

JUMAN/KNP を用いた 形態素・構文・格解析

河原大輔

情報通信研究機構

黒橋禎夫

京都大学

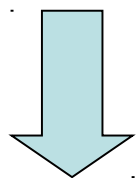
Closed class word の
振舞いは人手整備

文法
基本語彙(3万語)
- 代表表記
- 用言間の関係
- カテゴリ・ドメイン

Open class word の
振舞いは教師無し学習

未知語
語句の分布類似度
格フレーム
格フレーム間対応
未知語・複合語のカテゴリ・ドメイン

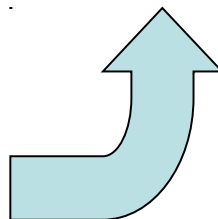
映像, 音声, 文字などの情報をデジタル処理で統合し, しかも双方向で情報交換ができるのがマルチメディア.



JUMAN 99 %

映像 えいぞう 映像 名詞 普通名詞 **
, , , 特殊 読点 **
音声 おんせい 音声 名詞 普通名詞 **
, , , 特殊 読点 **
文字 もじ 文字 名詞 普通名詞 **
など など など 助詞 副助詞 **
の の の 助詞 接続助詞 **
情報 じょうほう 情報 名詞 普通名詞 **
を を を 助詞 格助詞 **
デジタル でじたる デジタル 名詞 普通名詞 **
処理 しょり 処理 名詞 サ変名詞 **
で で で 助詞 格助詞 **
統合 とうごう 統合 名詞 サ変名詞 **
し し する 動詞 * サ変動詞 基本連用形
, , , 特殊 読点 **
しかも しかも しかも 接続詞 ***
...

映像, <p>
音声, <p>
文字などの <p> PARA
情報を
デジタル処理で
統合し, <p>
しかも
双方向で
情報交換が
できるのが <p> PAPA
マルチメディア.



KNP 90 %

目次

1. インストール確認
2. 環境設定
3. JUMAN の仕組み、使い方
4. KNP の仕組み、使い方
5. Perl 超入門
6. JUMAN/KNP と Perl を用いたいろいろな頻度統計の取り方
7. 自動構築した大規模格フレームとそれに基づく格解析

配布パッケージの内容

- C:\juman-knp-20090930
 - install
 - juman-6.0.exe, knp-3.0.exe, Perl 関連
 - src （ Perl スクリプト）
 - cut.pl, grep.pl, phrase.pl, sort.pl, uniq.pl
 - text （サンプルテキスト）
 - 料理: cook_small.txt, cook_large.txt
 - Web : web_small.txt, web_large.txt
 - small: 1,000 文 , large: 20,000 文

1. インストール確認

- 配布パッケージ
 - C:\juman-knp-20090930
- JUMAN
 - C:\Program Files\juman
- KNP
 - C:\Program Files\knp
- Perl (ActivePerl)
- Perl モジュール

2. 環境設定

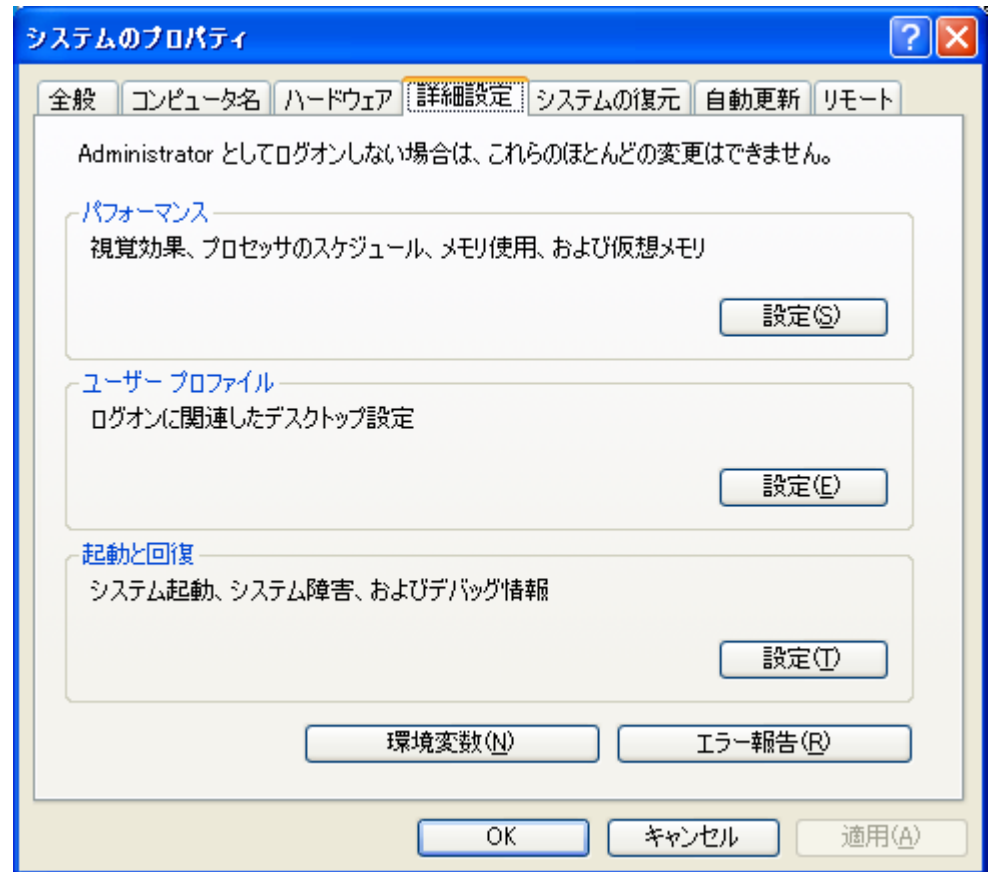
- PATH の設定
 - マイコンピュータを
右クリックしてプロ
パティを選ぶ



(Windows XP の場合)

2. 環境設定

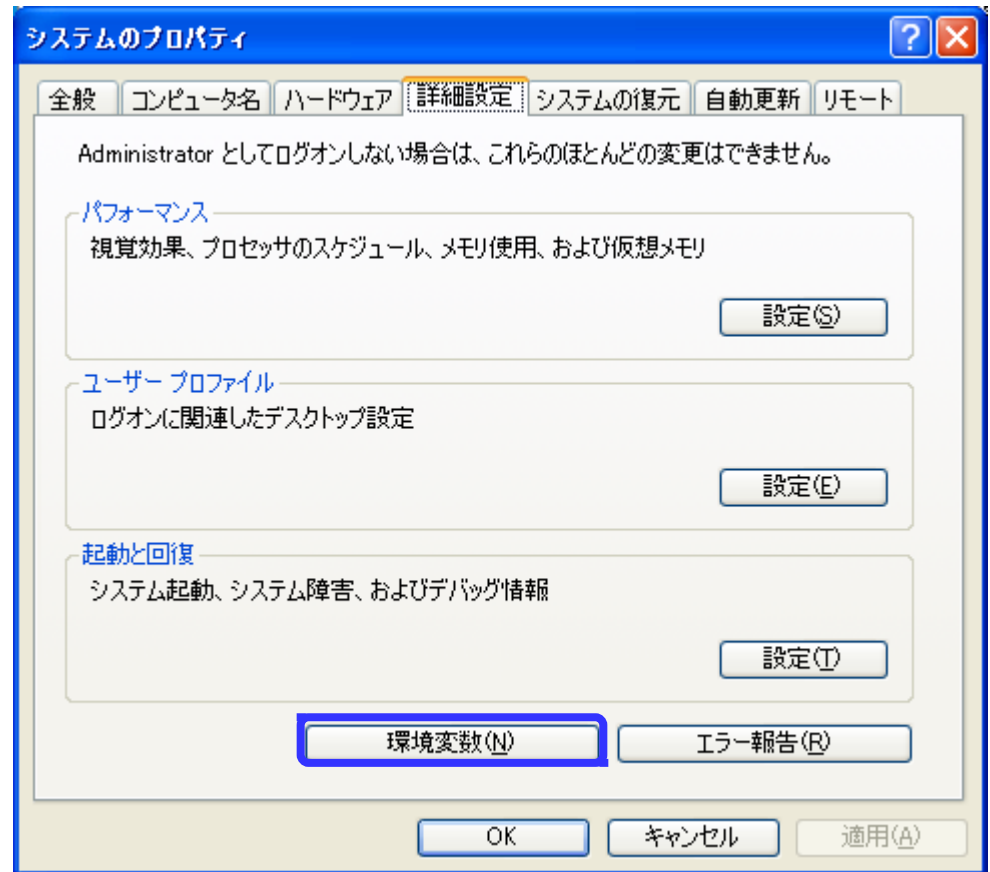
- PATH の設定
 - 詳細設定のタブを選ぶ



(Windows XP の場合)

2. 環境設定

- PATH の設定
– 環境変数を押す



(Windows XP の場合)

2. 環境設定

- PATH の設定
 - コンピュータを右クリックしてプロパティを選ぶ



(Windows Vista の場合)

