified as auxiliary verbs in school grammars.

例: 「(食べ) たい」

5.4.10 形容詞-接尾 形容詞・イ段 (adjective-suffix i-group)【infl.】

I-group adjectives classified as auxiliary verbs in school grammars.

例: 「(嫌味) たらしい」

5.5 Adverbs

5.5.1 副詞-一般 (adverb-misc)

Words that can be segmented into one unit and and where adnominal modification is not possible.

例: 「あいかわらず」「多分」など.

5.5.2 副詞-助詞類接続 (adverb-particle_conjunction)

Adverbs that can be followed by 「の」「は」「に」「な」「する」「だ」 etc.

例: 「こんなに」「そんなに」「あんなに」「なにか」「なんでも」

5.6 Adnominals

5.6.1 連体詞 (adnominal)

Words that only have noun-modifying forms.

例: 「この」「その」「あの」「どの」「いわゆる」「なんらかの」「何らかの」「いろんな」「こういう」「そういう」「ああいう」「どういう」「こんな」「そんな」「あんな」「どんな」「大きな」「小さな」「おかしな」「ほんの」「たいした」「(, も) さる (ことながら)」「微々たる」「堂々たる」「単なる」「いかなる」「我が」「同じ」「亡き」…

5.7 Conjunctions

5.7.1 接続詞 (conjunction)

Conjunctions that can occur independently.

例: 「が」「けれども」「そして」「じゃあ」「それどころか」…

5.8 Particles

5.8.1 助詞-格助詞-一般 (particle-case-misc)

Case particles.

1 "nite" is included as a case particle. "no" has usages as both a case particle and another joining mouns together. The latter is classified as 「助詞 連体化 (particle-adnominalizer)」.

例: 「から」「が」「で」「と」「に」「へ」「より」「を」「の」「にて」

5.8.2 助詞-格助詞-引用 (particle-case-quote)

the "to" that appears after nouns, a person's speech, quotation marks, expressions of decisions from a meeting, reasons, judgements, conjectures, etc.

例: 「(だ) と (述べた.)」「(である) と (して執行猶予...)」

5.8.3 助詞-格助詞-連語 (particle-case-compound)

Compounds of particles and verbs that mainly behave like case particles.

例: 「という」「といった」「とかいう」「として」「とともに」「と共に」「でもって」「にあたって」「に当たって」「に当って」「にあたり」「に当たり」「に当り」「に当たる」「にあたる」「において」「に於いて」「に於て」「における」「に於ける」「にかけ」「にかけて」「にかんし」「に関し」「にかんして」「に関して」「にかんする」「に関する」「に際し」「に際して」「にしたがい」「に従い」「に従う」「にしたがって」「に従って」「にたいし」「に対し」「にたいして」「に対して」「にたいする」「に対する」「について」「につき」「につけ」「につけて」「につれ」「につれて」「にとって」「にとり」「にまつわる」「によって」「に依って」「に因って」「により」「に依り」「に因り」「による」「に依る」「に因る」「にわたって」「にわたる」「をもって」「を以って」「を通じ」「を通じて」「を通して」「をめぐって」「をめぐり」「をめぐる」「って-口語/」「ちゅう-関西弁「という」/」「(何) ていう (人)-口語/」「っていう-口語/」「といふ」「とかいふ」

5.8.4 助詞-接続助詞 (particle-conjunctive)

例: 「から」「からには」「が」「けれど」「けれども」「けど」「し」「つつ」「て」「で」「と」「ところが」「どころか」「とも」「ども」「ながら」「なり」「ので」「のに」「ば」「ものの」「や (した)」「やいなや」「(ころん) じゃ(いけない)-口語/」「(行っ) ちゃ(いけない)-口語/」「(言っ) たって (しかたがない)-口語/」「(それがなく)ったって (平気)-口語/」

5.8.5 助詞-係助詞 (particle-dependency)

例: 「こそ」「さえ」「しか」「すら」「は」「も」「ぞ」

5.8.6 助詞-副助詞 (particle-adverbial)

