日本語形態素解析システム
JUMAN version 7.0

黒橋・河原研究室

平成 23 年 12 月

Japanese Morphological Analysis System JUMAN 7.0 Users Manual

Version 0.6	17 February 1992
Version 0.8	14 April 1992
Version 1.0	25 February 1993
Version 2.0	11 July 1994
Version 3.0b	11 June 1996
Version 3.0	16 October 1996
Version 3.1	15 November 1996
Version 3.11	5 December 1996
Version 3.2	22 May 1997
Version 3.3	5 September 1997
Version 3.4	11 November 1997
Version 3.5	23 March 1998
Version 3.6	2 December 1998
Version 3.61	26 May 1999
Version 4.0	19 July 2003
Version 5.0	23 March 2005
Version 5.1	7 September 2005
Version 6.0b	30 September 2007
Version 6.0	26 September 2009
Version 7.0	31 December 2011

目 次

1　はじめに

計算機による日本語の解析において，欧米の言語の解析と比べてまず問題になるのが次の 2 点です．一つは形態素解析の問題です．ワードプロセッサの普及などによって日本語の入力には大きな問題がなくなりましたが，計算機による日本語解析では，まず入力文内の個々の形態素を認識する必要があります．これには実用に耐えられるだけの大きな辞書も必要であり，これを如何に整備するかという問題も同時に存在します．もう一つの問題として，日本語には広く認められ同意を得られた文法，ないし，文法用語がないという現実です．学校文法の単語分類および文法用語は一般には広く知られていますが，研究者の間ではあまり評判がよくありませんし，計算機向きではありません．

日本語の解析に真っ先に必要な形態素解析システムは，多くの研究グループによって既に開発され技術的な問題が洗い出されているにも係わらず，共通のツールとして世の中に流布しているものはありません．計算機可読な日本語辞書についても同様です．

本システムは，計算機による日本語の解析の研究を目指す多くの研究者に共通に使える形態素解析ツールを提供するために開発されました．その際，上の二つ目の問題を考慮し，使用者によって文法の定義，単語間の接続関係の定義などを容易に変更できるように配慮しました．

本システムの辞書は，JUMAN2.0 までのバージョンでは，長尾研究室で行われた Mu プロジェクトで開発された辞書，Wnn かな漢字変換システムの辞書，および，ICOT から公開された日本語辞書を利用しました．また，JUMAN3.X では，(株) 日本電子化辞書研究所 (EDR) から許諾を得て EDR 日本語単語辞書の一部を利用しました．また，JUMAN4.0 以降のバージョンでは独自に構築した辞書を利用しています．

本システムの開発は京都大学長尾研究室・黒橋研究室を中心に，他の多くの機関の方々の協力を得て行われました．長尾研究室の松本裕治先生 (現 奈良先端大) には常に開発の指導的役割をはたして頂きました．妙木裕さん (現 キヤノン) には JUMAN の初期バージョンを開発していただきました．米国 MCC の自然言語グループの方々には JUMAN1.0 でハッシングによる辞書引きの高速化を実現していただきました．SICStus Prolog とのソケットによる接合部を東京工大の徳永健伸先生に作成していただきました．また，富士通の中村俊久さんには JUMAN2.0 で ndbm による辞書のハッシュデータベース化を実現していただきました．長尾研究室の山地治さんには連語処理をはじめとした JUMAN3.0 へのバージョンアップをしていただきました．NTT 基礎研の磯崎秀樹さんには JUMAN3.4 へのバージョンアップに伴う SICStus Prolog との接合部の調整をしていただきました．また，JUMAN3.1 以降では奈良先端大松本研究室で開発された茶筅のパトリシア木による辞書検索システムを利用させていただいています．また，JUMAN3.5 でのサーバ・クライアント方式への拡張は同じく茶筅を参考にさせていただき，その拡張作業には学術情報センターの影浦峡先生，吉岡真治先生に多大なご協力頂きました．言語メディア研 (旧長尾研) の工藤拓さんには Windows 版の作成と Perl モジュール化をしていただきました．

また，JUMAN6.0 への拡張においては，黒橋研の原島純さんに意味カテゴリの検討・付与，橋本力さん (現 NICT) にドメインの検討・付与，笹野遼平さんに連濁，反復形オノマトペ，小書き文字による非標準表記の自動認識，村脇有吾さんに未知語の自動獲得を中心となって行っていただきました．

JUMAN7.0 への拡張においては，柴田知秀助教に Wikipedia からの語彙の自動獲得を行っていただきました．

また，一人一人お名前を挙げることはできませんが，JUMAN システムに対して多くの方々からコメントと質問をいただきました．御協力頂いた各位にお礼申し上げます．

平成 23 年 12 月

本システムに関するお問い合わせは以下にお願いします.
〒 606-8501 京都府京都市左京区吉田本町
京都大学大学院情報学研究科 知能情報学専攻
黒橋 禎夫
Tel: (075)753-5344
Fax: (075)753-5962
Email: nl-resource@nlp.ist.i.kyoto-u.ac.jp

2　日本語形態素文法

本システムで仮定している日本語の形態素文法およびここで用いる用語について説明する．

2.1　形態品詞および品詞細分類

本システムでは形態素の集合の二階層の分類が仮定されている．これらの分類を**形態品詞**および**(形態) 品詞細分類**と呼ぶ．

例えば，名詞を普通名詞，サ変名詞，固有名詞などに細分類することができる．また，ある形態品詞を細分類したくない場合は細分類を指定しなくてもよい．例えば，品詞細分類を持たない形態品詞として動詞を定義してもよい．

この二階層に分類されたそれぞれの集合を**品詞**と呼ぶ．上の例では，名詞の細分類としての普通名詞，サ変名詞，固有名詞，および，動詞はそれぞれ品詞である．

2.2　活用型と活用形

学校文法における '動詞' '形容詞' '形容動詞' などのように，後接する形態品詞に応じてその形態が変化する形態素が存在する．このような形態変化を**活用** (conjugation) という．このような形態素の表層のうち変化しない部分を **(活用) 語幹**，変化する部分を **(活用) 語尾**と呼ぶ．大部分の活用は規則的である．その規則性に従って活用のタイプを分類し，これを**活用型**と呼ぶ．そして，連接に応じて実際にとり得る表層的な個々の形態を **活用形**と呼ぶ．

2.3　形態素構造

形態素 m の形態品詞，品詞細分類，活用型，活用形，表層形がそれぞれ $H1$, $H2$, $K1$, $K2$, M であるとき，これを 5 項組 ($H1$ $H2$ $K1$ $K2$ M) によって表わし,「形態素 m の **形態素構造**」と呼ぶ．

形態素構造の項 $H1, H2, K1, K2$ には，具体的な名称として，それぞれ形態品詞名・品詞細分類名・活用型名・活用形名を与える．また，項 M には，漢字または仮名で表記される表層形を与える．各形態素構造の記述には特別なシンボル '$*$' を使用することができる．'$*$' は，その項を考慮しない (未指定，'don't care') ことを表わす．形態素構造を表わすリストのある要素以降がすべて未指定でよい場合は，形態品詞の部分以外はそれらを省略してもよい．たとえば，形態素構造

$$\alpha_1 = (*)$$

は任意の形態素を表わす集合とみなすことができる．同様に，形態素構造

$$\alpha_2 = (\text{名詞})$$

は '名詞' という形態品詞に分類されたすべての形態素の集合を表わす．また，形態素構造

$$\alpha_3 = (* \ * \ * \ \text{未然形})$$

は '未然形' という活用形を取るすべての形態素の集合を表わす[1]．また，

$$\alpha_4 = (\text{助詞 格助詞} * \ * \ \text{が})$$

[1] この例では，$\alpha_1 \supseteq \alpha_2, \alpha_1 \supseteq \alpha_3$ という関係が成り立つ．

のように具体的な形態素を指定することも許される[2].

この記法は次に述べる連接規則を記述する際に使われる.

2.4 連接規則と連接関係

形態素構造の集合の組 A_1 と A_2 に含まれるそれぞれの要素が文中に連続して現れ得るとき，それらは連接可能であるといい，それを記述した規則を**連接規則**と呼ぶ．また，連接規則によって定義される 2 項関係を**連接関係**と呼ぶ.

連接規則は二種類の形態素の連接可能性が記述可能である．また，連接の強弱に関するコストをそれぞれの連接規則に対して記述することができる．連接規則は，次節で述べる連接規則辞書によって定義される.

3 辞書の定義とデータ構造

3.1 辞書の概要

本システムで用いる辞書の種類について述べる．本辞書システムの特徴は，利用者によって形態品詞名や連接関係などの文法情報やを自由に定義できることである.

日本語の形態素文法は，本システムに標準的に仮定されているファイルによって定義されなければならないが，この定義はユーザに任されている．ユーザが定義すべき辞書を総称して**ユーザ辞書**と呼ぶ．ユーザ辞書は，**文法辞書**と**形態素辞書**に大別される．文法辞書は 2 節で述べた日本語形態素文法を定義するための辞書で，**形態品詞分類辞書**，**活用関係辞書**，**活用辞書**，**連接規則辞書**から構成される (3.2 節)．形態素辞書は個々の形態素についての情報を実際に記述する辞書である (3.3 節).

形態素解析において使用する辞書を**システム辞書**と呼ぶ．システム辞書は，連接規則辞書と活用関係辞書から生成された**連接対応表**と**連接行列**，文法辞書を参照して形態素辞書に記述された情報を記号化した**データベース辞書**と**インデックスファイル**からなる.

ユーザ辞書およびシステム辞書を以下にまとめる．本システムには標準的なユーザ辞書が含まれているが，ユーザはこれを自由に書き直してよい．システム辞書は，本システムによってユーザ辞書から自動的に生成されるので，ユーザはこれらを意識する必要はない．形態素解析プログラムは変換して得られたシステム辞書のみを用いる.

ユーザ辞書

辞書の種類	辞書名称 (ファイル名)	説明
文法辞書	形態品詞分類辞書 (`JUMAN.grammar`)	品詞分類を定義する
	活用関係辞書 (`JUMAN.kankei`)	活用する各品詞に対して活用型の一覧を定義する
	活用辞書 (`JUMAN.katuyou`)	個々の活用型の具体的な活用形の一覧を定義する
	連接規則辞書 (`JUMAN.connect.c`)	連接関係を定義する連接規則の集合
形態素辞書	`ContentW.dic` など	個々の形態素の定義 (3.3 を参照)

システム辞書

[2] この例のように，活用しない形態素に対しても 5 項組によって形態素を表現する必要があり，表層形を指定する場合には活用型名，活用形名の部分を * としなければならない.

辞書の種類	辞書名称 (ファイル名)	説明
文法辞書	連接対応表 (jumandic.tab) 連接行列 (jumandic.mat)	形態品詞と接続番号との対応表 接続可能性表
形態素辞書	データベース辞書 (jumandic.dat) インデックスファイル (jumandic.pat)	各形態素の情報が格納される パトリシア木によるインデックス

本システムを使用する際に個人的にカスタマイズできる情報があり，その情報は，使用者のホームディレクトリにある.jumanrc というファイルで設定できる．このファイルが存在しない場合は，システムの標準的な設定ファイルが使用される．設定ファイルに書かれる情報の詳細は後で詳しく述べる．

また，本システムでは標準的なユーザ辞書を用意している．これについては，3.5 節で紹介する．

文法辞書および形態素辞書ファイルでは，各行中のセミコロン ';' で始まる文字列はその行末までがコメントとみなされる．

3.2 文法辞書の定義法

文法辞書の構成と内容は次のようになっている．各辞書の記述は S 式によって行う．付録 A に記述のフォーマットを BNF 式と例によって示す．

1. **形態品詞分類辞書：** (cf. JUMAN.grammar)

   本システムで用いる形態品詞およびその品詞細分類の名称を定義する．一つの形態品詞およびそれの品詞細分類を一つの S 式 (リスト構造) によって記述する．リストの第 1 要素に形態品詞を記述し，品詞細分類があれば第 2 要素に品詞細分類名をリストによって記述する．

   個々の形態品詞と品詞細分類はリストの形で表現されるが，それに属する形態素が活用する場合にはその第 2 要素として記号 % を付加する．形態品詞が記号 % をもつ場合には，その細分類に属するすべての形態素が活用するとみなされる．

   バージョン 3.4 以降では未定義語も一つの品詞として扱い，その細分類として「カタカナ」，「アルファベット」,「その他」をたてている．これらは形態品詞分類辞書で必ず定義する必要がある (4.2 節参照)．

   形態品詞分類辞書の記法を付録 A.1 に示す．

2. **活用関係辞書：** (cf. JUMAN.kankei)

   活用関係辞書は，形態品詞分類辞書によって活用すると定義された形態品詞または品詞細分類それぞれに対してそれが取り得る活用型の一覧を定義する．形態品詞分類辞書で活用すると定義されたものに対しては，必ず取り得る活用型の一覧が書かれていなければならない．活用型の名前付けはユーザが自由に行なってよい．

   個々の活用関係は，二つの要素からなるリストとして定義される．第一要素は形態品詞のリストまたは形態品詞と品詞細分類からなるリストであり，活用する品詞を示す．第二要素はその品詞が取り得る活用型の一覧を示すリストであり，任意個数のアトムを含む．

   活用関係辞書の記法を付録 A.2 に示す．

3. **活用辞書：** (cf. JUMAN.katuyou)

活用辞書は，活用関係辞書によって定義された個々の活用型に対して，それが取り得る活用形の名前およびその活用語尾を定義する．

この辞書の各項目も二つの要素からなるリストであり，第 1 要素に活用型の名前を記述し，第 2 要素として活用形の名前と活用語尾のリストより構成されるリストを記述する．この時，活用形の名前として「**基本形**」を必ず記述しなければならない．後で述べる形態素辞書には，活用する形態素はその基本形が登録されていると仮定されることになっているので，個々の形態素については，形態素辞書に登録された見出し表記からこの辞書で定義される基本形の語尾を取り除いたものがその形態素の不変化部分であるとみなされる．活用語尾の存在しない活用形 (例えば，語幹) に対しては，活用語尾として * と記述すればよい．

また，不変化部分が存在しない活用語については，見出し語と基本形語尾が同じになる (例えば，付録 A.3.2 のサ変動詞型の活用)．このような場合には，システムはすべての活用語尾を個別の形態素として形態素辞書に登録する．

活用辞書の記法を付録 A.3 に示す．

4. **連接規則辞書：** (cf. `JUMAN.connect`)

連接規則辞書は連接規則の集合である．一つの連接規則は形態素構造 (2.3 節を参照) の集合の対であり，2 要素のリストによって表現される．形態素構造の集合もリストによって表現される．連接規則には第 3 要素として連接のコストを記述することができる．コストの指定がない場合のデフォルト値は “10” である．コストの値は，0 から 255 までの整数値でなければならない．ただし, 連接規則辞書から得られる連接行列では, 二つの形態素が連接不可能であることを数値 0 によって表わしているので, 連接のコストを 0 と指定することはそれらが連接不可能であることを意味する. したがって, 連接条件の例外的な記述を連接コストを 0 にすることによって行うことができる.