2. 環境設定

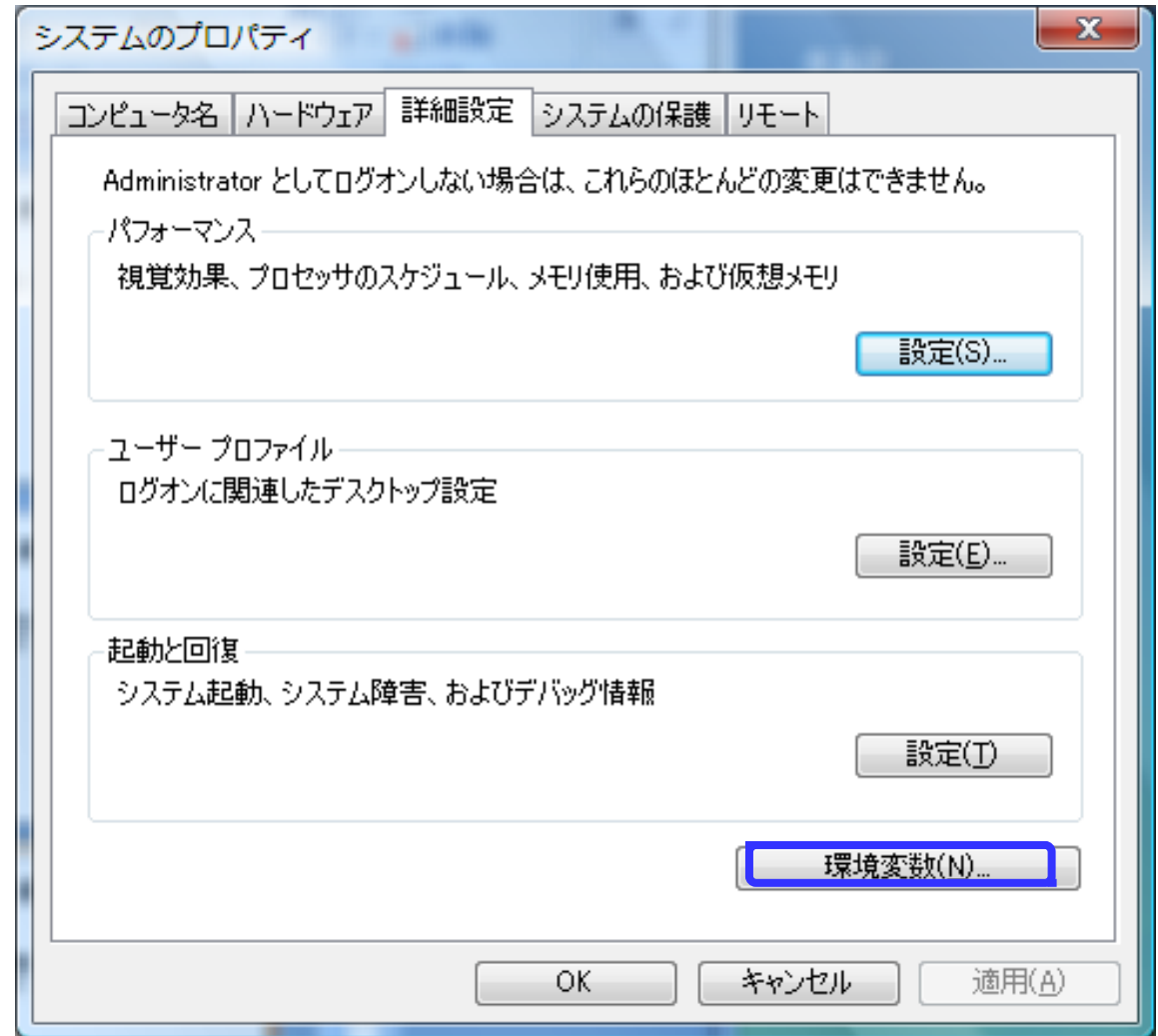
- PATH の設定
 - 「システムの
詳細設定」を
クリック



(Windows Vista の場合)

2. 環境設定

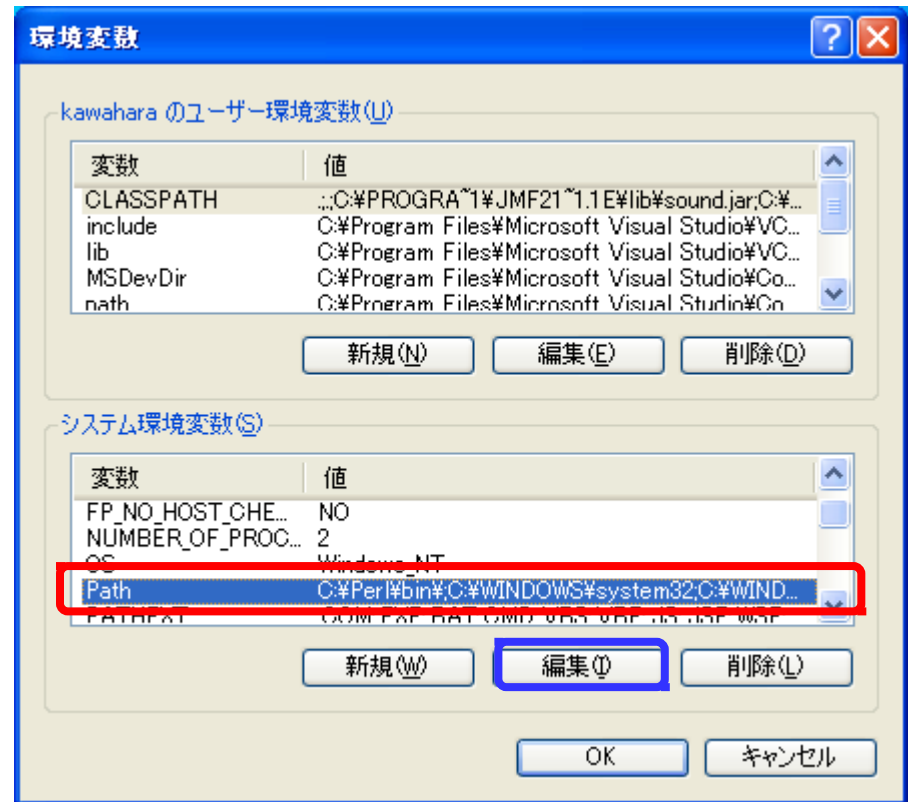
- PATH の設定
 - 環境変数を押す



(Windows Vista の場合)

2. 環境設定

- PATH の設定
 - Path をクリックして「編集」をクリック



2. 環境設定

- PATH の設定
 - Path をクリックして「編集」をクリック
 - 変数値の末尾に下記の文字列を追加する(元の文字列を消さないよう、まず右矢印を押す)



;C:\Program Files\juman;C:\Program Files\knp

目次

1. インストール確認
2. 環境設定
3. JUMAN の仕組み、使い方
4. KNP の仕組み、使い方
5. Perl 超入門
6. JUMAN/KNP と Perl を用いたいろいろな頻度統計の取り方
7. 自動構築した大規模格フレームとそれに基づく格解析

3.1 JUMAN の仕組み

コスト最小法

コスト = $\Sigma \{ (\text{形態素コスト} \times \text{形態素コスト重み})$

(品詞コスト × 見出し語コスト)

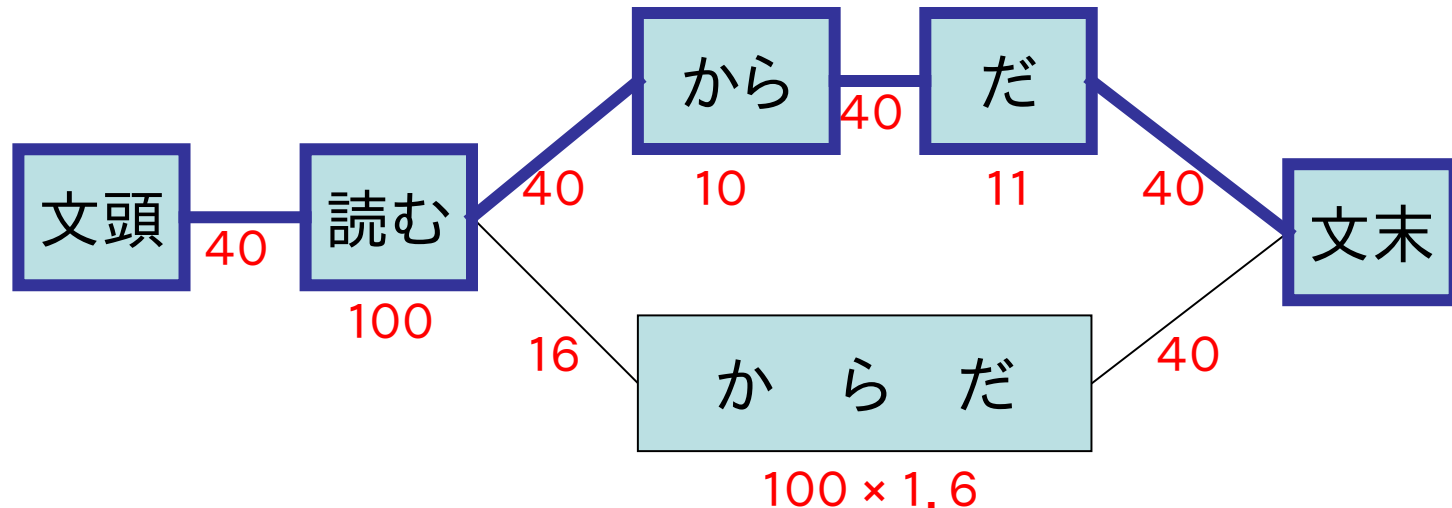
10~100 ~ 1.0~

1

+ (接続コスト × 接続コスト重み) }

~10~

4



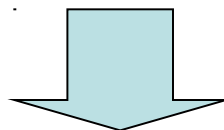
辞書・文法

文法辞書

JUMAN.grammar (品詞分類)
JUMAN.katuyou (活用)
JUMAN.kankei (活用関係)
JUMAN.connect.c (接続規則:
250)

形態素辞書

ContentW.dic など
自立語: 3 万語
付属語: 1500 語
固有名詞: 3 万語



コンパイル

jumandic.tab (接続対応表)
jumandic.mat (接続行列)

jumandic.dat (データベース)
jumandic.pat (インデックス)

ContentW.dic (形態素辞書)

(名詞 (普通名詞 ((読み からくさ)(見出し語 唐草 (から草 1.6)(からくさ 1.6)(意味情報 “代表表記：唐草 / からくさ”)))))
(名詞 (普通名詞 ((読み からくち)(見出し語 辛口 (から口 1.6)(からくち 1.6)(意味情報 “代表表記：辛口 / からくち”)))))
(副詞 ((読み からくも)(見出し語 辛くも からくも)(意味情報 “代表表記：辛くも / からくも”))))
(名詞 (普通名詞 ((読み からくり)(見出し語 からくり)(意味情報 “代表表記：からくり / からくり”)))))
(動詞 ((読み からす)(見出し語 枯らす からす)(活用型 子音動詞サ行)(意味情報 “代表表記：枯らす / からす”))))
(名詞 (普通名詞 ((読み からす)(見出し語 烏 カラス (からす 1.6)(意味情報 “代表表記：烏 / からす”)))))
(名詞 (普通名詞 ((読み からだ)(見出し語 身体 体 (からだ 1.6)(意味情報 “代表表記：身体 / からだ”)))))
(名詞 (普通名詞 ((読み からだつき)(見出し語 体付き 体付 体つき (からだつき 1.6)(意味情報 “代表表記：体付き / からだつき”)))))
(名詞 (普通名詞 ((読み からっかぜ)(見出し語 空っ風 (からっかぜ 1.6)(意味情報 “代表表記：空っ風 / からっかぜ”)))))
(副詞 ((読み からっきし)(見出し語 からっきし)(意味情報 “代表表記：からっきし / からっきし”))))
...