例: 「がてら」 「かも」「くらい」「位」「ぐらい」「しも」「(学校) じゃ(これが流行っている)-口語/」「(それ) じゃあ (よくない)-口語/」「ずつ」「(私) なぞ」「など」「(私) なり (に)」「(先生) なんか (大嫌い)-口語/」「(私) なんぞ」「(先生) なんて (大嫌い)-口語/」「のみ」「だけ」「(私) だって-口語/」「だに」「(彼)ったら-口語/」「(お茶) でも (いかが)」「等 (とう)」「(今後) とも」「ばかり」「ばっか-口語/」「ばっかり-口語/」「ほど」「程」「まで」「迄」「(誰) も (が)([助詞-格助詞] および [助詞-係助詞] の前に位置する「も」)

5.8.7 助詞-並立助詞 (particle-coordinate)

例: 「と」「たり」「だの」「だり」「とか」「なり」「や」「やら」

5.8.8 助詞-終助詞 (particle-final)

例: 「かい」「かしら」「さ」「ぜ」「(だ)っけ-口語/」「(とまってる) で-方言/」「な」「ナ」「なあ-口語/」「ぞ」「ね」「ネ」「ねぇ-口語/」「ねえ-口語/」「ねん-方言/」「の」「のう-口語/」「や」「よ」「ヨ」「よぉ-口語/」「わ」「わい-口語/」

1 The sentence-final particle 「や」 as in 「(まあいい) や」, 「(すごい) や」, etc. We treat the sentence-final particle from the Kansai dialect as a non-inflectional auxiliary verb.

5.8.9 助詞-副助詞/並立助詞/終助詞 (particle-adverbial/conjunctive/final)

The particle "ka," when unknown whether it is adverbial, conjunctive, or sentence final. For example,

(a)「A か B か」. Ex:「(国内で運用する) か,(海外で運用する) か (.)」

(b) Inside an adverb phrase. Ex:「(幸いという) か (, 死者はいなかった.)」「(祈りが届いたせい) か (, 試験に合格した.)」

(c)「かのように」. Ex:「(何もなかった) か (のように振る舞った.)」

例: 「か」

1 In the latest IPA part of speech hierarchy, this category is further classified into 「副助詞 (adverbial)」, 「並立助詞 (conjunctive)」, and「終助詞 (final)」, but we did not divide this category any further.

5.8.10 助詞-連体化 (particle-adnominalizer)

The "no" that attaches to nouns and modifies non-inflectional words.

1 In THiMCO97 this usage of "no" was also classified as a case particle.

5.8.11 助詞-副詞化 (particle-adnominalizer)

the "ni" and "to" that appear following nouns and adverbs that are giongo, giseigo, or gitaigo.

例: 「に」「と」

1 However, when a partible represents a change of state and modifies "suru" or "naru," it is classified as a case particle.

5.8.12 助詞-特殊 (particle-special)

A particle that does not fit into one of the above classifications. This includes particles that are used in Tanka, Haiku, and other poetry.

例: 「かな」「けむ」「(しただろう) に」「(あんた) にゃ(わからん)」「(俺) ん (家)」

5.8.13 助詞-間投助詞 (particle-interjective)

This part of speech was not present in Ipadic. It is for particles with grammatical roles and was defined in version 2.4.

例: 「(松島) や」

5.9 助動詞

5.10 Auxiliary Verbs

5.10.1 助動詞 五段・ラ行アル (aux group-5 aru)【infl.】

The auxiliary verb that acts as the inflection of "dearu." The "aru" of 「である」, 「ではある」, etc.

例: 「ある」

5.10.2 助動詞 五段・ラ行ゴザル (aux group-5 gozaru)【infl.】

The auxiliary verb that acts as the inflection of "gozaru."

例: 「ござる」

5.10.3 助動詞 形容詞・イ段 (aux adjective_i-group)【infl.】

The auxiliary verb that act as adjective inflections.

例: 「らしい」

5.10.4 助動詞 特殊・ナイ (aux special-nai)【infl.】

The inflection type of the negation auxiliary verb "nai."

例: 「ない」

5.10.5 助動詞 特殊・タ (aux special-ta)【infl.】

The inflection type of the auxiliary verb that indicates perfective.