一つの連接規則の第 1 要素に含まれる形態素は第 2 要素に含まれる形態素に連接可能であることを示す．一つの形態素が複数の連接規則内の形態素構造に含まれることがあってもよい．連接関係はこれらの規則の和集合であるとみなされる．したがって，ある特定の形態素構造の対が連接可能であることが複数の連接規則によって定義されることがある．このような場合，これらの連接のコストは，連接規則辞書内で後方に現れた規則によって指定されるコストを取るとみなされる．

形態素構造で用いられる品詞名や活用形名は文法辞書で定義されたものでなければならない．例外として，(文頭) および (文末) という形態素構造がある．前者は，文頭の直前に現れると考えられるダミーの形態品詞であり，少なくとも一つの連接規則の第 1 要素の中に現れなければならない．同様に，後者は文末の直後に繋がると考えられるダミーの形態品詞であり，少なくとも一つの連接規則の第 2 要素の中に現れる必要がある．

連接規則辞書の記法を付録 A.4 に示す．

連接規則においては，ある品詞のグループが同じ連接関係をもつことがよく起こる．連接規則の記述を省力化するために，gcc で用意されているマクロ機能を利用することができる．`JUMAN.connect.c` というファイルに次のようなマクロを定義することにより，以降の連接規則内で `NormalNominal` という記号を`#define` の第二引数と等価に扱うことができる．

```
#define NormalNominal (名詞 普通名詞) \
                      (名詞 サ変名詞) \
```

(名詞 固有名詞) \
(名詞 地名) \
(名詞 人名) \
(名詞 数詞) \
(指示詞 名詞形態指示詞) \
(動詞 * * 基本連用形)

3.3 形態素辞書の記述法

形態素辞書もリスト構造を用いて記述する．形態素辞書は，最後に 'dic' という拡張子をもつファイルに格納する．形態素辞書は複数のファイルに分割されていてもよい．

各形態素は，複数の見出し語を持ってもよい．また，それぞれの見出し語に対してコスト計算のための相対的な重みを指定することができる．形態素のもつコストについては，4 章で説明する．複数の見出し語が記述されている場合，見出し語以外の形態素情報は共有される．

次に形態素辞書の形態素定義の記述方法を BNF で示す．

〈形態素定義〉 ::= (〈#形態品詞名〉〈形態素情報の並び〉) |
(〈#形態品詞名〉(〈#品詞細分類名〉〈形態素情報の並び〉))
〈形態素情報の並び〉 ::= 〈形態素情報〉 | 〈形態素情報〉〈形態素情報の並び〉
〈形態素情報〉 ::= (〈見出し語情報〉〈読み情報〉〈活用型情報〉〈意味情報〉)
〈見出し語情報〉 ::= (見出し語 〈見出し語内容の並び〉)
〈見出し語内容の並び〉 ::= 〈見出し語内容〉 | 〈見出し語内容の並び〉
〈見出し語内容〉 ::= 〈#見出し語表記〉 | (〈#見出し語表記〉) |
(〈#見出し語表記〉〈#数値〉)
〈読み情報〉 ::= (読み 〈#読み表記〉)
〈活用型情報〉 ::= (活用型 〈#活用型名〉) | NIL
〈意味情報〉 ::= (意味情報 〈#意味記述〉) | NIL

- 〈形態素情報〉は，共通の形態品詞名または品詞細分類名ごとにまとめて〈形態素情報の並び〉として書くこともできるし，個別に記述してもよい．〈#形態品詞名〉,〈#品詞細分類名〉は,「形態品詞分類辞書」で定義されていなければならない．また，品詞細分類されている形態品詞に対しては品詞細分類名を指定しなければならない．

- 〈活用型情報〉は,〈#形態品詞名〉または〈#品詞細分類名〉が活用すると定義されている時は省略できない．

- 〈#見出し語内容〉は形態素の表層の形と相対的な重みを表わす数値の対として表現する．相対的な重みが 1 の場合は，重みの記述を省略してもよい．重みとして指定できるのは正整数または小数点 1 桁以内の正小数である．活用する形態素の表層の形は，その基本形を書かねばならない．

- 〈#読み表記〉には形態素の読みを記述する．活用する形態素については，その基本形の読みを書く．

- 〈#意味記述〉には意味情報を記述する．意味情報には任意の S 式を記述することができる．また，アトムとして，二重引用符 (") で囲まれたテキス トを自由に用いることができる．二

重引用符に囲まれた範囲では任意の文字が使用可能である．二重引用符で囲まれたアトムの中の二重引用符は '\"' によって記述可能である．記述されたデータは二重引用符を含んでそのままテキストデータとして扱われる．長さについての制限は設けていない．

付録 B に形態素辞書の記述例を示す．

3.4 複数辞書の使用

JUMAN では，必要に応じて複数の辞書を使用することが出来るようになっている．ここではその使用法について述べる．

辞書は 3.1 節で述べたように，文法辞書及び形態素辞書に分けられる．複数化出来るのは，データベース辞書 (jumandic.dat)・インデックスファイル (jumandic.pat) からなる形態素辞書である．

利用者が用いる形態素辞書の指定は，他の設定と同様にリソースファイルによって行われる．具体的な記述法については，3.6 節を参照のこと．また，複数の形態素辞書を使用する場合は，それらは同じ文法辞書にしたがっていなければならない．

3.5 システム標準文法

本システムのユーザ辞書として標準の文法が用意されている．これをシステム標準文法と呼ぶ．システム標準文法を定義する文法辞書のうち，形態品詞分類辞書，活用辞書，連接規則辞書は，益岡・田窪文法 [1] を参照し，それを拡張して作成した．

1. **形態品詞分類辞書：** (cf. `JUMAN.grammar`)

   益岡・田窪文法に「特殊」(句読点・記号・括弧など) を加え，接辞を「接頭辞」「接尾辞」に分けて，14 種類の形態品詞を定義した．

2. **活用辞書：** (cf. `JUMAN.katuyou`)

   益岡・田窪文法に対して，文語的表現・口語的表現・敬語表現に対応できるように拡張を行い，21 の一般的な活用型と 7 の特殊な活用型を定義した．

3. **活用関係辞書：** (cf. `JUMAN.kankei`)

   活用する形態構造の一覧とそれらが取り得る活用型の一覧を定義している．

4. **連接規則辞書：** (cf. `JUMAN.connect.c`)

   新たに作成した．

3.6 リソースファイル (jumanrc)

形態素解析プログラムに必要な様々な設定はリソースファイルによって指定する．ユーザのホームディレクトリに '`.jumanrc`' というファイルがあればこれがリソースファイルとして用いられる．ユーザのホームディレクトリにこの名のファイルが存在しない場合は，`$PREFIX/etc/jumanrc` が参照される．

以下に必要な設定項目を示す．これらはすべて指定されていなければならない．(リソースファイルの記述例を付録 C に示す．)

1. **文法辞書の位置**

   文法辞書 (JUAMN.grammar，JUMAN.kankei，JUMAN.katuyou，JUMAN.connect，jumandic.tab，jumandic.mat) が存在するディレクトリを指定する．指定は次のように行う．

   (文法ファイル /usr/local/share/juman/dic)

2. **形態素辞書の位置**

   形態素辞書 (jumandic.dat，jumandic.pat が存在するディレクトリを指定する．複数の形態素辞書ディレクトリを指定することも可能である．指定は次のように行う．

   (辞書ファイル /home/mydir/juman/dic
   /usr/local/share/juman/dic)

   この例では，二つ形態素辞書ディレクトリを指定している．辞書引きに際しては，これら両方のディレクトリ中の形態素辞書が用いられる[3]．

   使用する辞書の最大数は 5 個に設定されている．これを変更したい場合は，配布パッケージの lib/pat.h の MAX_DIC_NUMBER の値を変更してからインストールを行えばよい．

3. **形態素のコストの定義**

   4 章で述べるように，形態素解析プログラムでは，解析結果の優先情報をコストとして計算している．解析に曖昧性がある場合は，コストの総計が低いものを優先することになっている．文法辞書によって定義された形態素のコストを定義しなければならない．

   品詞は，品詞名，または，品詞名と品詞細分類名の対によって表わされる．品詞細分類が定義されている品詞については，品詞細分類名 (または '*') を指定する必要がある．コストは正整数の値をとる．

```
(品詞コスト
        ((*)                10)
        ((特殊 *)          100)
        ((動詞)            100)
        ((形容詞)          100)
        ((判定詞)           11)
        ((助動詞)           10)
        ((名詞 *)          100)
        ((名詞 数詞)        40)
        ((名詞 形式名詞) 70)
        ((名詞 副詞的名詞) 70)
        ((指示詞 *)         40)
        ((指示詞 副詞形態指示詞)          60)
        ((副詞 *)          100)
        ((助詞 *)           10)
        ((助詞 終助詞)     20)
```

[3] 一つの形態素辞書には同一の形態素の登録は行われないが，複数の形態素辞書に同じ形態素が登録されている場合はあり得る．このような場合は，同じ形態素が複数得られることになる．

```
        ((接続詞)          100)
        ((連体詞)          100)
        ((感動詞)          100)
        ((接頭辞 *)         50)
        ((接尾辞 名詞性述語接尾辞)        14)
        ((接尾辞 名詞性名詞接尾辞)        35)
        ((接尾辞 名詞性名詞助数辞)        35)
        ((接尾辞 名詞性特殊接尾辞)        35)
        ((接尾辞 形容詞性述語接尾辞)      14)
        ((接尾辞 形容詞性名詞接尾辞)      14)
        ((接尾辞 動詞性接尾辞)            14)
        ((未定義語 カタカナ)              5000)
        ((未定義語 アルファベット)        100)
        ((未定義語 その他)                5000)
)
```

同じ品詞に対してコストの定義が複数回指定されている場合は，後のものが優先される．上の例では，品詞名が「名詞」の形態素のコストは 100 であるが，細分類名が「数詞」の形態素のコストは 40 となる．また，先頭の ‘(*)’ の指定により，ここで明示的に定義されていない形態素のコストはすべて 10 となる．

4. **連接コストと形態素コストの相対的な重みの定義**

   形態素解析におけるコストの計算は形態素のコストと連接のコストの総計として計算される．これら二種類のコストに異なる重みを掛けたい場合には，それを指定することができる．解析結果のコストはそれぞれのコストにここで指定された重みを乗じた値の総計として計算される．

   ```
   (連接コスト重み 4)
   (形態素コスト重み 1)
   ```

   また，形態素解析の過程において，常にコストが最低の結果を出すのではなく，ある程度のコスト幅を許容したい場合がある．この許容幅を指定することができる．

   ```
   (コスト幅 0)
   ```

   コスト計算については，4 章を参照のこと．

* **未定義語の品詞定義**

バージョン 3.4 以降ではリソースファイルでの未定義語の品詞定義は不必要となった．

4 形態素解析

JUMAN は，UTF-8 コードの日本語文字列を標準入力から一行ごとに読み込んで入力とし，連接規則辞書によって許容された形態素からなる束 (lattice) 状の構造を出力とする．

4.1 形態素解析アルゴリズム

JUMAN の解析アルゴリズムは，入力として与えられた日本語の文字列に対する次の基本動作よりなる．改行をもって一つの入力文字列の終了とする．

- ある特定の位置から始まるすべての可能な形態素を辞書引きによって得る．
- 辞書引きによって得られた個々の形態素に対して，その直前の位置に存在するすべての形態素との連接可能性のチェック，および，コストの計算を行う．
  - 連接可能性のチェックによって連接不可能と分かった形態素間の接続は行われない．また，その位置で連接可能なもののうち最良 (最小) のコストと比較して jumanrc の (コスト幅 ??) によって定義される数値以上ののコスト差をもつ形態素の接続は行われない．

本形態素解析では次の二種類のコストが使用される．

形態素のコスト: 個々の形態素に与えられているコストである．形態素のコストは，jumanrc で定義された「品詞コスト」,「形態素コスト重み」，および，形態素辞書で定義された見出し語のコストに対する重み，の三つの数値の積として計算される．

連接コスト: 二つの形態素の連接のコストである．連接規則辞書で定義されたコストと jumanrc で定義された「連接コスト重み」の積として計算される．

形態素解析の一つの解析結果は形態素からなる列であり，その総コストは，それに含まれる形態素のコストの総計，および，各形態素間の連接コストの総計の和である．ただし，列の先頭と末尾には，それぞれ,「文頭」および「文末」と呼ばれるダミーの形態素があると仮定する．

一般に解析結果は，文頭および文末を両端点とする束状のグラフ構造になる．このグラフ内の文頭から文末へいたるそれぞれの経路が一つの可能な解析結果を表わすことになる．

連接可能性のチェックおよびコストの計算の詳細について説明する．最初の辞書引きは文字列の先頭で行われ，その直前には「文頭」が存在すると仮定される．また,「文頭」がもつコストは 0 である．

最初に辞書引きが行われるのは入力文字列の先頭 (すなわち,「文頭」の直後) である．以降，辞書引きが行われるのは入力文字列中において文頭からの有効な連接をもつ形態素の直後の位置である．形態素の有効性，無効性については，以下の説明の中で述べる．

ある位置で辞書引きによって得られるすべての形態素を $Mf_j(1 \leq j \leq n)$ とし，その位置の直前で終了する形態素が $Mb_i(1 \leq i \leq m)$ であるとする．各 Mb_i は文頭から自分自身へいたる束状のグラフの経路の中で最小のコストの値を保持している．そのコストを Cb_i とする．