JUMAN.connect.c (連接規則辞書)

...

((BunsetsuEndSentenceEnd
BunsetsuEnd
(助詞 接続助詞 * * の))
((名詞))

4

((VerbBasicForm
IAdjBasicForm
NaAdjAllBasicForm
AuxBasicForm
NaAdjGuessForm
(* * * タ系推量形)
(動詞 * * タ系連用テ形)
(接尾辞 動詞性接尾辞 * タ系連用テ形))
((助詞 接続助詞 * * から)))

...

基本語彙の選択(3万語)

- 事典的な語は排除
例) 倭寇 天動説 秋の七草 父の日
- 古語、ほとんど使われない語・読みは排除
例) 内生活 手ずから 生く(おいゆく) 夜間(よま)
- 3文字以上の複合語は基本的に排除
例) 印刷機 映写機 運動場 競技場 研究費
 - ただし以下は採用
 - 構成語が一般的でないもの: 感受性 一本化
 - 意味が構成的でないもの: 耳学問 銀世界
 - 切り方が?なもの: 工学部 全速力 海産物
 - 音訓の原則で読みが誤るもの(他との整合性も考慮し): オレンジ色
- 2文字の語は原則採用だが、
 - 構成性が明確なものは排除: 学内 市内
 - 以下は採用だが関係解析で問題: (経理)部長 (警察)署員

JUMAN 辞書に記述されている情報 (1/3)

- 代表表記
- 1 文字漢字について、音・訓の区別
例) 字 / じ → 音, 字 / あざ → 訓
- 可能動詞であることと、もとの動詞
例) 書ける → 可能動詞: 書く
- 尊敬動詞・謙譲動詞であることと、もとの動詞
例) おっしゃる → 尊敬動詞: 言う
- 動詞が付属動詞として振舞うかどうか
例) 合う
- カテゴリ、ドメイン情報
例) カテゴリ: 先生, 学生, 父 → 人
ドメイン: テニス, ラケット, サーブ → スポーツ

JUMAN 辞書に記述されている情報 (2/3)

- 自動詞・他動詞の対応関係
 - 例) 壊れる → 自他動詞:他:壊す
 - 壊す → 自他動詞:自:壊れる
- 授受動詞の対応関係
 - 例) 貸す → 授受動詞:受:借りる
 - 借りる → 授受動詞:授:貸す
- 反義
 - 例) 増える → 反義:動詞:減る
 - 大きい → 反義:動詞:小さい
- 種々の派生
 - 例) 大人びる → 名詞派生:大人
 - 高める → 形容詞派生:高い

JUMAN 辞書に記述されている情報 (3/3)

- 固有名詞に付与されている種々の情報

- 人名

- 例) 山田 → 人名 : 日本 : 姓 : 7:0.00607

- 太郎 → 人名 : 日本 : 名 : 45:0.00106

- 地名

- 例) 京都 → 地名 : 日本 : 府 , 地名 : 日本 : 京都府 : 市

- 東北 → 地名 : 日本 : 地方

- 英国 → 地名 : 国

- イギリス → 地名 : 国 : 別称 : 英国

- カリフォルニア → 地名 : 国 : 米国 : 州

- 組織名

- 例) パナソニック

- 民主党

代表表記

子ども こども 子ども 名詞 普通名詞 **

は は は 助詞 副助詞 **

リンゴ りんご リンゴ 名詞 普通名詞 **

が が が 助詞 格助詞 **

すきだ すきだ すきだ 形容詞 * ナ形容詞 基本形

EOS

かぜ かぜ かぜ 名詞 普通名詞 **

で で で 助詞 格助詞 **

おくれた おくれた おくれる 動詞 * 母音動詞 夕形

EOS

代表表記

子ども こども 子ども 名詞 普通名詞 ** “代表表記：子供 / こども”

は は は 助詞 副助詞 **

リンゴ りんご リンゴ 名詞 普通名詞 ** “代表表記：林檎 / りんご”

が が が 助詞 格助詞 **

すきだ すきだ すきだ 形容詞 * ナ形容詞 基本形 “代表表記：好きだ / すきだ”

EOS

かぜ かぜ かぜ 名詞 普通名詞 ** “代表表記：風 / かぜ”

@ かぜ かぜ かぜ 名詞 普通名詞 ** “代表表記：風邪 / かぜ”

で で で 助詞 格助詞 **

おくれた おくれた おくれる 動詞 * 母音動詞 夕形 “代表表記：送れる / おくれる 可能動詞：送る”

@ おくれた おくれた おくれる 動詞 * 母音動詞 夕形 “代表表記：遅れる / おくれる”

EOS

代表表記(同じ読み)

- 漢字と平仮名、送り仮名

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

- 漢字別表記

例) 狩人 / 獵人 色取る / 彩る 綺麗だ / 奇麗だ

- カタカナ表記

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 溝 / みぞ / ミゾ 眼鏡 / めがね / メガネ

代表表記(異なる読み)

- 音便関係

例) 私 / わたし / わたくし / あたし 皆 / みな / みんな
旅客機 / りょかくき / りょかっき ふわり / ふんわり
とびきり / 飛び切り / とびっきり / 飛びっ切り

- カタカナ表記のバリエーション

例) ソフトウェア / ソフトウェア コンピューター / コンピュータ

カテゴリ情報（22 種類）

カテゴリ名	例	カテゴリ名	例
人	学生, 先生, …	場所 - 施設	ビル, 公園, …
組織・団体	政府, 企業, …	場所 - 施設部位	天井, 床, …
動物	犬, 猫, …	場所 - 自然	山, 海, …
植物	桜, バラ, …	場所 - 機能	上, 下, …
動物 - 部位	手, 足, …	場所 - その他	都市, 村, …
植物 - 部位	葉, 枝, …	抽象物	思考, 理由, …
人工物 - 食べ物	パン, コーヒー, …	形・模様	円, 球, …
人工物 - 衣類	ズボン, セーター, …	色	赤, 青, …
人工物 - 乗り物	自動車, 飛行機, …	数量	複数, メートル, …
人工物 - 金銭	給料, 借金, …	時間	今日, 朝, …
人工物 - その他	鉛筆, 消しゴム, …		

ドメイン情報（12 種類）

ドメイン名	例	ドメイン名	例
文化・芸術	映画 音楽 御 興 ...	交通	駅 道路 アクセル
レクリエーション	観光 花火 カジノ . ..	教育・学習	先生 算数 開校 ...
スポーツ	選手 野球 角界 .. .	科学・技術	研究 理論 ウラン .. .
健康・医学	手術 診断 胃液 .. .	ビジネス	輸入 市場 売上 ...
家庭・暮らし	育児 家具 帰省 .. .	メディア	放送 記者 載る ...
料理・食事	箸 昼食 和え る ...	政治	司法 税 拳党 ...
ドメイン無し	青 感情 上が る ...		

連濁、反復形オノマトペ、非標準表記 の自動認識

上 海 | ガ ニ | を | ば く ば く | 食 べ る

かわいい

蟹 / かに

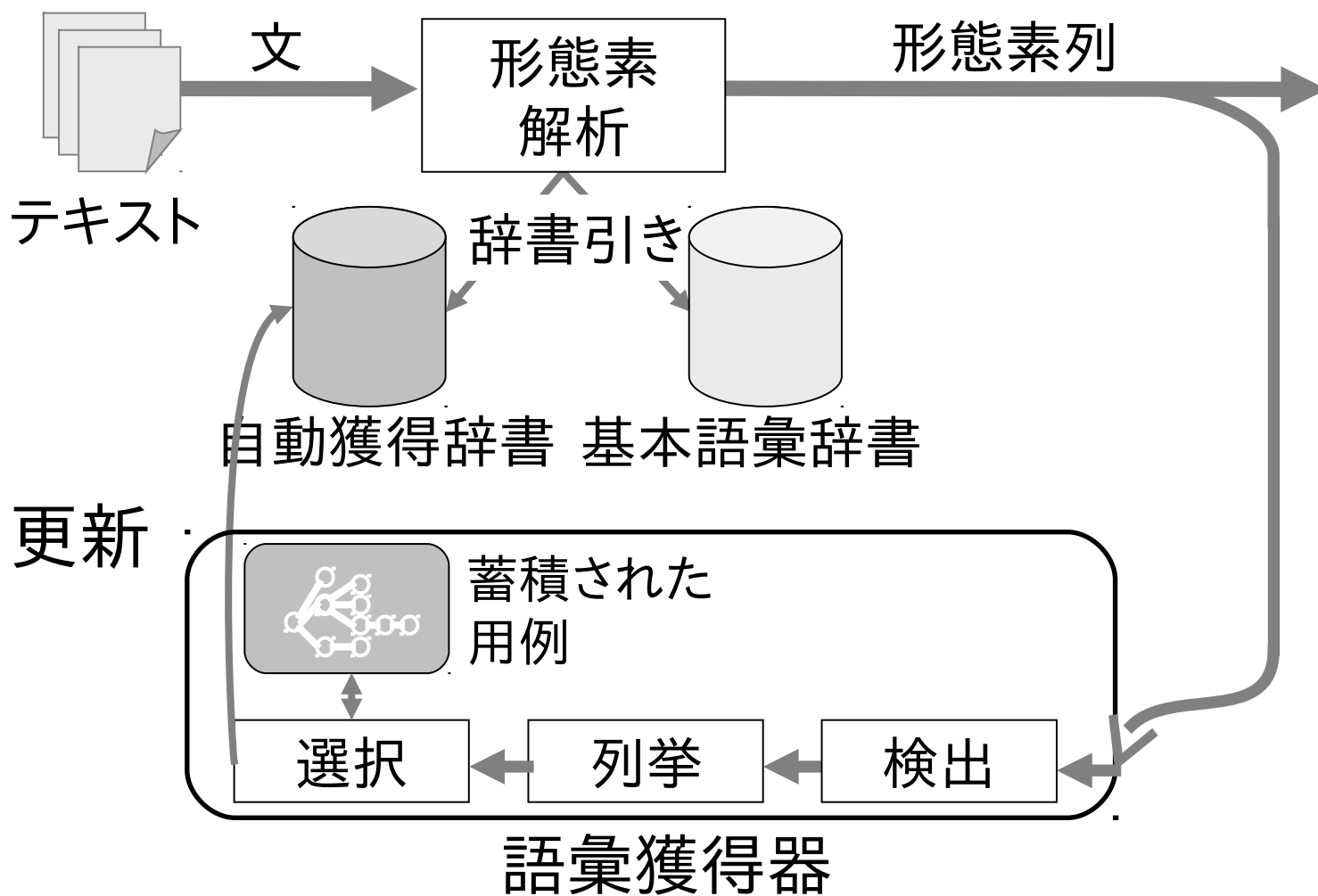
副詞

かわいい

(連濁) (反復形オノマトペ) (小文字を用いた非標準表記)