例: 「た」「だ」

1 Because group 5 consonant-g, consonant-n, consonant-b, and consonant-m verbs have the surface form "da" like in 「(学ん) だ」 and 「(泳い) だ」, we define "ta" and "da" as separate morphemes in our tagset.

5.10.6 助動詞 特殊・ダ (aux special-da)【infl.】

The inflection type of the assertive auxiliary verb "da."

例: 「だ」

5.10.7 助動詞 特殊・デス (aux special-desu)【infl.】

The inflection type of the assertive auxiliary verb "desu."

例: 「です」

5.10.8 助動詞 特殊・ドス (aux special-dosu)【infl.】

The inflection type of the assertive auxiliary verb "dosu."

例: 「どす」

5.10.9 助動詞 特殊・ジャ (aux special-ja)【infl.】

The inflection type of the assertive auxiliary verb "ja."

例: 「じゃ」

1 A softening of assertive "da."

5.10.10 助動詞 特殊・マス (aux special-masu)【infl.】

The inflection form of the auxiliary verb "masu" that indicates humility or politeness.

例: 「ます」

5.10.11 助動詞 特殊・ヌ (aux special-nu)【infl.】

The inflection of the negation auxiliary verb "nu."

例: 「ぬ」

5.10.12 助動詞 特殊・ヤ (aux special-ya)【infl.】

The assertive auxilary verb "ya" in the Kansai dialect.

例: 「(そう) や」

5.10.13 助動詞 不変化型 (aux non-inflectional)【infl.】

Auxiliary verbs that are non-inflectional in Modern Japanese. Includes spoken language or dialects where the inflection is unknown.

例: 「う」「まい」「(いざ行か) ん (む)」「(去り) ぬ」「(わから) ん-口語/」「(賜ラ) ン」「〜(美しい/学生) じゃん-口語/」「(いい)っす-口語/」「(負けてなら) じ」など.

5.10.14 助動詞 文語・?? (aux classical)【infl.】

Auxiliary verbs in Classical Japanese. Currently, the following inflection types are defined: classical-beshi, classical-gostoshi, classical-nari, classical-maji, classical-shimu, classical-ki, classical-keri, classical-ru, and classical-ri.

例: 「べし」「ごとし」「如し」「たり」「なり」「まじ」「き」「けり」「り」「る」

1 In the IPA part of speech tagset, an inflection type was prepared for "ji," but since it does not actually inflect, we group it in this category.

5.11 Exclamations

5.11.1 感動詞 (exclamation)

Greetings and other exclamations.

例: 「おはよう」「おはようございます」「こんにちは」「こんばんは」「ありがとう」「どうもありがとう」「ありがとうございます」「いただきます」「ごちそうさま」「さよなら」「さようなら」「はい」「いいえ」「ごめん」「ごめんなさい」…

5.12 Symbols

5.12.1 記号-一般 (symbol-misc)

A general symbol not in one of the categories below.

例: 「○」「◎」「@」「$」「〒」「→」「+」など.

5.12.2 記号-アルファベット (symbol-alphabetic

Upper- and lowercase English alphabetic characters.

例: 「A」「a」

5.12.3 記号-句点 (symbol-comma)

Commas.

例: 「,」「、」

5.12.4 記号-読点 (symbol-period)

Periods and full stops.

例: 「.」「.」「。」

5.12.5 記号-空白 (symbol-space)

full-width whitespace (cannot be displayed on-screen).

5.12.6 記号-括弧開 (symbol-open_bracket)

例: 「(」「{」「‘」「“」「『」「【」…

5.12.7 記号-括弧閉 (symbol-close_bracket

例: 「)」「}」「’」「”」「』」「】」…

5.13 Filler

5.13.1 フィラー (filler)

Aizuchi that occurs during a conversation or sounds inserted as filler.

例: 「あの」「うんと」「えと」

5.14 その他

5.15 Other Parts of Speech

5.15.1 その他-間投 (other-interjection)

Words that are hard to classify as noun-suffixes or sentence-final particles.

例: 「(だ)ァ」

参考文献

[1] 益岡隆志, 田窪行則: 『基礎日本語文法 –改訂版–』くろしお出版, 1992.