それぞれの Mf_j について，すべての Mb_i との連接可能性のチェックを連接行列により行う．Mb_i の後接情報と Mf_j の前接情報により，それらが連接可能であるか，また，連接可能であればその連接コストを求める．Mb_i と Mf_j の連接コストを $C_{i,j}$ とする (連接不可能な場合 $C_{i,j} = \infty$ とする)．したがって，Mf_j に対するすべての連接コストは，

$$Cb_1 + C_{1,j},\ Cb_2 + C_{2,j},\ \cdots, Cb_m + C_{m,j}$$

となる．これらの最小値からの差が jumanrc の「コスト幅」以内であるものだけが Mf_j に連接可能であると考えられる．連接可能な Mb_i が存在しない場合 (すべての $C_{i,j}$ が無限大の場合)，Mf_j は無効な形態素と考えられ，以後の解析結果のグラフには含まれない．Mf_j が無効でないと

き，Mf_j は 最小の $Cb_i + C_{i,j}$ の値に自身の形態素コストを加えた値を解析結果のコストとして保持することになる．

形態素解析の最後の段階では，入力文字列の末尾に現れる有効な形態素と「文末」との連接可能性のチェックとコスト計算が行われ，解析結果としてのグラフ構造が得られる．

出力としてはコスト最小 (それぞれの形態素の区切りで最小コストとの差が許容されるコスト幅以内) の解を求め，結果をオプションに従って表示する (オプションについては 6 節)．

4.2 未定義語の取り扱い

本システムでは，入力文字列中のあらゆる位置で未定義語が存在する可能性を考慮している．平仮名および漢字については一文字ずつを一語の未定義語としてきりだす．それ以外の文字については，同種の文字 (カタカナ，アルファベット，数字 等) の終りまでをまとまった一語の未定義語とする．そして，カタカナ文字列には「カタカナ」，アルファベット文字列には「アルファベット」，それ以外には「その他」という品詞細分類を与える．

未定義語という品詞と，その細分類である「カタカナ」,「アルファベット」,「その他」は形態品詞辞書で定義しておかなければならない．連接関係を連接規則辞書で定義すること，コストをリソースファイルの「品詞コスト」欄で定義することは通常の品詞と同様である．

5 インストール方法

インストール方法について説明する．

5.1 Unix 系 OS におけるインストール方法

`juman-7.0.tar.gz` を展開し，`juman-7.0` ディレクトリに移動して，次の手順を実行する．

1. '`./configure`' を実行する．
   '`./configure`' にオプションを与えることにより，インストール先やコンパイルオプションなどを変更することができる．詳細は，同梱の INSTALL または '`./configure --help`' の出力を参照のこと．

2. 'make' を実行する．
   JUMAN システムのコンパイルと辞書の構築が行われる．

3. (root で)'make install' を実行する．
   JUMAN システムと辞書がインストールされる．インストールされる場所は以下のとおりである．`$PREFIX` は，デフォルトでは`/usr/local` であるが，'`./configure --prefix`' で設定できる．

$PREFIX/bin/juman	実行ファイル
$PREFIX/share/juman/dic/	辞書
$PREFIX/etc/jumanrc	設定ファイル
$PREFIX/share/juman/doc/manual.pdf	マニュアル
$PREFIX/libexec/juman/	辞書作成プログラム
$PREFIX/lib/libjuman.*	ライブラリ
$PREFIX/include/juman*.h	ヘッダファイル

4. 品詞コストなどを個人で設定したい場合は，$PREFIX/etc/jumanrc を$HOME/.jumanrc としてコピーし，それを編集する．

5.2 Windows におけるインストール方法

juman-7.0.exe を実行し，表示される指示に従ってインストールを行う．以下のファイルがインストールされる．$PREFIX は，デフォルトでは C:¥Program Files¥juman であるが，インストール途中で設定できる．

$PREFIX/juman.exe	実行ファイル
$PREFIX/dic/	辞書
$PREFIX/dic/jumanrc	設定ファイル
$PREFIX/manual.pdf	マニュアル

現在の Windows 版は試験版であり未対応の機能がある．UNIX 版との違いは次のとおりである．

- 入出力は SJIS で行う．
- jumanrc での種々の設定はできない．辞書も一ヶ所 (1 フォルダ) しかもてず，ここに文法辞書と形態素辞書の双方をおく．辞書の場所の設定は juman.ini で行う．
- サーバ・クライアント機能は備えていない．

6 JUMAN の使用法

形態素解析の実行ファイルは juman という名前でインストールされている．オプション等の指定によって，この一つの実行ファイルをスタンドアロンモード (JUMAN3.4 以前のモード) とサーバ・クライアントモード (JUMAN3.5 以降で追加されたモード) で使い分けることができる．各モードの使用法を以下に説明する．

6.1 スタンドアロンモードでの使用

juman を実行する．標準入力から解析するテキストを一行一文として読み込み，解析結果を標準出力に出力する．このモードは，次節の方法でサーバ・クライアントモードを明示的に指定しない場合のデフォルトのモードである．

形態素解析に必要な文法・辞書の存在場所，ならびに解析中で用いられるコストに関する基本的な情報はリソースファイルから読み込まれる．リソースファイルを探す順番は以下の順である．

1. 実行時に-r オプションで指定されたファイル.
2. ユーザのホームにある.jumanrc ファイル.
3. $PREFIX/etc/jumanrc
4. 上記のいずれにもファイルが存在しない場合にはプログラムは強制終了される.

スタンドアロンモードでは以下のオプションが用意されている.

```
juman -[b|B|m|p|P] -[f|c|e|e2|E] [-i string] [-r rc_file] [-u [option]]
```

- 解が曖昧性を含む場合の表示方法 (曖昧性がない場合はどの方法も同じ表示となる)
  - -b : 後方最長一致の解を一つだけ表示する.
  - -B : -b オプションと同様に後方最長一致の解を表示するが, 同じ位置に複数の形態素がある場合はそれらをすべて表示する. ある位置からの 2 つ目以降の形態素は "@" をつけて表示する (default).
  - -m : 曖昧性のある部分だけ, 複数の形態素を表示する.
  - -p : 曖昧性の組合せを展開し, すべての解を個別に表示する.
  - -P : -p と -B を合わせたもの. すなわち-p と同様に曖昧性の全組合せを示すが, 同じ位置の複数の形態素は-B のように "@" でまとめて表示する.
- 各形態素の表示方法
  - -f : カラムを整えて表示.
  - -c : 完全な形態素情報をコードで表示.
  - -e : 完全な形態素情報を文字とコードで表示.
  - -e2 : -e の情報に加えて意味情報を表示 (default).
  - -E : -e の情報に加えて各形態素の文中での位置と意味情報を表示.
- その他
  - -i string : string で始まる入力行は解析せずにそのまま出力.
  - -r rc_file : rc_file をリソースファイルとして使用.
  - -u [option] : 連濁処理, オノマトペの自動認識等の未知語処理を行わない. option=[rendaku|lowercase|long-sound|onomatopoeia] を指定した場合は, それぞれ, 連濁, 小書き文字, 長音記号, オノマトペへの対処のみ行わない.
  - -h : ヘルプメッセージを出力.

以下にいくつかのオプション指定による使用例を示す.

```
% juman
私は昨日学校を休んだ
私          (わたし)      私          普通名詞
は          (は)          は          副助詞
昨日        (きのう)      昨日        時相名詞
学校        (がっこう)    学校        普通名詞
を          (を)          を          格助詞
休んだ      (やすんだ)    休む        動詞              子音動詞マ行    タ形
EOS
```

```
% cat > test
子どもはリンゴがすきだ
% juman < test
子ども こども 子ども 名詞 6 普通名詞 1 * 0 * 0 "代表表記:子供"
は は は 助詞 9 副助詞 2 * 0 * 0 NIL
リンゴ りんご リンゴ 名詞 6 普通名詞 1 * 0 * 0 "代表表記:林檎"
が が が 助詞 9 格助詞 1 * 0 * 0 NIL
すきだ すきだ すきだ 形容詞 3 * 0 ナ形容詞 21 基本形 2 "代表表記:好きだ"
EOS
```

※ JUMAN5.0 以降，代表表記を設定し，意味情報として出力することとした (オプションとして -e2 を指定)．詳細については E.12 節を参照．

6.2 サーバ・クライアントモードでの使用

JUMAN3.5 以降ではサーバ・クライアントモードでの使用が可能となった．このモードでは，あるマシンで形態素解析を行うサーバを起動しておき，他のマシン (同一マシンでも可) でテキストの入出力とサーバとの通信だけを行うクライアントを起動する．ネットワーク環境での複数ユーザによる利用などに適したモードである．

6.2.1 サーバの起動

オプション-S を指定して juman を実行することによりサーバとして起動される．

この場合のリソースファイルの探索順はスタンドアロンモードの場合と同様である．ただし，この時得られるリソース情報 (文法ファイルの位置，解析のコストなど) はクライアント側でリソースファイルが読み込まれない場合にのみ有効となる．

サーバに対するオプション指定は，サーバであることを示す-S オプションが追加されることを除いてスタンドアロンモードと同様である[4]．

6.2.2 クライアントの起動

オプション-C hostname を指定して juman を実行することによりクライアントとして起動される．hostname はサーバを起動したマシン名である．

もう一つの方法として，サーバを起動したマシン名を環境変数$JUMANSERVER に設定しておく方法がある．この環境変数が設定されていればオプションの指定なしでもクライアントとして起動される[5]．

クライアントとして起動された場合のリソースファイルの探索順は次のようになる．

1. 実行時に-r オプションで指定されたファイル．
2. ユーザのホームにある.jumanrc ファイル．

[4]サーバ・クライアント間の通信に用いられるポート番号のデフォルト値は 32000 である．この変更は-N number というオプションで行う．通常は変更の必要はない．

[5]サーバ・クライアント間の通信に用いられるポート番号がデフォルト値でない場合は-C オプションの場合も環境変数 $JUMANSERVER で指定する場合も，サーバのマシン名の後に (空白を入れずに)':' をはさんでポート番号を指定する．

3. 上記のいずれにもファイルが存在しない場合はサーバ側で指定されているリソース情報が使用される．

クライアントに対するオプション指定は，クライアントであることを示す-C オプションが追加されることを除いてスタンドアロンモードと同様である．

6.3 Perl からの利用

Perl モジュール “Juman” を用いることにより，Perl から JUMAN を利用することができる．

- インストール方法

  JUMAN を展開したディレクトリにある perl ディレクトリに移動し，次の手順を実行する．

  1. perl Makefile.PL
  2. make
  3. (root で) make install

- 使用例

  プログラムの例を以下に示す．このプログラムを実行すると,「この文を形態素解析してください.」という文の解析結果 (デフォルトのオプションは “-B -e2”) が表示される．

```
use Juman;
$juman = new Juman;
$result = $juman->analysis( "この文を形態素解析してください．" );
print $result->all();
```

  詳細は ‘perldoc Juman’ の実行結果を参照のこと．

参考文献

[1] 益岡隆志，田窪行則：『基礎日本語文法』くろしお出版，1989.

[2] 妙木裕，松本裕治，長尾眞：「汎用日本語辞書および形態素解析システム」情報処理学会第42 回全国大会予稿集，1991.

[3] 中村俊久，黒橋禎夫，長尾真「部分文字列情報の利用による日本語単語の高速検索」情報処理学会自然言語処理研究会NL-101, 1994.

[4] 山地 治，黒橋 禎夫，長尾 真「連語登録による形態素解析システム JUMAN の精度向上」言語処理学会 第 2 回年次大会1996.

付録

A　ユーザ辞書の記法の定義と例

ユーザ辞書の記法を BNF によって示す．以下，〈···〉は非終端記号名を表わす．〈#形態品詞名〉のように#が付けられた非終端記号はユーザによって定義される具体的な名前を表わすと仮定する．NIL は空リストではなく，空系列を表わす．

A.1　形態品詞分類辞書 (JUMAN.grammar) の記述

A.1.1　形態品詞分類辞書の個々の項目の記述

〈形態品詞分類辞書項目〉　::=　（〈形態品詞〉）| （〈形態品詞〉（〈品詞細分類の並び〉））

〈形態品詞〉　::=　（〈#形態品詞名〉）| （〈#形態品詞名〉 %）
〈品詞細分類の並び〉　::=　〈品詞細分類〉〈品詞細分類の並び〉
〈品詞細分類〉　::=　（〈#品詞細分類名〉）| （〈#品詞細分類名〉 %）

A.1.2　形態品詞分類辞書の定義例

```
((動詞 %))                          ; 動詞は形態品詞名であり，活用する．

((名詞)                             ; 名詞は形態品詞であり，
        ((普通名詞)                 ; 普通名詞，サ変名詞，固有名詞 ... に
         (サ変名詞)                 ; 細分類される．
         (固有名詞)
         (数詞)
         (形式名詞)
         (副詞的名詞)))

((接尾辞)
        ((名詞性述語接尾辞)
         (名詞性名詞接尾辞)
         (名詞性名詞助数辞)
         (形容詞性述語接尾辞 %)     ; 形容詞性述語接尾辞は活用する．
         (形容詞性名詞接尾辞 %)
         (動詞性接尾辞 %)))
```

A.2 活用関係辞書 (JUMAN.kankei) の記述

A.2.1 活用関係辞書の個々の項目の記述

〈活用関係辞書項目〉 ::= (〈品詞〉 (〈活用型の並び〉))

〈品詞〉 ::= (〈#形態品詞名〉) | (〈#形態品詞名〉〈#品詞細分類名〉)
〈活用型の並び〉 ::= 〈#活用型名〉 | 〈#活用型名〉〈活用型の並び〉

A.2.2 活用関係辞書の定義例

```
((形容詞                        ; 形容詞はイ形容詞とナ形容詞という活用型をもつ
        (イ形容詞
         ナ形容詞))

((助動詞)
        (イ形容詞
         ナ形容詞
         判定詞
         無活用型
         助動詞ぬ型
         助動詞だろう型
         助動詞そうだ型))

((接尾辞 形容詞性述語接尾辞)   ; 接尾辞の中で形容詞性述語接尾辞に細分類される
        (ナ形容詞                ; 形態素はナ形容詞またはイ形容詞という活用型を
         イ形容詞))              ; もつ
```

A.3 活用辞書 (JUMAN.katuyou) の記述

A.3.1 活用辞書の個々の項目の記述

〈活用辞書項目〉　::=　(〈#活用型名〉(〈活用形対の並び〉))

〈活用形対の並び〉　::=　〈活用形対〉| 〈活用形対〉〈活用形対の並び〉
〈活用形対〉　::=　(〈#活用形名〉〈語尾表示〉)
〈語尾表示〉　::=　〈#語尾〉 | *