- 連濁:
 - 辞書を検索する際、濁音から始まる場合は清音化したものも検索する
 - いくつかの制約を課す
 - 和語のみ濁音化する
 - 濁音を含む形態素の濁音化は考えない(ライマンの法則)
 - おお + とかげ ≠ おおどかげ
- 反復形オノマトペ:
 - 形態素の候補を検索する際、2 ～ 4 文字の繰り返し表現があれば候補に加える
- 小文字を用いた非標準表記:
 - 入力文に「あ、い、う、え、お、わ、か」があった場合、それらを大文字化したものも検索する

オンライン未知語獲得



語彙獲得器の仕組み

…何となくググってみた。…

(ググ - る, 動詞 - ラ行)
(ググ - る, 動詞 - ラ行, タ連用テ
形)

[BOS] ググらずに答えるのが…

(ググ - る, 動詞 - ラ行)
(ググ - る, 動詞 - ラ行, 未然形)

…いるだけで、ググるための…

(ググ - る, 動詞 - ラ行)
(ググ - る, 動詞 - ラ行, 基本形)

(ググ - る, 動詞 - ラ
行)

語彙獲得例:

ようつべ: 名詞

倅田來未: 名詞 - 人
名

ムカツ - く: 動詞 -
カ行

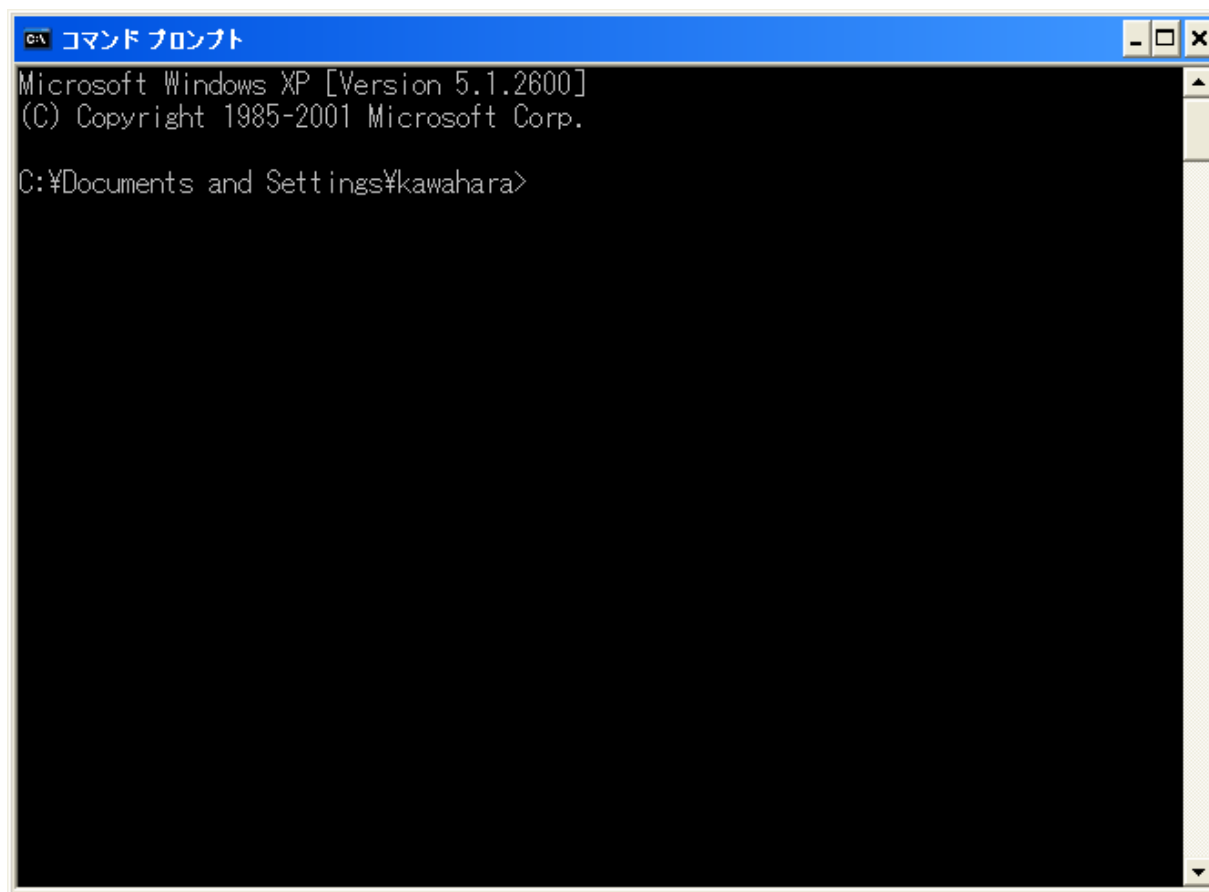
うざ - い: 形容詞

3.2 JUMAN を使ってみよう

- コマンド プロンプトを開く
 - スタート⇒すべてのプログラム⇒アクセサリ⇒
コマンド プロンプト

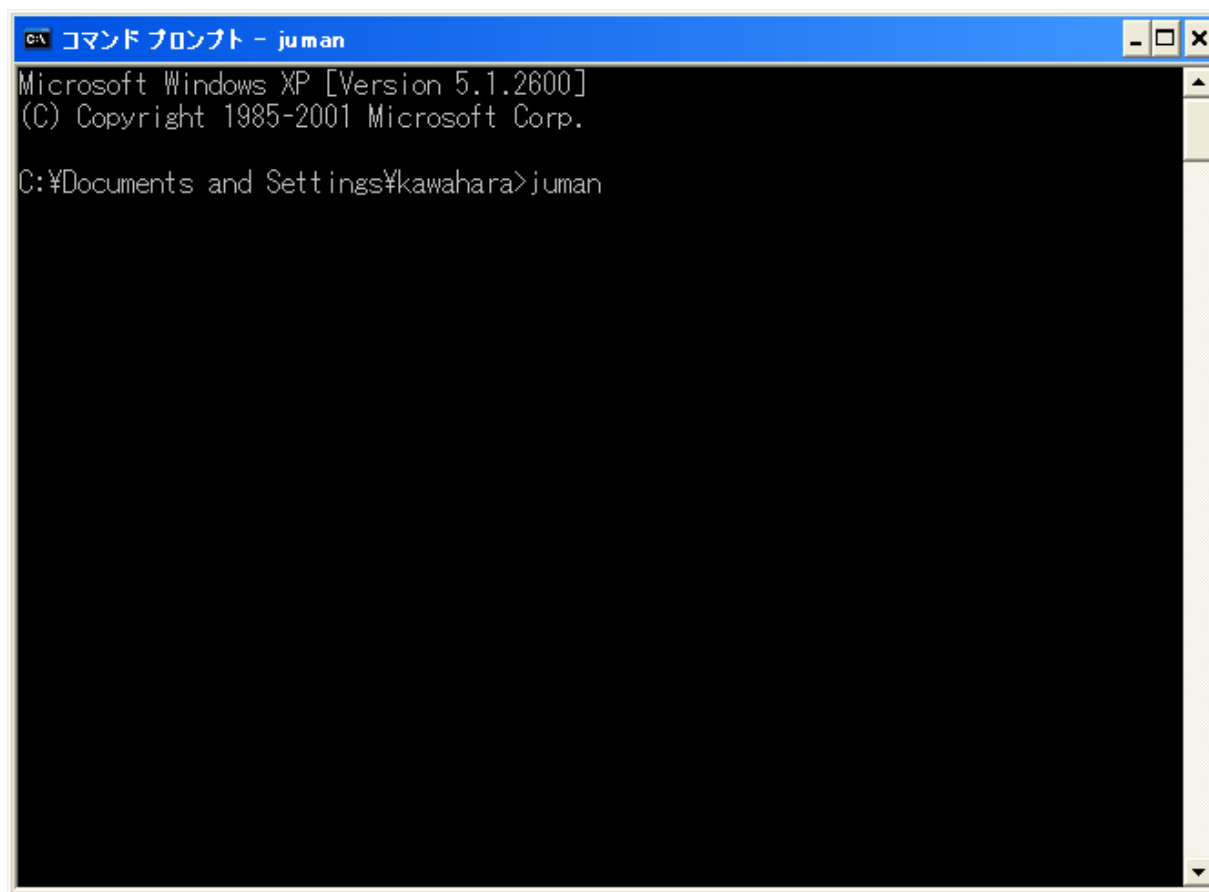
JUMAN を使ってみよう

- コマンド プロンプトを開く



JUMAN を使ってみよう

- juman と打つ

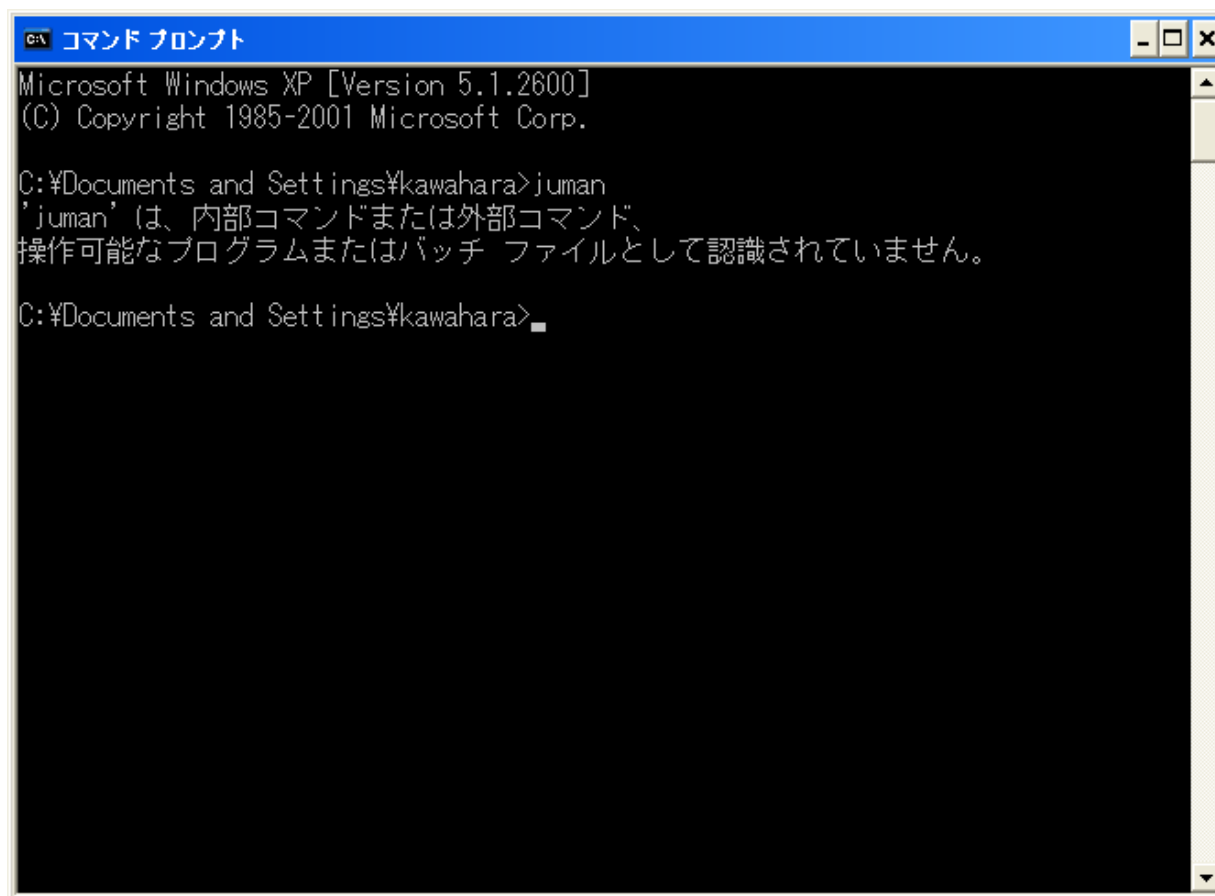


```
C:\ コマンド プロンプト - juman
Microsoft Windows XP [Version 5.1.2600]
(C) Copyright 1985-2001 Microsoft Corp.

C:¥Documents and Settings¥kawahara>juman
```

JUMAN を使ってみよう

- 環境設定ができていない場合



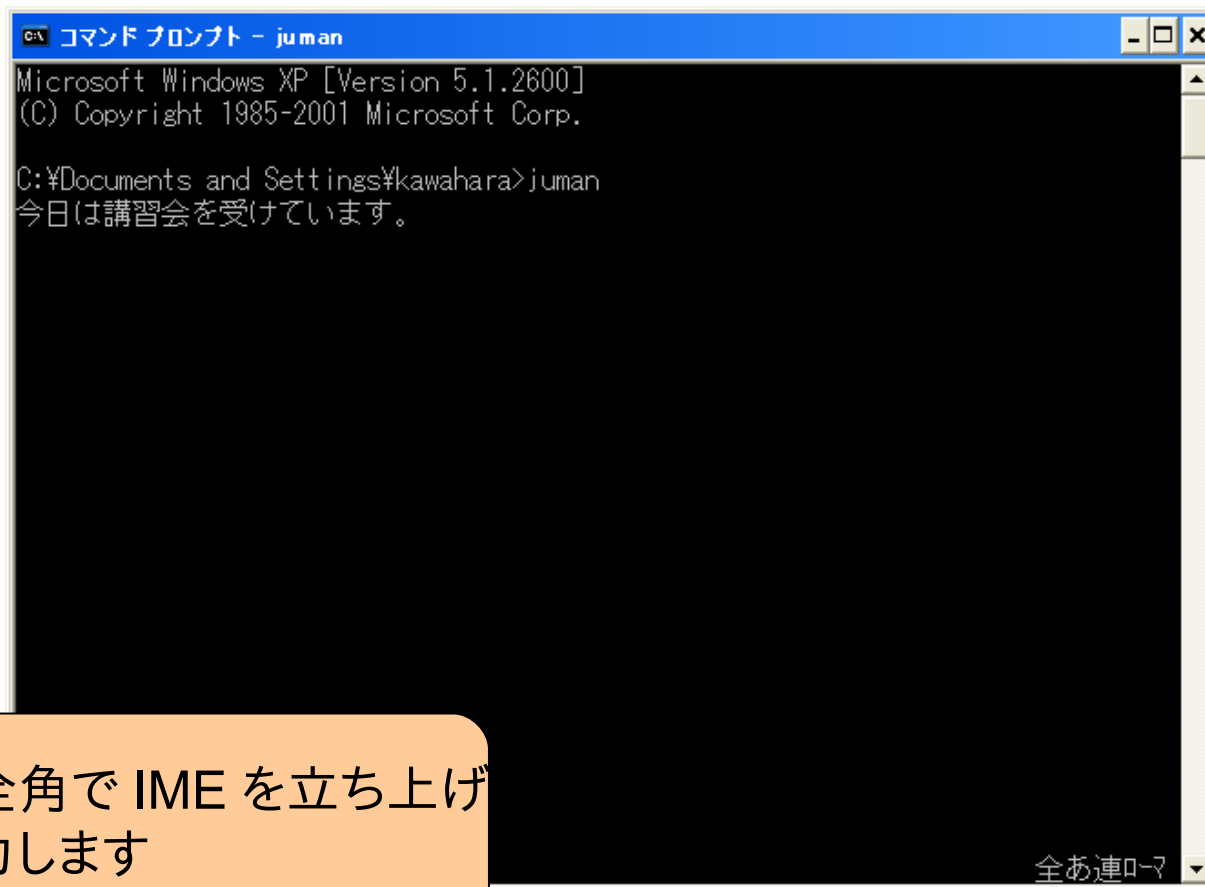