[2] 松本裕治, 黒橋禎夫, 宇津呂武仁, 妙木裕, 長尾真: 「日本語形態素解析システム JUMAN 使用説明書 version 2.0」, NAIST Technical Report, NAIST-IS-TR94025, 1994.

[3] 山下達雄: 「規則と確率モデルの統合による形態素解析」, 奈良先端科学技術大学院大学修士論文, NAIST-IS-MT9551119, March 1997.

[4] 山下達雄, 松本裕治: 「コスト最小法と確率モデルの統合による形態素解析」, 情報処理学会研究報告 96-NL-119, May 1997.

[5] 北内 啓, 山下 達雄, 松本 裕治: 「日本語形態素解析システムへの可変長連接規則の実装」, 言語処理学会第三回年次大会論文集, pp.437-440, 1997.

[6] 「研究開発用知的資源タグ付きテキストコーパス報告書」平成 9 年度, テキストサブワーキンググループ, 技術研究組合 新情報処理開発機構, 1998.

[7] 松田 寛: 「品詞タグ付きコーパス作成支援環境の構築」, 奈良先端科学技術大学院大学修士論文, NAIST-IS-MT9851103, March 1999.

[8] 北内 啓, 宇津呂 武仁, 松本 裕治: 「誤り駆動型の素性選択による日本語形態素解析の確率モデル学習」, 情報処理学会論文誌 Vol. 40, No. 5, p.p.2325-2337, May 1999.

[9] 松田 寛, 桐山 和久, 山田 悟史, 吉野 圭一, 松本裕治: 「部分形態素解析を用いたコーパスの品詞体系変換」, 情報処理学会研究報告 99-NL-134, p.p.23-30, Nov. 1999.

[10] Masayuki Asahara: Extended Statistical Model for Morphological Analysis, 奈良先端科学技術大学院大学修士論文, NAIST-IS-MT9851001, March 2000.

[11] 松田 寛, 松本 裕治: 「品詞タグ付きコーパス作成支援 GUI ツール VisualMorphs」, 情報処理学会研究報告 2000-NL-137, p.98, June, 2000.

[12] 浅原 正幸, 松本 裕治: 「統計的日本語形態素解析に対する拡張 HMM モデル」, 情報処理学会研究報告 2000-NL-137, p.p.39-46, June, 2000.

[13] Masayuki Asahara, Yuji Matsumoto: Extended Models and Tools for High-performance Part-of-Speech Tagger, Proceedings of COLING 2000, July, 2000.

[14] 浅原 正幸, 松本 裕治: 「誤り駆動による統計的品詞タグづけモデルの拡張」, 情報処理学会研究報告 2000-NL-139, p.p.25-32, Sep. 2000.

[15] 松本 裕治: 「形態素解析システム『茶筌』」, 情報処理 Vol.41 No.11, p.p.1208-1214, Nov. 2000.

[16] 伝 康晴, 浅原 正幸: 「リレーショナル・データベースによる統合的言語資源管理環境」, ワークショップ「話し言葉の科学と工学」, Feb. 2001.

[17] 松本 裕治, 伝 康晴: 「話し言葉の形態素解析」, 情報処理学会研究報告 2001-NL-143, p.p.49-54, May, 2001.

[18] Masayuki Asahara and Ryuichi Yoneda and Yuji Matsumoto: 「Use of a Relational Database in the Development and Maintenance of Linguistic Resources for Statistical Japanese Morphological Analysis」, IRCS Workshop on Linguistic Databases, Dec. 2001.

[19] 浅原 正幸, 松本 裕治: 「形態素解析のための拡張統計モデル」, 情報処理学会論文誌, Vol.43 No.03, Mar. 2002.

[20] 浅原 正幸, 米田 隆一, 山下 亜希子, 伝 康晴, 松本 裕治: 「リレーショナルデータベースによる品詞タグつきコーパスの管理手法」, SIG-SLUD-34 Mar. 2002.