A.3.2 活用辞書の定義例

```
(母音動詞
    ((語幹              *   )
     (基本形            る  )
     (未然形            *   )
     (意志形            よう)
     (命令形            ろ  )
     (命令形            よ  )
     (基本条件形        れば)
     (基本連用形        *   )
     (タ形              た  )
     (タ系条件形        たら)
     (タ系連用テ形      て  )
     (タ系連用タリ形    たり))
)

(サ変動詞
    ((基本形            する  )
     (未然形            さ    )
     (意志形            しよう)
     (命令形            しろ  )
     (命令形            せよ  )
     (基本条件形        すれば)
     (基本連用形        し    )
     (タ形              した  )
     (タ系条件形        したら)
     (タ系連用テ形      して  )
     (タ系連用タリ形    したり))
)
```

A.4 連接規則辞書 (JUMAN.connect) の記述

A.4.1 連接規則の記述

〈連接規則〉 ::= （（〈形態素構造の並び〉）（〈形態素構造の並び〉）） |
（（〈形態素構造の並び〉）（〈形態素構造の並び〉）〈#連接コスト〉）
〈形態素構造の並び〉 ::= 〈形態素構造〉| 〈形態素構造〉〈形態素構造の並び〉
〈形態素構造〉 ::= （〈#形態品詞名〉）|（〈#形態品詞名〉〈#品詞細分類名〉） |
（〈#形態品詞名〉〈#品詞細分類名〉〈#活用型名〉） |
（〈#形態品詞名〉〈#品詞細分類名〉〈#活用型名〉〈#活用形名〉） |
（〈#形態品詞名〉〈#品詞細分類名〉〈#活用型名〉〈#活用形名〉〈#見出し語〉）

〈#形態品詞名〉〈#品詞細分類名〉〈#活用型名〉〈#活用形名〉〈#見出し語〉については，*も指定できる．
〈#連接コスト〉は 0 から 255 までの整数値 (省略された場合はのデフォルト値は 10)．

A.4.2 連接規則の定義例

```
(((名詞)                          ; 名詞，指示詞の名詞形態指示詞などは
  (指示詞 名詞形態指示詞)         ; 判定詞および特殊の読点と連接可能
  (接尾辞 名詞性述語接尾辞)
  (接尾辞 名詞性名詞接尾辞)
  (接尾辞 名詞性名詞助数辞))
        ((判定詞)
         (特殊 読点))
)

(((名詞 サ変名詞)                        ; サ変名詞と接尾辞の「化」には，
  (接尾辞 名詞性名詞接尾辞 * * 化))      ; 「する」「できる」などが
        ((動詞 * サ変動詞 * する)        ; 後接可能である．
         (動詞 * 母音動詞 * できる)      ; また，その連接のコストは 5 であり
         (動詞 * 母音動詞 * 出来る))     ; 結合度が強い．
        5 )
```

B　形態素辞書の記述

B.1　形態素の記述

〈形態素定義〉　::=　(〈#形態品詞名〉〈形態素情報の並び〉) |
　　(〈#形態品詞名〉(〈#品詞細分類名〉〈形態素情報の並び〉))
〈形態素情報の並び〉　::=　〈形態素情報〉 | 〈形態素情報〉〈形態素情報の並び〉
〈形態素情報〉　::=　(〈見出し語情報〉〈読み情報〉〈活用型情報〉〈意味情報〉)
〈見出し語情報〉　::=　(見出し語 〈見出し語内容の並び〉)
〈見出し語内容の並び〉　::=　〈見出し語内容〉 | 〈見出し語内容の並び〉
〈見出し語内容〉　::=　〈#見出し語表記〉 | (〈#見出し語表記〉) |
　　(〈#見出し語表記〉〈#数値〉)
〈読み情報〉　::=　(読み 〈#読み表記〉)
〈活用型情報〉　::=　(活用型 〈#活用型名〉) | NIL
〈意味情報〉　::=　(意味情報 〈#意味記述〉) | NIL

〈活用型情報〉については，活用すると定義された形態素に対しては省略することができない．
〈意味情報〉は，省略可能．

B.2　形態素の記述例

```
(名詞
  (普通名詞
    ((見出し語 日本語)
     (読み にほんご))
    ((見出し語 英語)
     (読み えいご))
  )
)

(動詞
    ((見出し語 (歩く) (あるく 1.5))
     (読み あるく)
     (活用型 子音動詞カ行)
     (意味情報 "walk(X)")
    )
)
```

C リソースファイル (jumanrc) の記述例

```
(文法ファイル /usr/local/share/juman/dic)

(辞書ファイル /usr/local/share/juman/dic)

(品詞コスト
   ((*)                10)
   ((特殊 *)          100)
   ((動詞)            100)
   ((形容詞)          100)
   ((判定詞)           11)
   ((助動詞)           10)
   ((名詞 *)          100)
   ((名詞 数詞)        40)
   ((名詞 形式名詞) 70)
   ((名詞 副詞的名詞) 70)
   ((指示詞 *)         40)
   ((指示詞 副詞形態指示詞)          60)
   ((副詞 *)          100)
   ((助詞 *)           10)
   ((助詞 終助詞)    20)
   ((接続詞)          100)
   ((連体詞)          100)
   ((感動詞)          100)
   ((接頭辞 *)         50)
   ((接尾辞 名詞性述語接尾辞)        14)
   ((接尾辞 名詞性名詞接尾辞)        35)
   ((接尾辞 名詞性名詞助数辞)        35)
   ((接尾辞 名詞性特殊接尾辞)        35)
   ((接尾辞 形容詞性述語接尾辞)     14)
   ((接尾辞 形容詞性名詞接尾辞)     14)
   ((接尾辞 動詞性接尾辞)            14)
   ((未定義語 カタカナ)              5000)
   ((未定義語 アルファベット)        100)
   ((未定義語 その他)                5000)
)

(連接コスト重み 4)

(形態素コスト重み 1)

(コスト幅 0)
```

D JUMAN サーバのプロトコルの概要

JUMAN サーバに接続すると，最初の 1 行に JUMAN やサーバのバージョンが出力される．現在の JUMAN では，入出力ともに，文字コードは UTF-8，改行コードは LF(ASCII 10) となっている．

JUMAN サーバと通信するためのコマンドを以下に示す．

RUN [options]

形態素解析を実行する．

RUN を，必要があれば後ろにオプションをつけて入力すると '200 OK' と出力される．オプションは JUMAN 起動時のものと同じである．

つづいて，日本語の文章を UTF-8 で入力すると，改行を入力するごとにその行の解析結果が出力される．

入力は，Ctr-k (ASCII 0b) のみの行を入力することで終了する．その結果，終了のしるしとして ASCII 0b と改行が出力されるが，これは普通の端末では空行 2 行が出力されたようになる．

解析中に異常が発生した場合は，'500' の後に，エラーメッセージが表示される．

RC

設定ファイルの読み込みを行う．

RC を入力した後，設定内容 (jumanrc ファイルの内容) を入力する．入力は，Ctr-k (ASCII 0b) のみの行を入力することで終了する．

設定の読み込みが正常終了した場合は，'200 OK' と出力される．設定を誤った書式で入力した場合は，サーバプログラムは処理を中止し接続が切れる (より親切なエラーメッセージに変更の予定).

HELP

簡単な説明を表示する．

QUIT

接続を切断する．

E JUMAN の更新履歴

E.1 JUMAN 0.8 から JUMAN 1.0 への拡張点

1. ICOT 辞書との結合により語彙総数が大幅に増加した (異なり形態素数約 13 万語)
2. 複数の形態素辞書を同時に使用できるようにした．
3. MCC 自然言語処理研究グループにより辞書引きの高速化のためのハッシング機能が追加された．
4. 解析中に使用するコストの計算方法を統一的なものにした．

- 形態素自身にコストを持たせた.
- 連接コストを連接規則個別に定義できるようにした.
- 形態素コストと連接コストに重みを付けることができるようにした.
- 形態素間の連接点で許容可能なコストの幅を指定できるようにした.

5. 複数の見出し語 (かな表記を含む) をもつ形態素を単一の形態素として記述できるようした. また，それぞれの見出し語に相対的な重みを記述できるようにした.
6. 連接規則辞書中でマクロ定義を行えるようにした.
7. 個人ユーザが記述するオプショナルな情報を `.jumanrc` というファイルにまとめて記述するように変更した.
8. C 版のシステムからソケットを介して結果が SICStus Prolog に渡るようになった．これに伴い，Prolog 版の形態素解析プログラムが不要になった.
9. システム標準文法における助詞の分類のうち,「提題助詞」と「取り立て助詞」の区別を止め,「副助詞」に統一した.

E.2 JUMAN 1.0 から JUMAN 2.0 への拡張点

1. UNIX に標準装備のハッシュデータベース ndbm を用い，さらに，疑似的な TRIE 構造を実現することにより，辞書引きを高速化した.
2. システムのコンパイルを従来の `gmake` ではなく，`make` により実行できるようにした.

E.3 JUMAN 2.0 から JUMAN 3.0 への拡張点

(黒橋 禎夫　山地 治　大石 巧　坂口 昌子)

以下の機能拡張，および整備を行った.

1. 連語処理
2. 数詞の連結処理
3. 括弧，空白の透過処理
4. カタカナ語の扱い
5. 解析結果の表示オプションの追加
6. 標準辞書の整備
7. 標準文法の整備
8. 辞書システムの GDBM バージョンの追加

それぞれについて以下で詳しく説明する.

E.3.1 連語処理

形態素並びを連語として一括登録し，連語としての固有のコスト，固有の接続を指定できる機能を追加した．このことの利点は次の 2 点である．

- これまで品詞コスト，連接コストの調整では対処しきれなかった (対処すると副作用を生むような) 解析誤りを，連語の登録によって解決することができる．連語登録にはほとんど副作用がない．
- 2 語の連接可能性を指定するという，これまでの枠組では表現できなかった連接条件を，連語の連接条件として指定することができる．たとえば，ナ形容詞テ形に連語「は (副助詞) + ない (接尾辞)」が接続可能であるというような指定ができる．これを，ナ形容詞テ形と副助詞「は」，副助詞「は」と接尾辞「ない」がそれぞれ連接可能であるとしてしまうと大変な副作用がある．

連語の辞書記述は 'dic' という拡張子をもつファイルで行う．中間辞書，システム形態素辞書への変換はこれまでどおりである (`makeint *.dic; makehd *.int` とする場合，コンパイル作業はこれまでとまったく同じ)．辞書記述は次の形式で行う．

```
(連語
 ((副詞 ((読み より)(見出し語 より)))
  (形容詞 ((読み よい)(見出し語 よい)(活用型 イ形容詞アウオ段)(活用形 *)))
) 0.6 )
```

はじめに形態品詞名に相当するものとして「連語」と指定し，次に連語を構成する各形態素の情報を通常の形態素辞書と同じ形式で並べる．連語内に活用する形態素がある場合は，その活用型と活用形を指定する必要がある (このとき見出し語と読みには基本形を与える)．連語の末尾が活用する形態素で，その活用形が何でもよいという場合 (上の例で「よりよい」だけでなく「よりよく」，「よりよき」なども可能な場合) は "(活用型 *)" とする．

最後の数字は連語コスト ($0 <$ 連語コスト ≤ 1.0) で，省略された場合は 0.5 となる．連語全体のコストは，連語を構成する各形態素のコストと，各形態素間の連接コストの総和に，この連語コストをかけた値として計算される．

一方，連語固有の接続規則を指定する場合は，次のように行う．

```
(((形容詞 * ナ形容詞 ダ列タ系連用テ形))
        ((連語 * イ形容詞アウオ段 * はない)
         (連語 * イ形容詞アウオ段 * もない)))
```

形態品詞名を「連語」，品詞細分類名を「*」とし，活用型，活用形は連語末尾の形態素の活用型，活用形を与える．連語固有の連接規則の指定がない場合には，左の連接については連語中の先頭の形態素の連接，右の連接については連語中の末尾の形態素の連接がそのまま使われる．

E.3.2 数詞の連結処理

入力文中に数詞 (辞書に数詞として登録されている語) が連続して現れる場合，それらを連結して一語の数詞を出力する機能を追加した．これは標準文法依存の機能である．この機能を削除したい場合には `makefile` 中の `NEUMERIC_P` というフラグを削除すればよい．

E.3.3 括弧，空白の透過処理

括弧は連接する 2 語の間にほぼ任意に現れる，これを括弧とその前後の語との連接規則で規定することはほとんど不可能であった．そこで，括弧は任意の語の間に現れえるとして，その前の語と後の語の連接を調べるという透過処理の機能を設けた．たとえば “書いて」いた” の場合，“書いて” と “いた” が連接可能であるかどうかが調べられる．この時，加算されるスコアは “書いて”，“」”，“いた” の形態素スコア，および “書いて–いた” の連接スコアとなる．これも標準文法依存の機能であるので，この機能を削除したい場合には makefile 中の `THROUGH_P` というフラグを削除すればよい．

なお，通常は許されない連接が括弧が入ることで許されるようになるという現象もあるので，これまでどおり括弧を一形態素として扱う処理 (パス) も残し，そのような現象は括弧を通した連接規則として規定している (詳しくは連接規則を参照)．

一方，丸括弧の場合は一つの括弧ではなく括弧のペア，“（...）” が任意に挿入される．この場合には “解（わか）る” のように 2 語の間ではなく 1 語の内部に挿入される用法もあるので，JUMAN の中で統一的に扱うことは極めて困難である．丸括弧を処理 (削除) するようなプリプロセッサを用意するなどの方法で対処して頂きたい．(標準文法では一応 丸括弧も透過処理の対象としている)

テキスト中の 2 語の間に任意に現れる同じ問題として，空白の問題がある．2.0 版では入力文中の空白は自動的に削除され，形態素解析結果としては空白の存在は何も示されなかった．3.0 版では，空白についても連接の透過処理の枠組で扱うことに変更し，入力文中に空白があれば形態素解析結果としても空白が表示されるようにした．(この情報は形態素解析結果から元テキストを復元する際などに必要となる)．

これに伴い，品詞「特殊」に「空白」という細分類を追加し (`JUMAN.grammar`)，全角空白は「特殊–空白」という品詞の形態素として辞書登録した (`Special.dic`)．半角空白については，JUMAN が基本的に全角文字列に対する処理システムであり，半角文字列の辞書登録，解析をサポートしていなかったため，解析システム (プログラム) 内部で定義しておくという方法をとった．出力としては，全角空白はそのまま “　”，半角空白は “\ ” と表示される．これらの空白処理の機能についても，makefile 中のフラグ `THROUGH_P` を削除することで無効にできる．