```
コマンド プロンプト
Microsoft Windows XP [Version 5.1.2600]
(C) Copyright 1985-2001 Microsoft Corp.

C:\Documents and Settings\kawahara>juman
'juman' は、内部コマンドまたは外部コマンド、
操作可能なプログラムまたはバッチ ファイルとして認識されていません。

C:\Documents and Settings\kawahara>
```

JUMAN を使ってみよう

- 「今日は講習会を受けています。」と入力

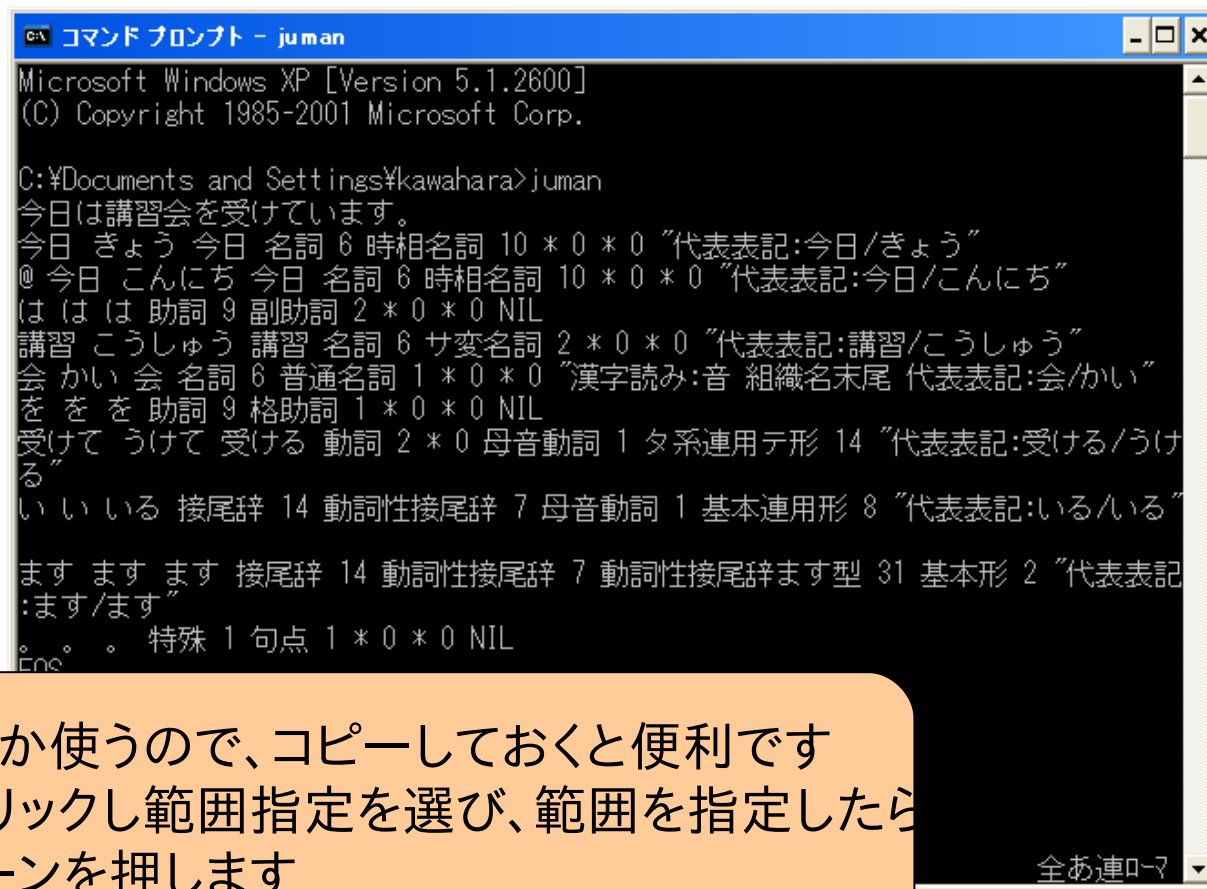


Tips

Alt+ 半角 / 全角で IME を立ち上げ
日本語を入力します

JUMAN を使ってみよう

- 「今日は講習会を受けています。」と入力



```
コマンドプロンプト - juman
Microsoft Windows XP [Version 5.1.2600]
(C) Copyright 1985-2001 Microsoft Corp.

C:\Documents and Settings\kawahara>juman
今日は講習会を受けています。
今日 きょう 今日 名詞 6 時相名詞 10 * 0 * 0 "代表表記:今日/きょう"
@ 今日 こんにちは 今日 名詞 6 時相名詞 10 * 0 * 0 "代表表記:今日/こんにちは"
は は は 助詞 9 副助詞 2 * 0 * 0 NIL
講習 こうしゅう 講習 名詞 6 サ変名詞 2 * 0 * 0 "代表表記:講習/こうしゅう"
会 かい 会 名詞 6 普通名詞 1 * 0 * 0 "漢字読み:音 組織名末尾 代表表記:会/かい"
を を を 助詞 9 格助詞 1 * 0 * 0 NIL
受けて うけて 受ける 動詞 2 * 0 母音動詞 1 タ系連用テ形 14 "代表表記:受ける/うける"
いい いる 接尾辞 14 動詞性接尾辞 7 母音動詞 1 基本連用形 8 "代表表記:いる/いる"
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"
End。 。 特殊 1 句点 1 * 0 * 0 NIL
```