[21] Masayuki Asahara and Ryuichi Yoneda and Akiko Yamashita and Yasuharu Den and Yuji Matsumoto: 「Use of XML and Relational Databases for Consistent Development and Maintenance of Lexicons and Annotated Corpora」, LREC 2002, May. 2002.

[22] 浅原 正幸, 米田 隆一, 山下 亜希子, 伝 康晴, 松本 裕治: 「語長変換を考慮したコーパス管理システム」, 情報処理学会論文誌, Vol.43 No.07, Jul. 2002.

[23] 浅原 正幸, 松本 裕治: 「形態素解析とチャンキングの組み合わせによる日本語テキスト中の未知語出現箇所同定」, 情報処理学会研究報告 2003-NL-154, Mar. 2003.

付録

A Revision History

A.1 Changes from ipadic-2.6.3 to ipadic-2.7.0

A detailed list of chages can be found at: `http://chasen.aist-nara.ac.jp/~masayu-a/ipadic/arch/ipadic-2.7.0-dif`

- Received a report of unregistered words and errors from Mr. Mizushima of NU-HIStory.
- Moved one-byte characters to a separate dictionary file Onebyte.dic.
- Changes to inflected forms and inflectional types: added 「形容詞・イイ」 and 「特殊・ドス」

A.2 Changes from ipadic-2.6.2 to ipadic-2.6.3g

A detailed list of chages can be found at: `http://chasen.aist-nara.ac.jp/~masayu-a/ipadic/arch/ipadic-2.6.3-dif`

- Received a report of errors from Mr. Mine of the National Institute of Japanese Language.
- Received a report of errors from Mr. So of the National Institute of Japanese Language.
- Received a report of unregistered words and errors from Mr. Mizushima of NU-HIStory.

A.3 Changes from ipadic-2.6.1 to ipadic-2.6.2

Improved the part of speech connection file. No changes to the dictionary entries.

A.4 Changes from ipadic-2.6.0 to ipadic-2.6.1

No changes to the dictionary entries.

- Updated chasenrc
- Support for ChaSen 2.3.2.

A.5 Changes from ipadic-2.5.1 to ipadic-2.6.0

ipadic-2.5.1 and ipadic-2.6.0 use the exact same part of speech tagset.

- Our thanks to Professor Tsurumaru of Nagasaki University who helped us register 6,000 new words.
  A detailed list of chages can be found at: `http://chasen.aist-nara.ac.jp/~masayu-a/ipadic/arch/ipadic-2.6.0-`
- A list of new vocabulary can be found at: `http://chasen.aist-nara.ac.jp/~masayu-a/ipadic/arch/ipadic-2.6.0`
- Support for ChaSen 2.3.1.
  - Deleted .pat and .ary generated portions.
  - Support for double array dictionaries (dadic)
  - Support for the character encoding option -i

A.6 Changes from ipadic-2.5.0 to ipadic-2.5.1

ipadic-2.5.0 and ipadic-2.5.1 use the exact same part of speech tagset.

- Moved compound case particles and other similar words to `Postp-col.dic`.
- Updated parameters surrounding group 1 verbs and auxiliary verbs
- A list of new vocabulary can be found at: `http://chasen.aist-nara.ac.jp/~masayu-a/ipadic/arch/ipadic-2.5.1`

A.7 Changes from ipadic-2.4.X to ipadic-2.5.0

ipadic-2.4.X and ipadic-2.5.0 use the exact same part of speech tagset.

B Dictionary Copyright

incidental and consequential damages, arising out of or otherwise in connection with the use or inability to use the program or any product, material or result produced or otherwise obtained by using the program, regardless of whether they have been advised of, or otherwise had knowledge of, the possibility of such damages at any time during the project or thereafter. Each user will be deemed to have agreed to the foregoing by his or her commencement of use of the program. The term "use" as used herein includes, but is not limited to, the use, modification, copying and distribution of the program and the production of secondary products from the program.

In the case where the program, whether in its original form or modified, was distributed or delivered to or received by a user from any person, organization or entity other than ICOT, unless it makes or grants independently of ICOT any specific warranty to the user in writing, such person, organization or entity, will also be exempted from and not be held liable to the user for any such damages as noted above as far as the program is concerned.