E.3.4 カタカナ語の扱い

これまでは入力文中のカタカナ連続を未定義語 (大きなペナルティが与えられる) として切り出していたが，これを通常のサ変名詞として扱うように変更した．また，標準辞書からカタカナのみからなる語を削除した．カタカナ表記の場合たえず新語がつくり出されるので，辞書で対処することは実質上不可能であると判断したためである．

E.3.5 解析結果の表示オプションの追加

以下の解析結果表示オプションとして，`-B`，`-P`，`-E` を追加した (従来からのオプションとともに 6 節で説明)．

E.3.6 標準辞書の整備

語彙総数を増大し (異なり形態素数約 20 万)，また一つの形態素に対しても，次のように種々の表記の見出し語を用意した．

(名詞
  (普通名詞
    ((読み ふりそで)
     (見出し語 振り袖 (振りそで 1.6) (ふり袖 1.6) (ふりそで 1.6)))))

しかし，平仮名を含む表記は解析誤りの原因ともなるので，名詞に関しては以下のような制限を設けた．

- 基本的な漢字 (小学校 1,2 年で習う漢字) を平仮名書きする表記は作成しない．
- 漢字の部分を平仮名書きする表記に対しては，その長さに応じて 1 文字の場合 3, 2 文字の場合 2.0, 3 文字以上の場合 1.6 という相対重みを与える．

3.0 標準辞書のコンパイル時間は SPARCstation20 上で約 2 時間半，作成される JUMANTREE.dbm.pag のサイズは約 130M バイトである．なお，もちろん JUMAN2.0 標準辞書を JUMAN3.0 システムで用いることも可能である．

E.3.7 標準文法の整備

標準文法の整備を行った．まず，連接規則の精密化を行った．特に，助詞間の連接については大幅な見直しを行った．

また，活用等に関しても，これまで扱っていなかった文語形，省略形 (口語的) などに対応した．

動詞	タ系連用チャ形	ちゃ	(書いちゃ 困る)
動詞	音便条件形	りゃ	(泳ぎゃ 直る)
イ形容詞	文語連用形	ゅう	(美しゅう ございます)
イ形容詞	文語基本形	し	(広し といえども)
イ形容詞	文語命令形	かれ	(美しかれ と願う)
ナ形容詞	ダ列タ系連用ジャ形	じゃ	(きれいじゃ ない)
ナ形容詞	ダ列文語連体形	なる	(雄大なる アフリカ)
ナ形容詞	ダ列文語条件形	なれば	(平和なれば こそ)
ナ形容詞	デス列省略推量形	でしょ	(静かでしょ)
助動詞だろう型	デス列省略推量形	でしょ	(走らない でしょ)
動詞性接尾辞 る	(活用型 母音動詞)		(食べて る，食べて た)
動詞性接尾辞 う	(活用型 子音動詞ワ行)		(書いちゃう，書いちゃった)

判定詞に関しては特殊連体形「の」をたて，他の活用形もふくめて接続を正確なものに修正した．標準的に用意した.jumanrc(jumanrc.sample) では判定詞の品詞コストを 11，助詞の品詞コストを 10 としているので，明確に判定詞である場合 (接続表で規定されている場合) は判定詞，それ以外は助詞と解釈される．たとえば,「学生 の ように〜」「学生 で はない」「学生 な わけがない」などは判定詞と解析される．

判定詞とナ形容詞関係の活用型の違いは次のようにまとめられる．

	基本連体形	特殊連体形	基本連用形
ナ形容詞	な (静かな)	—	に (静かに)
ナノ形容詞	な (大量な)	の (大量の)	に (大量に)
ナ形容詞特殊	な (同じな (ので))	* (同じ)	に (同じに)
判定詞	な (学生な (ので))	の (学生の (ようだ))	—

E.3.8 辞書システムの GDBM バージョンの追加

辞書システムのもとになっているハッシュ・ライブラリ NDBM は OS 依存で，移植性などに問題があったため，別に GDBM のバージョンを用意した．`Makefile` で`# --- Use GDBM File ?`以下の 2 行をコメントアウトすれば NDBM，コメントアウトしなければ GDBM として動作する (配布時は NDBM)．

E.4 JUMAN 3.0 から JUMAN 3.1 への拡張点

1. 奈良先端科学技術大学院大学松本研究室の日本語形態素解析システム『茶筅』の辞書システム (パトリシア) を取り込んだ．これに伴いインストール方法 (5 節参照)，システム辞書名，プログラム名などに変更があるので，注意が必要．

JUMAN3.0		JUMAN3.1
`JUMANTREE.table`	↔	`jumandic.tab`
`JUMANTREE.matrix`	↔	`jumandic.mat`
`JUMANTREE.dbm.dir`		`jumandic.dat`
`JUMANTREE.dbm.pag`	↔	`jumandic.pat`
`JUMANTREE.imis`		
`makehd`	↔	`makepat` (正確には対応しない，辞書のコンパイルは Makefile による)
`.jumanrc` の辞書ファイル指定 ファイル名 (`.../JUMANTREE`)	↔	ディレクトリ名

2. 解析オプションとして，`-i`，`-r` を追加した (6 節参照)．

3. 固有名詞についてコスト調整を行い，ある程度正しく解析されるようにした．

E.5 JUMAN 3.1 から JUMAN 3.2 への拡張点

1. 従来，品詞定義として名詞の細分類に固有名詞，地名，人名があったが，標準辞書ではこれらを区別せずにすべて固有名詞としていた．これを地名，人名，その他 (固有名詞) に分類した．

2. 助詞の細分類の中に名詞接続助詞と述語接続助詞があったが，この分類は不統一であったので，これらをまとめて接続助詞という細分類とした．また,「と同時に」などの複合的な助詞を削除した．

3. 京都大学でのコーパス作成プロジェクトでの使用を通して，標準文法，標準辞書の整備を行った．品詞コストの調整も若干行ったので，個人用に.jumanrc を設定している場合は jumanrc.sample を参照のこと．

E.6 JUMAN 3.2 から JUMAN 3.3 への拡張点

1. 活用形に漢字が含まれる場合に (標準文法では「カ変動詞来」と「動詞性接尾辞得る型」), その読みを活用型の定義に含めるようにした．これによって,「来る」などの読みの部分に漢字が含まれることがなくなった．
2. 長い文を入力しても「Too many morphs」のエラーが出ないよう改善した (形態素の配列の記憶領域の確保を動的なものに変更した).
3. 標準文法および標準辞書の整備.

E.7 JUMAN 3.3 から JUMAN 3.4 への拡張点

1. 固有名詞の分類として組織名をふやし，固有名詞辞書を整備した．(SRI の亀山さん，NYU の関根さんに情報を頂きました．感謝いたします)
2. 未定義語を品詞として JUMAN.grammar で定義し，細分類として「カタカナ」,「アルファベット」,「その他」を設けた.「カタカナ」はカタカナ列,「アルファベット」はアルファベット列,「その他」はそれ以外の同一文字種列 (ただし平仮名，漢字は一文字) がプログラムによって切り出される．これらの接続，重みなどは通常の品詞と同様に JUMAN.connect，.jumanrc などで定義する．
3. 標準文法および標準辞書の整備.

E.8 JUMAN 3.4 から JUMAN 3.5 への拡張点

1. サーバ・クライアント方式への拡張 (学術情報センターの影浦峡先生，吉岡真治先生にご協力頂きました．感謝いたします).
2. プログラム中のリソースファイルの扱いを以下のように整理した.
   - 関数 read_jumanrc_fn → set_jumanrc_fileptr に変更
     (リソースファイルのファイル名 (Jumanrc_File) ではなくファイルポインタ (Jumanrc_Fileptr) を設定)
   - 関数 read_jumanrcfile → set_jumangram_dirname に変更
     (文法ファイルのディレクトリ名を設定，変数名は Jumanpath_rc から Jumangram_Dirname に変更)
   - 関数 read_jumanpathrc → 削除
   - 関数 juman_init → juman_init_rc に変更
     (リソースファイル関係の初期化)
   - 関数 juman_init2 → juman_init_etc に変更

E.9 JUMAN 3.5 から JUMAN 3.6 への拡張点

1. Windows 版の作成.
2. Perl モジュール (Juman.pm) の作成.
3. 連語処理の bug fix.

E.10 JUMAN 3.6 から JUMAN 3.61 への拡張点

1. iotool.c の修正 (`$HOME` が設定されていなくて動作するように)
2. trans.c の修正 (Pentium II マシンで浮動小数の丸め誤差の影響について報告があったので対処)
3. 辞書・接続表などの若干の修正
   (連語「しかいない」,「はしたものの」,「においてよく」など)

E.11 JUMAN 3.61 から JUMAN 4.0 への拡張点

1. 辞書の変更：EDR 日本語単語辞書の利用をやめ，大規模テキスト中の統計量をもとに，約 6 万 8 千語の独自の辞書を構築した．これに関連して次の変更を行った.
   (a) ナ形容詞語幹と名詞が重複するものについては，名詞見出し語を削除し，ナ形容詞語幹の連接を名詞に準ずるものとした.
   (b) 一般的なカタカナ見出し語を与えたので,「未定義語カタカナ」のコストを「未定義語その他」と同じく 5000 とした．これにより，例えば「テレビゲーム」は「テレビ」「ゲーム」となる.
   (c) 形態素辞書の「意味情報」として”可能動詞”の記述を与えた.
2. juman_lib.c の bug fix

E.12 JUMAN 4.0 から JUMAN 5.0 への拡張点

概要

- 日本語の基本的語彙，約 3 万語 (固有名詞を除く) を選定した.
- 表記バリエーションの整備を行い，代用表記を出力することとした.
- その他の整備（読みの音訓情報の付与,「読ます」「読まされる」などの使役形への対応）

E.12.1 見出し語の整備

辞書の整備（表記バリエーション，意味情報付与など）を適切に行っていくために，また，不必要に大きな語彙による副作用（解析誤り）を防ぐために，標準的日本語で用いられる基本的語彙を選定した．ほとんど使われることのない古語などは排除し，複合語についても，構成性原理が成り立つものについては原則的に見出し語から排除した．判断基準は，京都コーパスでの頻度，新聞 2 千万文の自動解析での頻度，国語辞典の見出し語となっているかどうか，などから総合的に判断した．その結果，固有名詞を除く自立語 (名詞，動詞，形容詞，副詞，連体詞，接続詞，感動詞) について，約 3 万語の辞書となった．

（※ JUMAN4.0 からこの方向で整理を行ったが，今回，一応の整理を完成した．）

なお，京都コーパス 4.0 は JUMAN5.0 の語彙に対応している．

判断基準について

句，事典的な語は排除

例）育ての親，断腸の思い
元寇 倭寇 遣隋使 遣唐使 太郎冠者

連用形＋名詞も基本的には排除（相互情報量などで KNP で一語にする必要あり）

例）操り人形 打ち上げ花火 埋め立て地

2 文字のものは原則採用，3 文字以上のものは原則排除

例）印刷機，映写機，航空機...
観光地 原産地...
運動場 競技場...
組合員 銀行員...
観覧車 機関車 救急車...
無重力 無期限...
重金属 重工業...

しかし，以下のようなものは若干採用した．

- 例解小学国語辞典の見出しとなっているもの（の一部）

  例）絵葉書

- 切り方が？なもの

  例）工学部 医学部 全速力 水産物 海産物 林産物

- 音訓の原則（後述）では読みが分からなくなるもの

  例）記念日
  オレンジ色（青色 赤色 黄色 橙色などもあるので）

  しかし，次のようなものは読みが分からなくなっても排除した．

  例）基準日（にち）
  金メダル 金メッキ（かね）

- 構成語が濁音となり，平仮名表記の場合に解析できなくなるもの．

  例）旗印（じるし） 溜まり醤油（じょうゆ） 掛け布団（ぶとん） 毒蜘蛛（ぐも）
  計画倒れ（連用形でも濁音では解析できないので）

  なお，構成語の濁音の平仮名表記は，一部，独立の見出し語とし，代表表記はもとの漢字表記とした．

  例）じるし（印） じょうゆ（醤油） ぶとん（布団） げんか（喧嘩）

- カタカナの判断は難しい．例えば「フルタイム」「パートタイム」「フルセット」は排除,「フルコース」は採用した．

- その他，排除して問題となっているもの

  例）小 (しょう)/悪魔
  ロシア/正教（まさのり）
  有田/焼（未定義語）
  十（じゅう）/進法
  トップ/１０（いちぜろ）
  トップテン（未定義語） ベストテン（未定義語）
  一（いち）/手 一（いち）/局
  万（まん）/国旗

- とりあえず採用したが？のもの

  例）添え

２文字の「〜長」「〜員」などは採用したが，関係解析などで「〜」を独立させたい場合もあり，今後検討する．

例）（経理）部長 （駒ヶ根）署員 （民主）党員

動詞についてはまだ整理が不十分である．

例）取り外す 取り付ける などは採用している．しかし，可能形（取り外せる）はなく，2 語に分割される．

E.12.2 自立語辞書の基本 format の整理

表記バリエーションとコストの扱い，それらの例外処理，代表表記，種々のコメントなどを一元的に管理できるように，自立語辞書の基本 format を定め，JUMAN 辞書はそこから自動生成することとした．

表記バリエーションは JUMAN3.0 以降自動生成していたが，例外処理などは，その結果に対してフィルタを用意しており，若干不透明であった．

また，表記バリエーション・代表表記の整理は今回新たに行ったものである．

この基本 format については，ドキュメントを整理中であり，別の機会に発表する予定である．また，基本 format のデータについては現在のところ公開していない．

E.12.3 表記バリエーションの整理

同じ語の表記バリエーションであることを扱うために，代表表記を設定し，これを意味情報（の一部）として出力することとした (-e2 オプション).