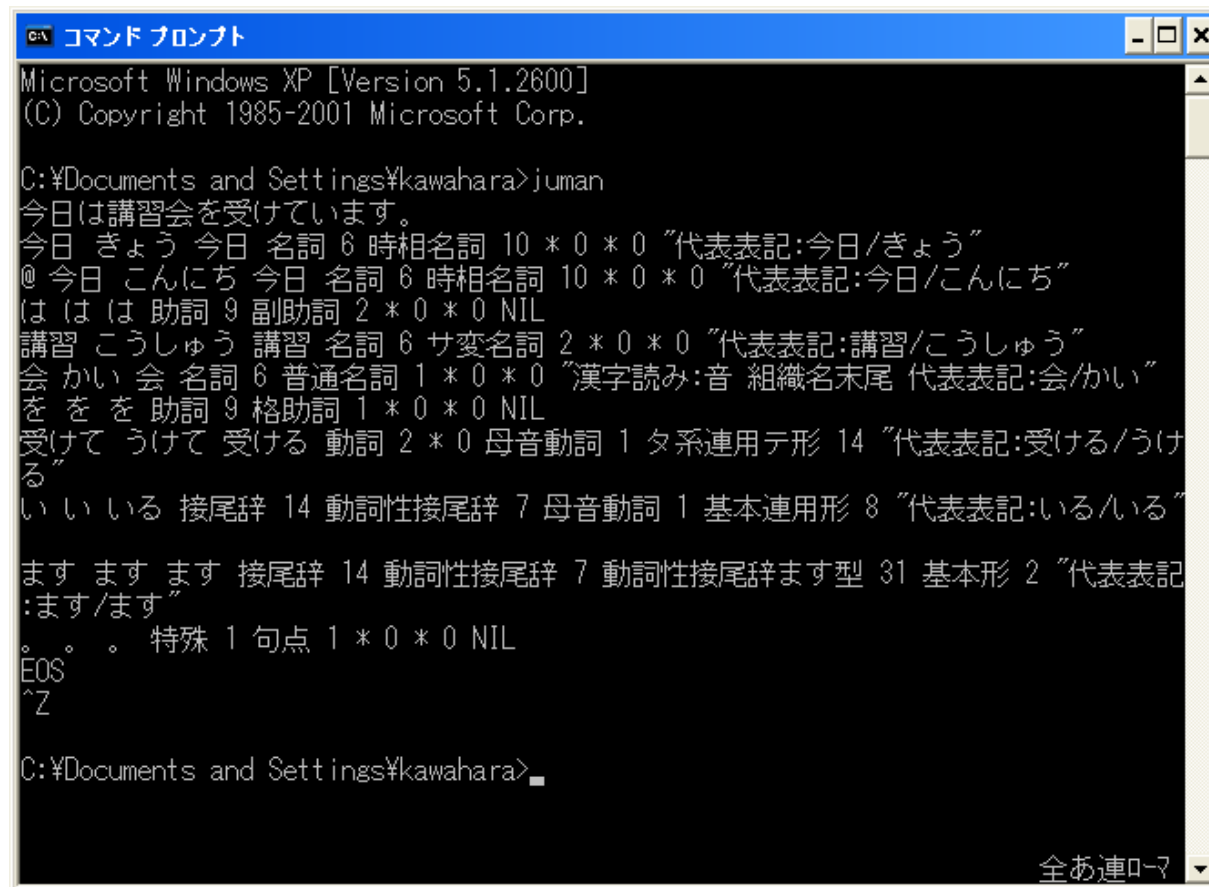
Tips

この文は何回か使うので、コピーしておくと便利です

- ⇒ 右クリックし範囲指定を選び、範囲を指定したら
リターンを押します

JUMAN を使ってみよう

- コントロール Z リターン で終了



```
コマンド プロンプト
Microsoft Windows XP [Version 5.1.2600]
(C) Copyright 1985-2001 Microsoft Corp.

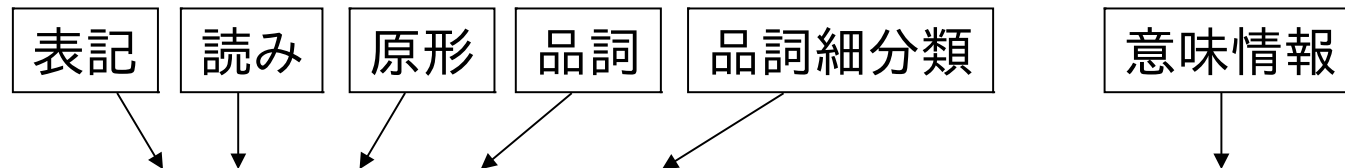
C:\Documents and Settings\kawahara>juman
今日は講習会を受けています。
今日 きょう 今日 名詞 6 時相名詞 10 * 0 * 0 "代表表記:今日/きょう"
@ 今日 こんにちは 今日 名詞 6 時相名詞 10 * 0 * 0 "代表表記:今日/こんにちは"
は は は 助詞 9 副助詞 2 * 0 * 0 NIL
講習 こうしゅう 講習 名詞 6 サ変名詞 2 * 0 * 0 "代表表記:講習/こうしゅう"
会 かい 会 名詞 6 普通名詞 1 * 0 * 0 "漢字読み:音 組織名末尾 代表表記:会/かい"
を を を 助詞 9 格助詞 1 * 0 * 0 NIL
受けて うけて 受ける 動詞 2 * 0 母音動詞 1 タ系連用テ形 14 "代表表記:受ける/うける"
いい いる 接尾辞 14 動詞性接尾辞 7 母音動詞 1 基本連用形 8 "代表表記:いる/いる"

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"
。 。 。 特殊 1 句点 1 * 0 * 0 NIL
EOS
^Z

C:\Documents and Settings\kawahara>
```

全あ連ローマ

juman フォーマット



今日 きょう 今日 名詞 6 時相名詞 10 * 0 * 0 "代表表記: 今日 / きょう カテゴリ: 時間"
@ 今日 こんにち 今日 名詞 6 時相名詞 10 * 0 * 0 "代表表記: 今日 / こんにち カテゴリ"
は は は 助詞 9 副助詞 2 * 0 * 0 NIL
講習 こうしゅう 講習 名詞 6 サ変名詞 2 * 0 * 0 "代表表記: 講習 / こうしゅう ドメイン: 教
会 かい 会 名詞 6 普通名詞 1 * 0 * 0 "漢字読み: 音 組織名末尾 代表表記: 会 / かい"
を を を 助詞 9 格助詞 1 * 0 * 0 NIL
受けて うけて 受ける 動詞 2 * 0 母音動詞 1 夕系連用テ形 14 "代表表記: 受ける / うけ
いい いる 接尾辞 14 動詞性接尾辞 7 母音動詞 1 基本連用形 8 "代表表記: いる / い
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
ます"
。。。特殊 1 句点 1 * 0 * 0 NIL
EOS

活用型

活用形

曖昧性のある形態素を表す

JUMAN のカスタマイズ

- 辞書エントリの追加

- 例えば、「ジンギスカン」を追加する

- まず、今の解析がどうなるかを確認する
- C:\Program Files\juman\dic 以下に usr.dic というファイルを作り、以下の内容を記述する

(名詞 (普通名詞 ((読み じんぎすかん)
(見出し語 ジンギスカン じんぎすかん 成吉思汗))))

- C:\Program Files\juman\dic にある makedic.bat を実行する
- 解析してみる

自動獲得辞書を利用する方法

1. C:\Program Files\juman\autodic\Auto.dic を
C:\Program Files\juman\dic にコピーする
2. C:\Program Files\juman\dic にある
makedic.bat を実行する
3. 例えば、「2ちゃんねる」を juman で解析する

ブログのドメイン推定

台湾の新米主婦: 白TSUBAKI

http://sen0633.269g.net/article/5687598.html

Google

kirakira*clover

台湾の新米主婦

<<happy wedding> Main | コスモス>>

2007年10月15日

PR

プレゼント&キャンペーン

500円分 ほか
ポイントゲット!!