解析例）

```
% cat sample.txt
子どもはリンゴがすきだ
かぜでおくれた

% juman < sample.txt
子ども こども 子ども 名詞 6 普通名詞 1 * 0 * 0 "代表表記:子供"
は は は 助詞 9 副助詞 2 * 0 * 0 NIL
リンゴ りんご リンゴ 名詞 6 普通名詞 1 * 0 * 0 "代表表記:林檎"
が が が 助詞 9 格助詞 1 * 0 * 0 NIL
すきだ すきだ すきだ 形容詞 3 * 0 ナ形容詞 21 基本形 2 "代表表記:好きだ"
EOS
かぜ かぜ かぜ 名詞 6 普通名詞 1 * 0 * 0 "漢字読み:訓 代表表記:風"
@ かぜ かぜ かぜ 名詞 6 普通名詞 1 * 0 * 0 "代表表記:風邪"
で で で 助詞 9 格助詞 1 * 0 * 0 NIL
おくれた おくれた おくれる 動詞 2 * 0 母音動詞 1 タ形 8 "可能動詞 代表表記:送れる"
@ おくれた おくれた おくれる 動詞 2 * 0 母音動詞 1 タ形 8 "代表表記:遅れる"
EOS
```

これによって，日本語処理の（初期段階における）最大の問題である表記揺れの問題を，形態素解析を行うだけである程度取り除くことが可能となる.

また，平仮名書き等による曖昧性も，適切な範囲で複数の可能性（代表表記）を挙げるようにしており，以後の構文解析・意味解析などへの適切な入力を提供することができる（上記の「かぜ」の例など）.

これらの整理は，不必要な大規模語彙について機械的に行っているのではなく，基本語彙について注意深く行っている．そのため，情報検索においても，形態素解析結果の代表表記のマッチングとすることで適合率・再現率を向上させることが期待できる.

この枠組みの中では，次のような現象を扱っている.

漢字と平仮名，送り仮名

漢字とするか平仮名とするか，また，送り仮名のバリエーション．その組み合わせ．(※ JUMAN3.0 から扱っているが，代表表記を設け，再整理した)

例）拳銃 けん銃 拳じゅう けんじゅう
表す 表わす あらわす
落とす 落す おとす

基本的なアルゴリズムは以下のとおり.

- 名詞，ナ形容詞以外はすべてのバリエーションを作り，コスト（ペナルティ）はない.

- 名詞，ナ形容詞は（完全平仮名表記以外で）小学 1〜3 年生学習漢字を平仮名書きする表記は作らない．また，漢字を平仮名書きした表記にはコストを与える．
- 漢字 1 字＋平仮名 1 字の表記は（平仮名部分のもとの漢字が常用漢字であれば）ほとんど使われず，副作用のもとなので作らない

  例）開き（開基） 合い（合意） 厚み（厚身）

漢字別表記

例）狩人／猟人 朝日／旭
色取る／彩る 哀れむ／憐れむ
綺麗だ／奇麗だ 気掛りだ／気懸かりだ

動詞でどこまでまとめるかは微妙な問題であり，一般の国語辞典などでは意味の一致の観点からはまとめすぎである．例えば「下りる/降りる」「下ろす/降ろす」はまとめないこととした．

カタカナ表記のバリエーション

ソフトウェアとソフトウエアのようなカタカナ表記のバリエーションに対しても代表表記を設けた．これは，カタカナ語のコーパス中の出現頻度をもとに，表記バリエーションの自動認識と複合語の自動分割を行い，その結果を人手でチェックして整理した．

日本語固有語のカタカナ表記

動物，植物，食べ物などで，日本語固有語であってもカタカナ表記されるものについて，代表表記のもとに整理した．

例）大根 だいこん ダイコン
餃子 ぎょうざ ギョウザ ギョーザ (最後はカタカナ表記バリエーション)

上記のカテゴリに関わらず，カタカナ表記があるものをまとめた

例）溝 ミゾ みぞ
眼鏡 メガネ めがね
奴 ヤツ やつ

代表表記は，原則として，新聞記事での高頻度表記としたが，この選択には強いこだわりはない（妥当性を主張するものではない）．重要なことは，同じ語の表記集合がまとめられ，これを通してテキストマッチングなどが適切に行われることであり，代表表記はこの集合の ID の役割を果たすものである．しかし，ID を数字などにすることは人間にとって管理しやすいものではないので高頻度の表記を採用した．ただし，現在の代表表記は ID として一意性を持つものではなく，これについては今後さらに整理する予定である．

なお，読みが複数ある漢字表記に対しては現時点では何も特別な扱いはしていない．

例）「旅客機／りょかくき」と「旅客機／りょかっき」は別見出し語
※ただし代表表記はともに旅客機

E.12.4　読みの音訓の情報の付与

常用漢字 1 文字の見出し語に対して読みの音，訓の情報を与えた．

JUMAN では複数の読みを出すが，KNP の最初で，独立の名詞，最初の名詞は訓読み，名詞に続く名詞は音読みを選択することで，大部分の読みは正しく扱える（この原則で対応できない場合は，一部，複合語として見出し語とした）．

例）運動場（じょう），救急車（しゃ），自由米（まい）

E.12.5　「読ます」の形の使役,「読ませる」の形の使役受身への対応

「す」「さす」という動詞性接尾辞を作り対応した．

解析例）

```
読ま よま 読む 動詞 2 * 0 子音動詞マ行 9 未然形 3 "代表表記:読む"
す す す 接尾辞 14 動詞性接尾辞 7 子音動詞サ行 5 基本形 2 NIL
EOS
読ま よま 読む 動詞 2 * 0 子音動詞マ行 9 未然形 3 "代表表記:読む"
さ さ す 接尾辞 14 動詞性接尾辞 7 子音動詞サ行 5 未然形 3 NIL
れる れる れる 接尾辞 14 動詞性接尾辞 7 母音動詞 1 基本形 2 NIL
EOS
知ら しら 知る 動詞 2 * 0 子音動詞ラ行 10 未然形 3 "代表表記:知る"
す す す 接尾辞 14 動詞性接尾辞 7 子音動詞サ行 5 基本形 2 NIL
EOS
知ら しら 知る 動詞 2 * 0 子音動詞ラ行 10 未然形 3 "代表表記:知る"
さ さ す 接尾辞 14 動詞性接尾辞 7 子音動詞サ行 5 未然形 3 NIL
れる れる れる 接尾辞 14 動詞性接尾辞 7 母音動詞 1 基本形 2 NIL
EOS
```

以下，判断基準等の詳細

- 「読ます」は「読ま＋す（接尾辞）」とする.「読ま＋せる（接尾辞）」に近い.「〜＋さ＋れる」とよく使うのはこれに相当する場合が多い.
- 「驚かす」「怒らす」「悩ます」「喜ばす」「悲します」など「人を〜」のものも二語にする.
- 「知らす」「聞かす」も二語，ただし「知らせる」は使役度が極めて低く「伝える」に近いので一語としておく.
- 「励ます」は「励ませる」の意味と「元気づける」の意味があるが，後者が major なので，一語とする.
- 人が関係しない「乾かす」「（たばこ・大臣風を）吹かす」などは一語とする.「（肩を）いからす」も一語とするが，漢字表記の「怒らす」は「おこらす（二語）」への副作用があるのでやめる.
- 「（飛行機を）飛ばす」「（心を）踊らす」は一語とし,「人を飛ばす・踊らす」は（本来二語としたいが）あきらめる．逆に,「働かす」は普通人なので二語とし,「頭を働かす」はあきらめる.

E.13 JUMAN 5.0 から JUMAN 5.1 への拡張点

1. 従来 KNP で付与していた語の意味情報を JUMAN で付与することに変更. (付属動詞候補, 補文ト, ～を～に構成語 など)
2. プログラムでの変数定義の修正 (MacOS X での問題解消)
3. 実行時のデフォルトのオプションを -B -e2 に変更
4. 辞書・接続表の若干の修正

E.14 JUMAN 5.1 から JUMAN 6.0 への拡張点

概要

- 意味カテゴリの付与
- ドメインの付与
- 固有名詞辞書の語彙と意味情報の整理・付与
- 語の意味関係の整理と意味情報の付与
- 連濁，反復形オノマトペ，小書き文字による非標準表記の自動認識
- 未知語の自動獲得
- 意味情報の表記法の整理
- その他の整備（辞書・接続表の若干の修正等）

E.14.1 意味カテゴリの付与

「人」「動物」「植物」「人工物」「抽象物」などの意味カテゴリ 22 種を名詞の意味情報として付与した．以下に各カテゴリの基準，具体例を示す.

人 「幽霊」「神」「河童」「人魚」「半魚人」などの人間に近い生物（架空の生物も含む）を表す単語も含む.

> 学生 先生 歌手 父 兄 大人 子供 赤ちゃん 私 僕 あいつ 我々 誰 何者 個人 主人公 跡取り 逸材 語り手 手先 神 幽霊 故人 河童 人魚 半魚人 妖精

組織・団体 「コンビ」や「トリオ」などの複数人数を表す単語も含む.「我々」などの複数人数を表す人称代名詞は＜人＞に含める.

> 政府 軍 国家 党 委員会 組合 企業 マスコミ 警察 家族 チーム クラス コンビ カップル 一座 一同

動物　「怪獣」などの架空の動物を表す単語も含む.

犬 猫 鳥 ふな めだか 金魚 かえる ほ乳類 虫 さなぎ 恐竜 三葉虫 細菌 微生物 竜 しゃちほこ 怪獣 ペット 愛犬 害虫 天敵 類人猿

植物

木 草 桜 紫陽花 バラ ひまわり 朝顔 チューリップ 稲 盆栽 牧草 一年草 大木 果樹 針葉樹 こけ カビ 切株 国花 まりも

動物-部位　人間の部位も含む.「垢」や「かさぶた」など体に付着している物質はこれに含めるが,「涙」や「血」などの分泌物は＜自然物＞に含める.

手 皮膚 傷 毛 指紋 肉 アラ 尾 羽 くちばし 内臓 筋肉 ほくろ ひげ 爪 たてがみ 甲羅 うろこ 角 牙 骨 関節 かさぶた

植物-部位　＜動物-部位＞と同様に「樹液」などは＜自然物＞に含める.

葉 茎 枝 根 実 種 花片 年輪 落葉 花粉 わら 樹皮 おしべ 菌糸 葉緑素 胞子 葉脈 落葉 つた

人工物-食べ物　「りんご」や「さんま」などの人工物でない食べ物も含む.

料理 アイス パン 菓子 焼き肉 ラーメン 豆腐 コーヒー ワイン 調味料 ソース ごちそう お弁当 主食 大好物 冷凍食品 風邪薬 エサ

人工物-衣類　「眼鏡」や「コンタクト」などは＜人工物-その他＞に含める.

セーター ワイシャツ ズボン スカート レインコート ネクタイ マフラー 靴 アクセサリー 手袋 ハチマキ ベルト

人工物-乗り物

自動車 飛行機 船 自転車 三輪車 ヘリコプター 御輿 エレベーター エスカレーター ソリ 車椅子 観覧車 いかだ 宇宙船 ロケット

人工物-金銭

給料 ボーナス 借金 運賃 謝礼 切手 馬券 つけ お宝 金券 金貨 小判 食券 カード 埋蔵金 遺産 保険金 賄賂 チップ 祝儀 香典 お年玉

人工物-その他　上の４つの人工物のカテゴリに属さない単語.　ただし建築物は＜場所-施設＞に含める.

鉛筆 消しゴム 箸 椅子 テーブル コップ おもちゃ カメラ 時計 鏡 傘 農薬 石鹸 布団 洗濯機 テレビ カーテン 電球 ランドセル 眼鏡

自然物　「山」などは＜場所-自然＞,「津波」などは＜現象-自然＞に含める.(「宇宙」は＜場所-自然＞,「星」は＜自然物＞とする)「アルコール」や「アミノ酸」などの物質名を表す単語も含む.

石 岩 砂 泥 空気 雲 湯気 水滴 炭素 石油 太陽 月 隕石 灰 ちり ほこり けむり 鉱物 さび 宝石 元素 原子 電子 イオン 地下水

場所-施設　＜「門」「天安門」などはこのカテゴリに含むが,「窓」「塀」などは＜場所-施設部位＞とする

ビル マンション 駅 港 遊園地 プール 橋 道路 公園 部屋 台所 風呂 トイレ 庭 門 屋台

場所-施設部位　建築物の部位である単語.ただし,「トイレ」「台所」などの「部屋」を表す単語は＜場所-施設＞とする.

天井 床 壁 屋根 窓 塀 廊下 階段 縁側 席 下座 ベランダ

場所-自然

山 海 池 沼 空 島 森 林 ジャングル がけ 地層 海底 水脈 水溜まり 平野 半島 岬 岸 草むら 砂漠 野原 山頂 火口 山脈 日向 日陰 茂み

場所-機能

上 下 左 右 中 外 前 うしろ 奥 表面 ふち 境界 方向 範囲 頂点 辺り そば あいだ 端 隅 角 隣 先 東 西 向こう 起点 終点

場所-その他

都市 村 里 首都 先進国 海外 全国 世界 天国 生産地 農地 畑 牧場 領土 空き地 隣国 選挙区 いなか 北極 戦場 人込み 行き止まり 現場 吹き溜まり 踏み場 ゴール 本塁

抽象物　抽象物の中でもう一段階細かいカテゴリの検討を行ったが，意味情報としては＜抽象物＞だけを与えている．下記の具体例は細かいカテゴリごと.

《現象-自然》雨 風 霧 霞 地震 津波 噴火 高潮 つらら 高気圧 音 光 開花 紅葉 発芽 刺激 反射 蒸発 火 酸化 凝固 昇華 乾燥 夕焼け 発光 日差し スペクトル

《現象-生命》あくび しゃっくり いびき 出産 病気 風邪 癌 死 頭痛 睡眠 けが 貧血 遺伝 呼吸 排泄 老化 孵化 回復 羽化 声 命

《動作》(物理的な動作を伴った単語) 運動 仕事 電話 コピー 拍手 転倒 移動 到着 通過 貫通 落下 閉鎖 包囲 炊事 洗濯 使用 採取 運転 飲食

《出来事》(出来事を表す単語) 戦争 大会 試合 事故 事件

《様子》(「混乱」や「緊張」などの様子そのものを表す単語,「味」などの述語が付属して様子を表す（「味が良い」など）単語,「長所」などの様子を表す文字（「長」）が付属して様子を表す単語) 混乱 緊張 身勝手 表情 実情 寝相 味 長所 短所 欠点 汚点

《気持ち》(明らかに感情が伴っている単語，また「食欲」なども含む) 愛情 勇気 敵意 敬愛 あこがれ ためらい 嫌悪 遠慮 良心 自信 執着 我慢 同情 感謝 おごり 誠意 忠節 信用 心服 蔑視 哀悼 恥 決心 尊敬 賞賛

《制度・規則》法律 掟 保険 条約 校則 約束 方式 法則 文法 書式 流儀 しきたり マナー 権利 義務 戒律 摂理 ルール 鉄則 規定 政令 契約 公約

《知的生産物》(「物語」「演劇」「学問」「伝統」などの知的生産物を表す単語，ただし，＜制度・規則＞に属する単語は除く) 言語 ニュース 噂 文章 音楽 映像 演劇 学問 芸術 文化 伝承 宗教 名前 故事ことわざ

《力》(「能力」や「五感」なども含む) 能力 魅力 五感 引力 圧力 火力 エネルギー 威力 底力 動力 念力

《抽象-機能》(抽象概念の中でも特に機能的な単語) 理由 原因 結果 目的 関係 対象 条件 基本 基礎 例 内容 相互 概念 対応 由来 基準

《抽象-その他》

思考 評価 誇張 協力 想像 予測 注意 成立 証明 安定 優遇 奨励 遵守 放任 要求 推薦 誘惑 交渉 保証 承諾 許可 妥協 逃げ場 作戦 戦略 思想 思い出 本音 意見 仮説 民意 容疑 主義 主張 魂胆 異心 謀略 案 解答 方針 意志 人心 所存 予想 打算 要約 知識 広告 指示 連絡 証言 告示 評判 証拠 勝利 敗北 捌け口 とっかかり

形・模様

円 球 線 正方形 直角 縦じま ぶち まだら シルエット 凹凸 粒 列 大型 小型 フォーム 十字 いびつ スパイラル ジグザグ 流線形

色

赤 青 黄 緑 ピンク 白 黒 ベージュ

数量 「和」や「差」などの数量の関係や，単位もこれに含む.

複数 多数 和 差 比 速度 番号 ボリューム 面積 余分 速度 勾配 頭数 沢山 無数 数多 以上 以下 回 倍 些細 半分 最大 最小

時間

年 月 朝 晩 時刻 今日 休日 未来 過去 期間 季節 時期 瞬間 チャンス 途端 間髪 順序 先後 永遠 締切 一生 王朝 世代 史上 将来

以下に，複数の意味カテゴリに属する，または判断が難しいと考えられる単語の具体例とその見解を示す.

- 「魚」「野菜」などの＜動物＞や＜植物＞とも＜人工物-食べ物＞とも考えられる単語
  例えば「金魚」という単語には＜動物＞のカテゴリだけを付与するが,「さんま」という単語には＜動物＞と＜人工物-食べ物＞の両方のカテゴリを付与する.
- 「学校」「会社」「市役所」などの＜組織・団体＞とも＜場所-組織＞とも考えられる単語
  両方のカテゴリを付与する.
- 「間」などの＜場所-機能＞とも＜時間＞とも考えられる単語
  両方のカテゴリを付与する.
- 「青二才」「意気地無し」などの＜人＞とも＜様子＞とも考えられる単語
  ＜様子＞のカテゴリだけを付与する.
- 「罠」「右腕」「圧力」「像」などの実体を表す単語が比喩的に抽象概念も表す単語
  例えば「罠」という単語には＜人工物-その他＞のカテゴリだけを付与し,「右腕」という単語には＜人＞と＜動物-部位＞両方のカテゴリーを付与する.

E.14.2 ドメインの付与

意味カテゴリは語の上位下位関係に基づくものであり，これを意味の縦糸と考えるとすれば，意味の横糸として,「文化・芸術」「スポーツ」「健康・医学」「科学・技術」などのドメイン 12 種を設定し，これを語の意味情報として付与した (主に名詞，一部 動詞，形容詞).

これにより，たとえば「土俵」という語には「カテゴリ:場所，ドメイン:スポーツ」,「医者」には「カテゴリ:人，ドメイン:健康・医学」という意味情報が付与されている．以下，各ドメインの基準等を具体例とともに説明する.

文化・芸術 文化，芸術，芸能に関わる単語.「文学」や「美術」などの抽象物を表す語も,「書籍」や「ギター」,「女優」などの具体物を表す語もこのドメインに含める．また,「教会」や「仏壇」などの宗教関係の語もこのドメインに含める．ただし，葬儀関係の語は，＜文化・芸術＞と＜家庭・暮らし＞の両方に含めることとする.

写真 映画 音楽 文学 アニメ 曲 映画 デザイン 展示 映像 美術 芸術 コンサート ピアノ 演出 劇場 楽器 レ コード 芸能 劇 アーティスト 教会 仏壇

レクリエーション 遊びや趣味，娯楽に関わる語．ただし，趣味や娯楽の対象となりうるものであっても，＜レクリエーション＞以外のドメインのいずれかと強く関連するものは除く．例えば,「音楽」は趣味や娯楽となりうるが，＜文化・芸術＞と強く関連するので除外する．同様に,「ゴルフ」は＜スポーツ＞と強く関連するので除外する．一方，囲碁将棋関係は，＜レクリエーション＞と＜スポーツ＞の両方に含める.

遊園地 ゲーム 遊ぶ 旅行 温泉 観光 旅 趣味 パーティー おもちゃ 花火 カラオケ 競馬

スポーツ　スポーツや格闘技に関する語.「サッカー」などのスポーツ名はもちろん,「ドリブル」「ホームラン」などの動作に関わる語や,「ラケット」「土俵」などの道具に関わる語も含める．囲碁将棋関係は，＜レクリエーション＞と＜スポーツ＞の両方に含める.

選手 試合 スポーツ 野球 サッカー レース ボール スキー ゴルフ 競技 対戦 決勝 投手 トレーニング 予選

健康・医学　健康，医療，衛生に関わる単語．このドメインでも,「診察」「予防」などの抽象的なものから,「包帯」「医師」などの具体的なものまで，健康，医療，衛生に関わる単語を広く含める.

医療 病院 患者 感染 癌 医師 ウイルス 診断 症状 看護手術 痛む 予防 薬 風邪 医学 栄養 診療 医者 療法 傷 疾患 ダイエット 歯科 臨床 移植 外科 体重 治る 病気

家庭・暮らし　日常生活に関わる単語．朝起きて，歯を磨き，外出し，帰宅して，風呂に入って寝る．これらの合間に掃除や洗濯，買いもの等の家事を済ませ，子供の面倒を見る．こういった活動に関わる語が全てこのドメインに含まれる．また,「父」「母」「兄弟」「親戚」などの人間関係を現す語もこのドメインに含める．これら以外にも，多くの人がその人生で直面する，結婚や出産，引っ越しなども含まれる．ただし，他のドメインと強く関連する語は除外する．例えば「朝食」は日常生活の一部だが＜料理・食事＞ドメインに含め，このドメインには含めない.「出勤」や「通学」はそれぞれ＜ビジネス＞，＜教育・学習＞に含める．葬儀関係の語は，＜家庭・暮らし＞と＜文化・芸術＞の両方に含めることとする.「インターネット」「パソコン」は＜科学・技術＞とするが,「メール」「ホームページ」は＜家庭・暮らし＞とする.

結婚 出産 引越し 家 家族 住宅 家庭 風呂 暮らす ゴミ トイレ 買い物 保育 洗う 水道 掃除 帰宅 散歩 実家 玄関 世帯 家具

料理・食事　料理，食事，食べ物に関する語．料理名や料理法はもちろん,「箸」「フォーク」などの道具,「レストラン」「料亭」などの店,「炊く」「煮る」などの動作も含める.「たばこ」などの嗜好品もこのドメインに含める.

菓子 食品 食事 料理 箸 味噌 夕食 皿 カフェ 醤油 食べ物 昼食 朝食 炊く 煮る レストラン 冷蔵

交通　陸海空を問わず,「車」「船」「飛行機」などの乗り物,「信号」「標識」などの交通に関わる設備機器,「駐車」「離陸」などの動作,「運転手」「乗客」などの人物,「歩道」「駅」などの場所，その他関連する単語.

駅 道路 交通 運転 空港 航空 鉄道 信号 路線 国道 地下鉄

教育・学習　教育，勉強，学校に関わる語.「先生」「生徒」などの人物や「算数」「国語」などの科目,「成績」「留学」などの抽象的な語をこのドメインに含める．ただし，以下の＜科学・技術＞により強く関連すると考えられる語，例えば「論文」「研究」「学会」などは除外する.

たし算 教育 先生 授業 生徒 学ぶ 勉強 卒業 小学校 教室 テスト 教授 レポート 受験 教科 入学 留学 成績 学科 数学 学年 国語 塾 教材 演習 校長 大学

科学・技術　科学，技術，研究，開発，その他各種理工系専門分野に関わる語.「博士」「学者」などの人物,「解析」「実験」などの活動,「原子」「アンペア」「変数」などの専門用語的なものなどが含まれる.「メール」「ホームページ」は＜家庭・暮らし＞とするが,「インターネット」「パソコン」は＜科学・技術＞とする.

論文 研究 開発 データ 通信 科学 ネットワーク 電子 実験 コンピューター 分析 エネルギー 機械 学会 解析 理論 工学 博士 ロボット 原子 発電 回路 電波 学者

ビジネス　ビジネスあるいは仕事，経済に関する語.「販売」「経営」「契約」などの活動,「社長」「スタッフ」などの人物,「資本」「ニーズ」などの経済用語的なものなどがこのドメインに属する.ただし，個別的な仕事に関する語，例えば「農業」や「水産」などの語は含めない．あくまでビジネスや経済一般に関わる語のみを対象とする.

仕事 企業 販売 商品 経営 価格 産業 株式 市場 働く 営業 注文 メーカー 職員 組合 投資 広告 社長 資金 コスト 就職 株 職業 顧客 資本 需要 証券 退職 貿易

メディア　メディアあるいは報道機関，ジャーナリズムに関する語.「報道」「社説」「論評」などの抽象物以外にも,「キャスター」「アナウンサー」などの人物や「テレビ」「新聞」などの具体物も含める.

ニュース ラジオ テレビ 記事 放送 新聞 番組 メディア 報道 記者 マスコミ

政治　政治あるいは行政，司法，警察あるいは犯罪，福祉，人権，戦争などに関する語.「役所」「兵器」などの具体物,「民主」「法律」などの抽象物,「大統領」「判事」などの人物,「デモ」「投票」などの活動等が含まれる．一方,「合議」「対案」「代案」などは＜政治＞的なニュアンスもあるが，社会生活全般に関わるので＜ドメイン無し＞とする.

司法 行政 政府 軍 税 法律 議員 金融 国家 選挙 警察 裁判 財政 大臣 議会 党 人権 厚生 国会 交渉 テロ 役所 憲法 首相 民主 政権 都道府県 訴訟 逮捕 デモ 国連

名詞であっても，以下のように特定のドメインとの関係がない・薄いものはドメインを与えない.なお，下記の自然，天候などの分類は見通しをよくするために与えたもので，これらに分類される語に常にドメインを与えないという意味ではない.「社長」は人間だが＜ビジネス＞に,「駅」は建造物だが＜交通＞に含める.

自然： 岩 河川 宇宙 津波
天候： 雨 台風 寒気 日照り
人間： 彼女 私 太郎 善人
身体： 目 腕 筋肉 声
感情： 愛 憎悪 不安 歓喜
建造物：ビル 小屋 倉庫
その他：白 明後日 物品 北西 諸々 少量 弱点 合議 対案 代案

一方，複数のドメインに関連しうる語は多いが，いずれかがドミナントである場合はそのドメインだけを与え，複数に同程度に関連すると考えられる語であれば複数ドメインを与える．複数のドメインを与えている例は以下のようなものである．

大学院 → ＜教育・学習＞ ＜科学・技術＞
円高 → ＜ビジネス＞ ＜政治＞
薬膳 → ＜健康・医学＞ ＜料理・食事＞
登山 → ＜スポーツ＞ ＜レクリエーション＞

E.14.3　固有名詞辞書の語彙と意味情報の整理

固有名詞辞書の語彙をいくつかの基準で約 8000 語規模に限定し，種々の意味情報を整理・付与した．

人名 (約 4000 語)　Web の自動解析結果（もちろん解析誤りも含まれる）をもとに，日本人の姓の頻度上位 1500 位まで，名の 2000 位までを抽出・登録した．さらに，順位，相対頻度を意味情報として付与した．

山田 → 人名:日本:姓:7:0.00607
太郎 → 人名:日本:名:45:0.00106

英語名は，姓名の区別なく，Web の頻度上位 150 位までを登録した．

ジョン → 人名:外国

日本の地名 (約 2000 語)　日本の地名については，まず，都道府県，区，市，町までを意味情報とともに登録した．

京都 → 地名:日本:府
→ 地名:日本:京都府:市
上京 → 地名:日本:京都府:区
長岡京 → 地名:日本:京都府:市

※ 郡，村は登録していない．区については，東，西，南，北，中央，中 は登録していない．今後の固有名解析，形態素・構文解析との関係で検討する．

さらに，地方名，主要な地名（銀座など，意味情報は地区）を登録した．

東北 → 地名:日本:地方
銀座 → 地名:日本:地区

また，主要な山，湖，島について，地名＋山などで解析できないものを若干登録した．

比叡山 → 地名:日本:山
琵琶湖 → 地名:日本:湖

例外的なものとして，東名，関越，名神を施設として登録した．

東名 → 地名:日本:施設

世界の地名 (約 700 語)　世界の地名については，国，首都，米国の州，中国の省，各国の主要都市を登録した．また主要な国の別称，略称等も登録した．

英国 → 地名:国
イギリス → 地名:国:別称:英国
ロンドン → 地名:国:英国:首都
英 → 地名:国:略称:英国
ケンブリッジ → 地名:国:英国:市
カリフォルニア → 地名:国:米国:州
四川 → 地名:国:中国:省

主要な地域，海，山などを登録した．意味情報では複数の国にまたがるものは国を与えず，一国内のものは国を与えた．なお，地域は国をこえるもの，地方は国内で市よりも大きなもの，地区は市よりも小さなものとした．

アジア → 地名:地域
太平洋 → 地名:海
アルプス → 地名:山脈
シベリア → 地名:国:ロシア:地方
マンハッタン → 地名:国:米国:地区
グアム → 地名:国:米国:島
キリマンジャロ → 地名:国:タンザニア:山

例外的なものとして，ホワイトハウス，天安門などを施設として登録した．

ホワイトハウス → 地名:国:米国:施設
天安門 → 地名:国:中国:施設

組織名 (約 400 語)　会社，大学，政党などを組織名として登録した．現在は意味情報は与えていない．

パナソニック
京大
民主党

固有名詞 (約 60 語)　商品名，民族名，年号などを固有名詞（人名，地名，組織名以外の固有名詞という意味）として登録した．現在は意味情報を与えていない．

八ツ橋
アングロサクソン
平成

E.14.4　語の意味関係の整理と意味情報の付与

従来より「書ける」が「書く」の可能動詞であるという情報は付与していたが，このような見出し語間の意味関係を整理し，約 3 万語の見出し語について網羅的に情報を付与した（一部，未整理の問題もある）．以下にいくつかの具体例を示す．