プロフィール

Author: せん
台湾人の老公と日本生活3年目。
2009年1月にヒメが産まれ子育て奮闘

白TSUBAKI

最近ハマっています〜

TSUBAKIシリーズ

最近日本では白TSUBAKIが発売されてから、私の周りでは、みんなこれを使っています。

台湾ではもう、TSUBAKIは発売してるのかな？
去年私がいた時にはなかったけど、
もう売ってるのかな〜

新しい物大好き♥な私は、発売されると買っちゃうのでヤバイです※
化粧品とかも、使い切る前に買っちゃうので、
結構たまってます

買い物して帰ると、老公に、
「えー？また買い物？」って呆れられています・・・

完了

テキスト抽出→形態素解析 →ドメイン推定

台湾の新米主婦

2007年10月15日

白TSUBAKI

最近ハマっています～

TSUBAKIシリーズ

最近日本では白TSUBAKIが発売されてから、私の周りでは、みんなこれを使ってます。

台湾ではもう、TSUBAKIは発売してるのかな？

去年私がいた時にはなかったけど、もう売ってるのかな～

新しい物大好きな私は、発売されると買っちゃうのでヤバイです **化粧品**とかも、使い切る前に買っちゃうので、結構たまってます

買い物して帰ると、老公に、「えー？また**買い物**？」って呆れられています……

老公の**買い物**にはうるさい私ですが、自分の物**買う**ときは、別なんです――

ドメイン推定結果：

22 家庭・暮らし

10 料理・食事

10 ビジネス

目次

1. インストール確認
2. 環境設定
3. JUMAN の仕組み、使い方
4. KNP の仕組み、使い方
5. Perl 超入門
6. JUMAN/KNP と Perl を用いたいろいろな頻度統計の取り方
7. 自動構築した大規模格フレームとそれに基づく格解析

4.1 KNP の仕組み

処理の流れ

0. JUMAN の出力を入力に
1. 同形異義語の処理 (`mrph_homo.rule:60`)
→ 一意の形態素列に変換
2. 形態素への feature 付与 (`mrph_basic.rule:300`)
→ 文節列に変換
3. 文節への feature 付与 (`bnst_*.rule:200, 650`)
4. 並列構造解析
5. 係り受け可能性チェック (`kakari_uke.rule:40`)
6. 構文・格解析

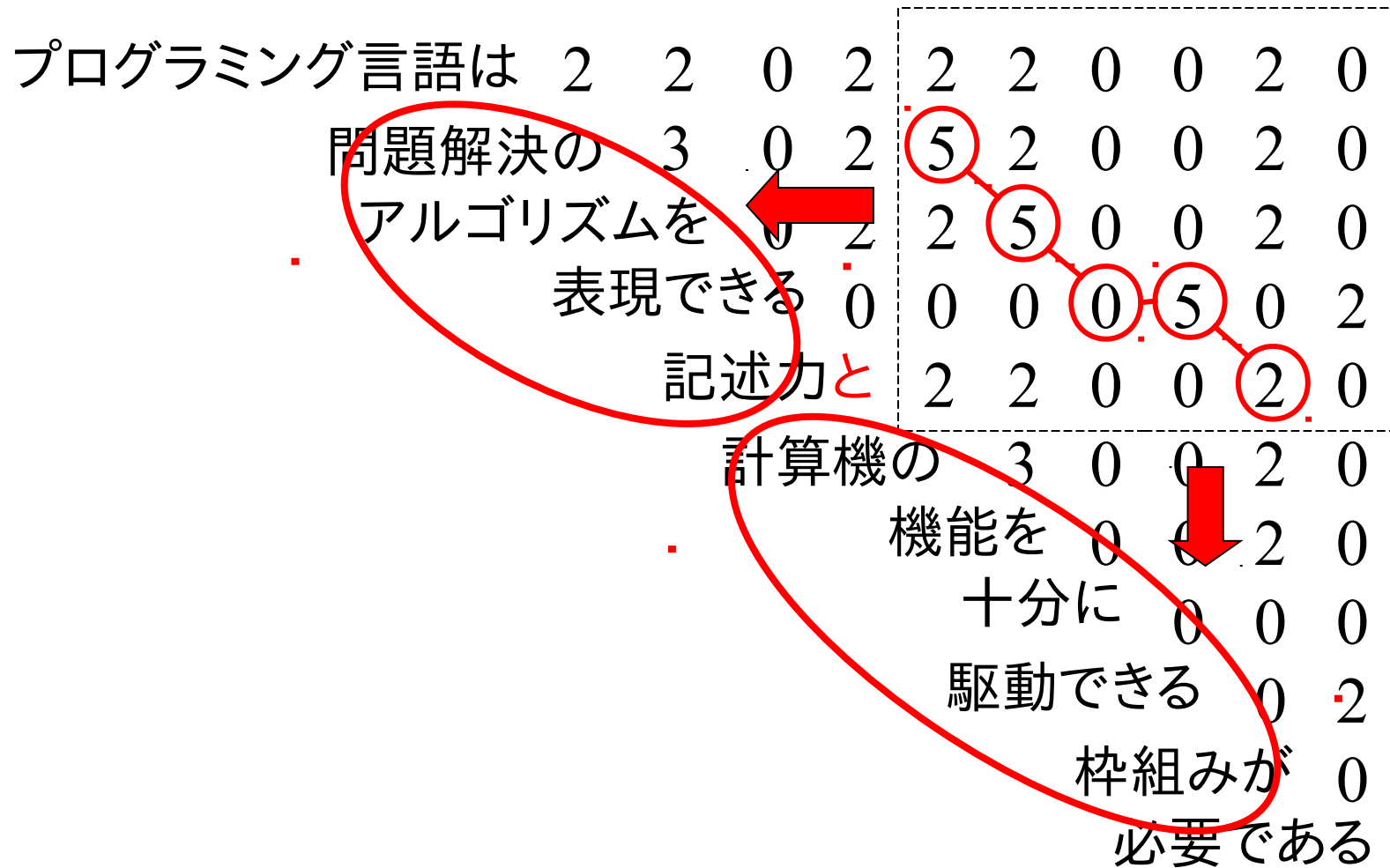
ルールの例 (bnst_basic.rule)

「・・・ 30年も 前から・・・」などを解析するためのルール

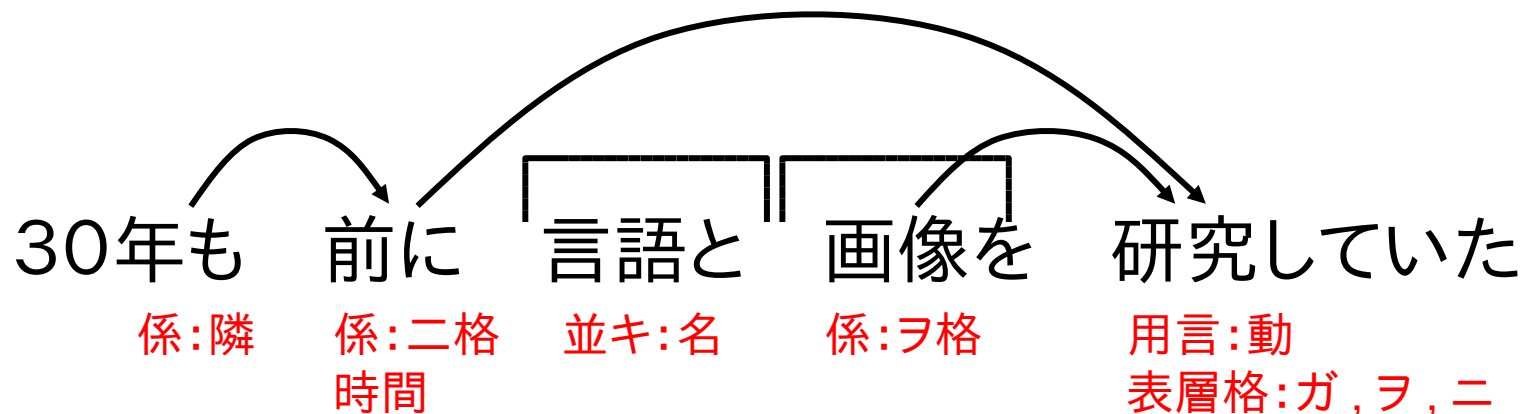
(
(?*) ← 前の文節列

(< (?* [助詞 * * * も]) ((時間)) >) ← 自分自身
形態素列:「～も」 <時間> feature を持つ
(< ([名詞 * * * (昔 前 先)] ?*) > ?*) ← 後ろの文節列
形態素列:「昔|前|先～」
係:隣 ← 与える feature
)

並列構造の解析



係り受け可能性チェック



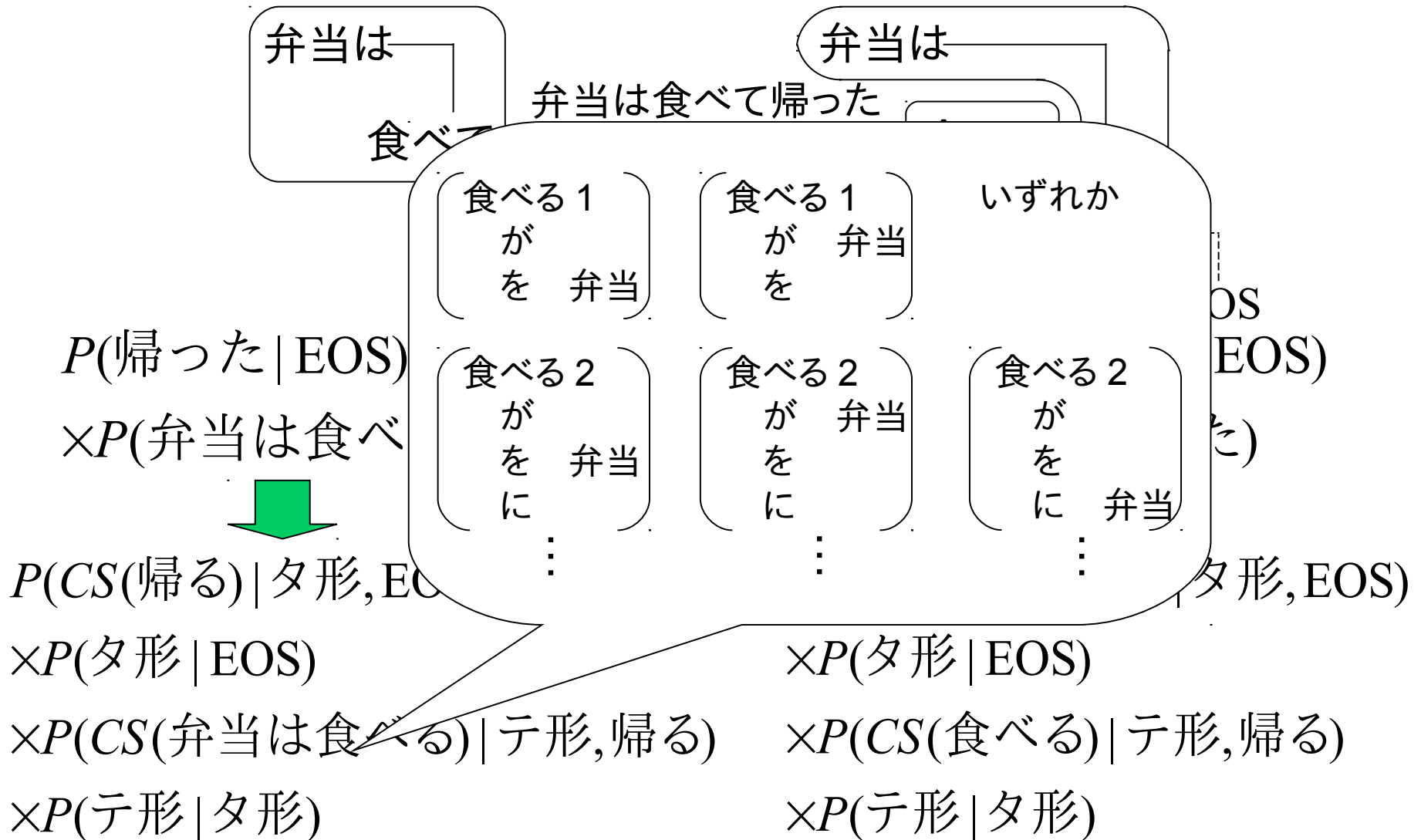
kakari_uke.rule:

係:ヲ格	→	用言	表層格:ヲ
係:二格	→	用言	表層格:ニ
係:二格 時間	→	用言	
係:隣	→	*	
...			