尊敬動詞・謙譲動詞
仰る → 尊敬動詞:言う
申し上げる → 謙譲動詞:言う

自動詞・他動詞
壊れる → 自他動詞:他:壊す
壊す → 自他動詞:自:壊れる

授受動詞
貸す → 授受動詞:受:借りる
借りる → 授受動詞:授:貸す

反義
増える → 反義:動詞:減る
大きい → 反義:形容詞:小さい

種々の派生
大人びる → 名詞派生:大人
愚痴る → 名詞派生:愚痴
白い → 名詞派生:白
高める → 形容詞派生:高い
小さな (連体詞) → 形容詞派生:小さい

E.14.5　連濁，反復形オノマトペ，小書き文字による非標準表記の自動認識

連濁，反復形オノマトペ，小書き文字による非標準表記について，辞書にその表記を登録するのではなく，プログラムによって動的に認識を行う．
連濁：「上海ガニ」の解析例

上海 しゃんはい 上海 名詞 6 地名 4 * 0 * 0 "代表表記:上海/しゃんはい 地名:国:中国:市"
ガニ がに カニ 名詞 6 普通名詞 1 * 0 * 0 "代表表記:蟹/かに カテゴリ:動物; 人工物-食べ物 ドメイン:料理・食事 濁音化"
EOS

反復形オノマトペ:「ばくばく食べる」の解析例

ばくばく ばくばく ばくばく 副詞 8 * 0 * 0 * 0 "自動認識"
食べる たべる 食べる 動詞 2 * 0 母音動詞 1 基本形 2 "代表表記:食べる/たべる ドメイン:料理・食事"
EOS

小書き文字による非標準表記:「かわいぃ子供」の解析例

かわいぃ かわいい かわいい 形容詞 3 * 0 イ形容詞イ段 19 基本形 2 "代表表記:可愛い/かわいい 非標準表記"
子供 こども 子供 名詞 6 普通名詞 1 * 0 * 0 "代表表記:子供/こども カテゴリ:人"
EOS

E.14.6 未知語の自動獲得

JUMAN の見出し語を 3 万語規模としているのは，新語・専門用語等への対応は人手で行うのではなく自動獲得によって行うべきであるとの考えに基づいている．

JUMAN6.0 では，未知語の自動獲得（品詞，活用形の推定を含む）を行い，その結果から自動的に構築した辞書 (約 1 万 3 千語) を試験版として付属している．この中には「ググる」「ようつべ」「ドラえもん」などの語がある．

自動辞書は$PREFIX/autodic/Auto.dic である．JUMAN の標準のインストールではこの辞書は設定ファイルでコメントアウトされており使用されない．使用するためには，$PREFIX/etc/jumanrc をユーザのホームに.jumanrc としてコピーし，autodic のコメントアウトをはずす必要がある (3.6 節，6.1 節参照)．

なお，現在の品詞推定では，名詞は普通名詞とサ変名詞の区別だけを行っており，固有名詞と普通名詞の区別は行っていない（本来固有名詞であるものも，普通名詞となっている）．将来的には，固有名詞の区別，さらに意味カテゴリ，ドメインの自動推定も行う予定である．また，与えられたコーパスから未知語の自動獲得を行うプログラムの整備・公開も検討している．

E.14.7 意味情報の表記法の整理

意味情報の表記法の整理を行った．「:」は意味情報の階層を示すものとし，最初の階層が同じものは一つの語に対して一つしか与えず，最初の階層が同じものが複数ある場合は 2 階層目以降を「;」で並列に並べることとする．

加える → 反義:動詞:引く; 動詞:減らす
貯金 → カテゴリ:人工物-金銭; 抽象物

E.15 JUMAN 6.0 から JUMAN 7.0 への拡張点

大きな変更点は，解析のロバスト化と UTF-8 化である．解析のロバスト化は次の三つからなる．

- 非反復形オノマトペ，長音記号による非標準表記，長音記号・小書き文字を用いた長音化の自動認識
- Wikipedia から抽出した辞書の追加
- 自動辞書の改良

システム・辞書の変更点は次のとおりであり，UTF-8 化が大きな変更点である．

- UTF-8 化
- 辞書構築プログラム (makepat.c) のバグ修正
- その他の整備（辞書・接続表の若干の修正等）

以下では，解析のロバスト化に関する三点と UTF-8 化について説明する．

なお，自動辞書と Wikipedia から抽出した辞書は定期的にアップデートし，JUMAN の公式サイト (http://nlp.ist.i.kyoto-u.ac.jp/index.php?日本語形態素解析システム JUMAN) にて配布する予定である．

E.15.1 非反復形オノマトペ，長音記号・小書き文字による長音化・非標準表記の自動認識

連濁，反復形オノマトペ，小書き文字による非標準表記と同様に，非反復形オノマトペ，長音記号による非標準表記 (長音記号による置換)，および，長音化 (長音記号・小書き文字の挿入) について，辞書にその表記を登録するのではなく，プログラムによって動的に認識を行う．

非反復形オノマトペ 以下のパターンに適合した文字列を非反復形オノマトペの候補とする．パターン中のH，K，Yはそれぞれ平仮名，片仮名，ヤ行拗音字 (平仮名，片仮名を含む) を表す．

- HっHり　　例） もっさり，ざっくり
- HっHYり　例） ぐっちょり，ぺっちゃり
- KッKリ　　例） モッサリ，ドッサリ
- KッKYリ　例） ズッチョリ，ポッチャリ
- KKっと　　例） ピタっと，キュっと
- KKッと　　例） ピタッと，ホロッと

`$PREFIX/lib/const.h` 中の mrph_pattern にパターンとその生成コストを追加することで，自動認識する非反復形オノマトペのパターンを新たに設定することが可能である．

非反復形オノマトペ：「ぺっちゃりしてる」の解析例

```
ぺっちゃり ぺっちゃり ぺっちゃり 副詞 8 * 0 * 0 * 0 "自動認識"
して して する 動詞 2 * 0 サ変動詞 16 タ系連用テ形 14 "代表表記:する/する 付属動詞候補（基本） 自他動詞:自:成る/なる"
る る る 接尾辞 14 動詞性接尾辞 7 母音動詞 1 基本形 2 "代表表記:る/る"
```

長音記号・小書き文字による長音化　長音記号「ー」,「〜」や，小書き文字「ぁ」,「ぃ」,「ぅ」,「ぇ」,「ぉ」が挿入された語はこれらを除いた表記で辞書の検索を行うことにより認識する．

長音記号の挿入：「報告しま〜〜す」の解析例

報告 ほうこく 報告 名詞 6 サ変名詞 2 * 0 * 0 "代表表記:報告/ほうこく 補文ト カテゴリ:抽象物"
し し する 動詞 2 * 0 サ変動詞 16 基本連用形 8 "代表表記:する/する 付属動詞候補（基本） 自他動詞:自:成る/なる"
ま〜〜す ます ます 接尾辞 14 動詞性接尾辞 7 動詞性接尾辞ます型 31 基本形 2 "代表表記:ます/ます 長音挿入"

小書き文字の挿入：「行きたぁあぁい」の解析例

行き いき 行く 動詞 2 * 0 子音動詞カ行促音便形 3 基本連用形 8 "代表表記:行く/いく 付属動詞候補（タ系） ドメイン:交通 反義:動詞:帰る/かえる"
たぁあぁい たい たい 接尾辞 14 形容詞性述語接尾辞 5 イ形容詞アウオ段 18 基本形 2 "代表表記:たい/たい 長音挿入"

長音記号による非標準表記　長音記号「ー」,「〜」により本来の仮名が置換された語は元の表記で辞書の検索を行うことにより認識する．

長音記号による置換：「おはよーございます」の解析例

おはよー おはよう おはよう 感動詞 12 * 0 * 0 * 0 "代表表記:おはよう/おはよう 非標準表記"
ございます ございます ございます 接尾辞 14 動詞性接尾辞 7 動詞性接尾辞ます型 31 基本形 2 "代表表記:御座います/ございます"

E.15.2 Wikipedia から抽出した辞書の追加

Wikipedia のエントリのうち，一形態素である可能性が高いものを JUMAN の辞書に追加した．具体的には以下の条件を満たすものを辞書に追加した．

- JUMAN の解析結果が未定義語一語になるもの

  (名詞 (地名 ((読み ミニストップ)(見出し語 (ミニストップ 1.1))(意味情報 "自動獲得:Wikipedia Wikipedia 上位語:コンビニエンスストア"))))

- JUMAN の解析結果が一文字形態素のみからなるもの[6]

[6]JUMAN の解析では以下のようになる．このような解析になる場合は解析誤りである可能性が高いので，辞書に登録する．
爽 爽 爽 未定義語 15 その他 1 * 0 * 0 NIL
健 けん 健 名詞 6 人名 5 * 0 * 0 ”人名:日本:名:22:0.00134”
@ 健 たけし 健 名詞 6 人名 5 * 0 * 0 ”人名:日本:名:22:0.00134”
@ 健 たける 健 名詞 6 人名 5 * 0 * 0 ”人名:日本:名:22:0.00134”
@ 健 つよし 健 名詞 6 人名 5 * 0 * 0 ”人名:日本:名:22:0.00134”
美 び 美 名詞 6 普通名詞 1 * 0 * 0 ”代表表記:美/び 漢字読み:音 カテゴリ:抽象物”
茶 ちゃ 茶 名詞 6 普通名詞 1 * 0 * 0 ”代表表記:茶/ちゃ 漢字読み:音 カテゴリ:人工物-食べ物 ドメイン:料理・食事”
EOS

(名詞 (普通名詞 ((読み そうけんびちゃ)(見出し語 (爽健美茶 1.1))(意味情報 "自動獲得:Wikipedia Wikipedia 上位語:清涼飲料水"))))

辞書は$PREFIX/wikipediadic/Wikipedia.dic にあり，この辞書を使用しないようにするには $PREFIX/etc/jumanrc をユーザのホームに.jumanrc としてコピーし，wikipediadic の行をコメントアウトすればよい．

コストは以下のように設定し，この辞書を入れることによる副作用が少なくなるようにしている．

- ひらがなを含む: 1.6 (例: おひつ，うっ血)
- アルファベットのみ: 1.0 (例: ＳＰＹ)
- その他: 1.1

意味情報　「自動獲得:Wikipedia」という素性を付与し，定義文から獲得された上位語があれば，Wikipedia 上位語という素性を付与し，また，リダイレクトがあれば Wikipedia リダイレクトという素性を付与する．

品詞細分類　品詞細分類としては普通名詞，人名，地名，組織名を考える．Wikipedia の定義文から上位語を獲得し，その主辞の JUMAN カテゴリにより，品詞細分類を決定する．

見出し語の例	上位語	JUMAN カテゴリ	品詞細分類
ロナウジーニョ	サッカー 選手	人	人名
兼六園	日本 庭園	場所-施設	地名
ＯＤＮ	インターネット プロバイダ	組織・団体	組織名
	...		
(上記にマッチしなければ普通名詞)			
インクィジター	アクション 小説	抽象物	普通名詞

代表表記の付与　Wikipedia において，エントリ A からエントリ B にリダイレクトがあり，カタカナをひらがなに正規化し編集距離が小さい場合に，B の「表記/読み」を代表表記として A，B に付与する．例えば,「スパゲッティ」「スパゲティ」「スパゲティー」に対してすべて代表表記「スパゲッティ/スパゲッティ」が付与される．

(名詞 (普通名詞 ((読み スパゲッティ)(見出し語 (スパゲッティ 1.1))(意味情報 "自動獲得:Wikipedia 代表表記:スパゲッティ/スパゲッティ"))))
(名詞 (普通名詞 ((読み スパゲティ)(見出し語 (スパゲティ 1.1))(意味情報 "自動獲得:Wikipedia Wikipedia リダイレクト:スパゲッティ 代表表記:スパゲッティ/スパゲッティ"))))
(名詞 (普通名詞 ((読み スパゲッティー)(見出し語 (スパゲッティー 1.1))(意味情報 "自動獲得:Wikipedia Wikipedia リダイレクト:スパゲッティ 代表表記:スパゲッティ/スパゲッティ"))))

E.15.3　自動辞書の改良

Wikipedia 辞書の利用にともない，Wikipedia 辞書に収録された語彙は自動辞書から削除した．また，名詞について,「人名」,「地名」などの品詞細分類を自動推定した．ただし，分類精度は高くない．

既知語と獲得語との間の異表記関係を自動認識した．例えば，獲得語「スポーティだ」は既知語「スポーティーだ」,「ばーちゃん」は「ばあちゃん」,「ムレる」は「むれる」の異表記である．異表記関係の自動認識には，表記上の近さと振る舞いの近さを手がかりとして利用した．まず，表記の上で似ている獲得語と既知語のペアを編集距離を用いて列挙する．しかし，表記上の類似度だけでは,「アワー」と「アワ」のように実際には無関係なペアも候補となる．そこで，次に分布類似度を用いて実際に同じ形態素らしいかを判定した．分布類似度としては，名詞ペアについては共起する用言，動詞，形容詞のペアについては共起する格要素の名詞を用いた．こうした非標準表記はJUMAN によって動的にも認識されるが，自動辞書の語彙が優先される．振る舞いの近さという手がかりを利用している分，自動辞書の語彙の方が信頼できると期待されるからである．

テキストから獲得した未知語の辞書 (自動辞書) は，バージョン 6.0 ではデフォルトではオフとしていたが，7.0 からデフォルトで使用されるようにした．辞書は`$PREFIX/autodic/Auto.dic` にある．使用しないようにするには`$PREFIX/etc/jumanrc` をユーザのホームに`.jumanrc` としてコピーし，`autodic` の行をコメントアウトする．

E.15.4 UTF-8 化

入出力の文字コードは，従来は EUC-JP であったが，UTF-8 に変更した．それに伴い，辞書，jumanrc およびソースコードの文字コードを UTF-8 にした．

これまでホームディレクトリに`.jumanrc` を置いていた場合は，このファイルの文字コードを EUC-JP から UTF-8 に変換する必要がある．また，ユーザ辞書を利用していた場合は，辞書ファイルの文字コードを UTF-8 に変換し，辞書を再コンパイルする必要がある．