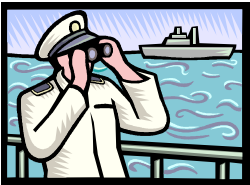
多様性・個別性に対応

- 句(文節)のバリエーション
 - 「書くといえども」「書くとはいえ」
 - 「書き(次第)」「書いた(途端)」「書く(余り)」
 - 「学生ではなく」
- 係り受けの種々のパターン
 - 「彼が学生かどうかは～」
 - 「これを基に～」
 - 「48キロ級で61連勝中の～」
 - 「首より下を～」

構文・格解析の統合的確率モデル



言語の理解 ⇔ 知識



クロールで 泳いでいる女の子を見た
望遠鏡で 泳いでいる女の子を見た



格フレーム

{人, 子, ...} が
{クロール, 平泳ぎ, ...}
で
{海, 大海, ...} を泳ぐ

{人, 者, ...} が
{双眼鏡, 望遠鏡, ...} で
{姿, 人, ...} を見る

格フレーム構築

人手では不可能

- 「数万語の自立語 × 多義性」を幅広くカバーできない
- 「がをに」の単純なパターンばかりではない
e.g. ～によって，～について，二重主語構文
- 新語・専門用語
e.g. ググる，サチる

格フレームの自動構築

- 超大規模コーパス (**Web16 億文**) の利用
- 構文解析→述語項関係をクラスタリング
 - 構文的曖昧性
 - 意味的曖昧性
 - 複雑な関係
 - 二重主語構文
 - 外の関係
- 並列計算環境の利用
 - **300CPU グリッド**
 - グリッドシェル GXP
 - 構文解析 3 日間
 - クラスタリング 10 日間



Web からの日本語文抽出

1. クローラーによるページ収集
 2. エンコーディング情報による日本語ページ候補抽出
 - Charest 情報、perl Encode::guess_encoding() 関数
 3. 言語情報による日本語ページ判定(約 1 億ページ)
 - が, を, に, は, の, で を 0.5% 以上含む
 4. ページの文リストへの分割(HTML タグと句点)
 5. 日本語文の抽出
 - ひらがな、カタカナ、漢字を 60% 以上含む文
 6. 重複文の削除
- 約 16 億日本語文 (妥当な日本語: 995 文 / 1,000 文)

Web からの日本語文抽出

しょうがないので駅のレストランで食事をしようとした所、1日数本しかない山田線の存在に思い当たる。

もれなくプレゼント！

でも僕はTシャツの上に長袖のシャツ。

今回は某アイドルの高橋一也も参加したので客が若い。

団体Aが「まちづくり」をテーマにインターネット上で公開講座を開催しようとしている。

htaccessを置いたとたんそのディレクトリ以下で、

去年の没後400年祭を機に復元した井戸を紹介する木下さん

恋は、真剣勝負。

ほめ言葉が多くって嬉しいですね。

開校式並びに入学式を挙行、初代校長佐治勝弥、職員10名沖館小学校校舎一部を併用す。

佐治勝弥校長青一中学校長に任命される。

いまだに言うでしょう。

「買いパラ」を見た伝えれば、お買い上げ合計金額より5%引きいたします。

政治も危機的状況ですし、物資も不足しています。

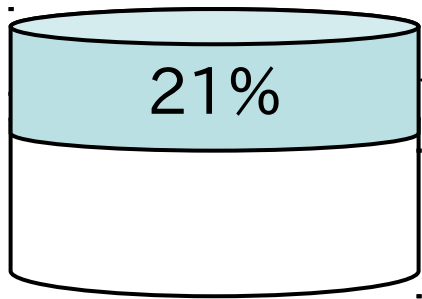
そういう長期的な存在理由とか、長期的なビジョンとか、何故ここが国のお金で、我々の税金でやらなければいけないのか、その辺を評価する上で何かお考えになられていますでしょうか。

...

構文的曖昧性

⇒ 構文解析結果から確実な関係だけを集める

- 買いたい 本を たくさん 見つけたので、東京に 送った。
- 被害者を 早く 救い出すべきだ。
- 火の 回りが 早く 救いだせなかった。
- その 議員は 法案を 提出した。
- 議員が 提出している 法案は...



精度 98.3 %で格関係を抽出

意味的曖昧性

⇒ 直前格に注目してクラスタリング

従業員 が 車 に 荷物 を 積む

作業員 が 荷物 を 積む

飛行機 に 荷物 を 積む

彼 が 車 に 物資 を 積む

トラック に 物資 を 積む

従業員 が 経験 を 積む

選手 が 経験 を 積む

意味的曖昧性

⇒ 直前格に注目してクラスタリング

従業員	が	車	に	荷物	を 積む
作業員	が			荷物	を 積む
		飛行機	に	荷物	を 積む
彼	が	車	に	物資	を 積む
		トラック	に	物資	を 積む
従業員	が			経験	を 積む
選手	が			経験	を 積む

意味的曖昧性

⇒ 直前格に注目してクラスタリング

従業員	が	車	に	荷物	を積む
作業員	が			荷物	を積む
		飛行機	に	荷物	を積む
彼	が	車	に	物資	を積む
		トラック	に	物資	を積む
従業員	が			経験	を積む
選手	が			経験	を積む

複雑な構文(二重主語構文・外の関係) ⇒ 漸進的な学習

- この法案はA議員が提出した。

{議員, 委員..}が {法律, 案..}を 提出する

- その車はエンジンがよい。 → 二重主語構文

{エンジン}が よい

- 法案を提出した議員...

{議員, 委員..}が {法律, 案..}を 提出する

- 法案を提出する見通し... → 外の関係

{議員, 委員..}が {法律, 案..}を 提出する

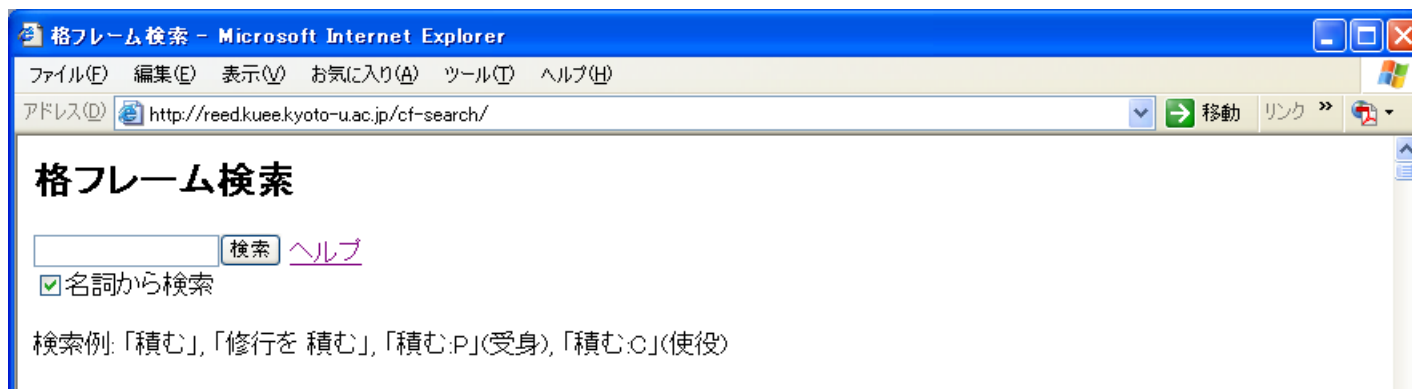
構築した格フレームの例

用言	格	用例
焼く(1)	ガ格	私 :18, 人 :15, 職人 :10, ...
	ヲ格	パン :2484, 肉 :1521, ケーキ :1283, ...
	デ格	オーブン :1630, フライパン :1311, ...
焼く(2)	ガ格	先生 :3, 政府 :3, 人 :3, ...
	ヲ格	手 :2950
	ニ格	攻撃 :18, 行動 :15, 息子 :15, ...
焼く(3)	ガ格	メーカー :1, ディストリビューター :1, ...
	ヲ格	データ :178, ファイル :107, コピー :9, ...
	ニ格	R:1583, CD:664, CDR:3, ...

構築した格フレームの例

用言	格	用例
泳ぐ	ガ格	イルカ :142, 生 :50, 魚 :28, ...
	ヲ格	海 :1188, 水中 :281, 海中 :101, ...
	デ格	クロール :86, 平泳ぎ :49, 泳法 :24, ...
磨く	ガ格	私 :4, 男性 :4, 人 :4, おれ :4, ...
	ヲ格	歯 :5959, 奥歯 :27, 前歯 :12
	デ格	ブラシ :38, 塩 :13, 粉 :12, ...
録画する	ガ格	旦那 :4, 妹 :2, 知人 :2, 友人 :2, ...
	ヲ格	番組 :1435, 放送 :521, 特番 :26, ...
	ニ格	ビデオ :3753, ディスク :256, ...

格フレーム検索



<http://nlp.kuee.kyoto-u.ac.jp/nl-resource/caseframe.html>

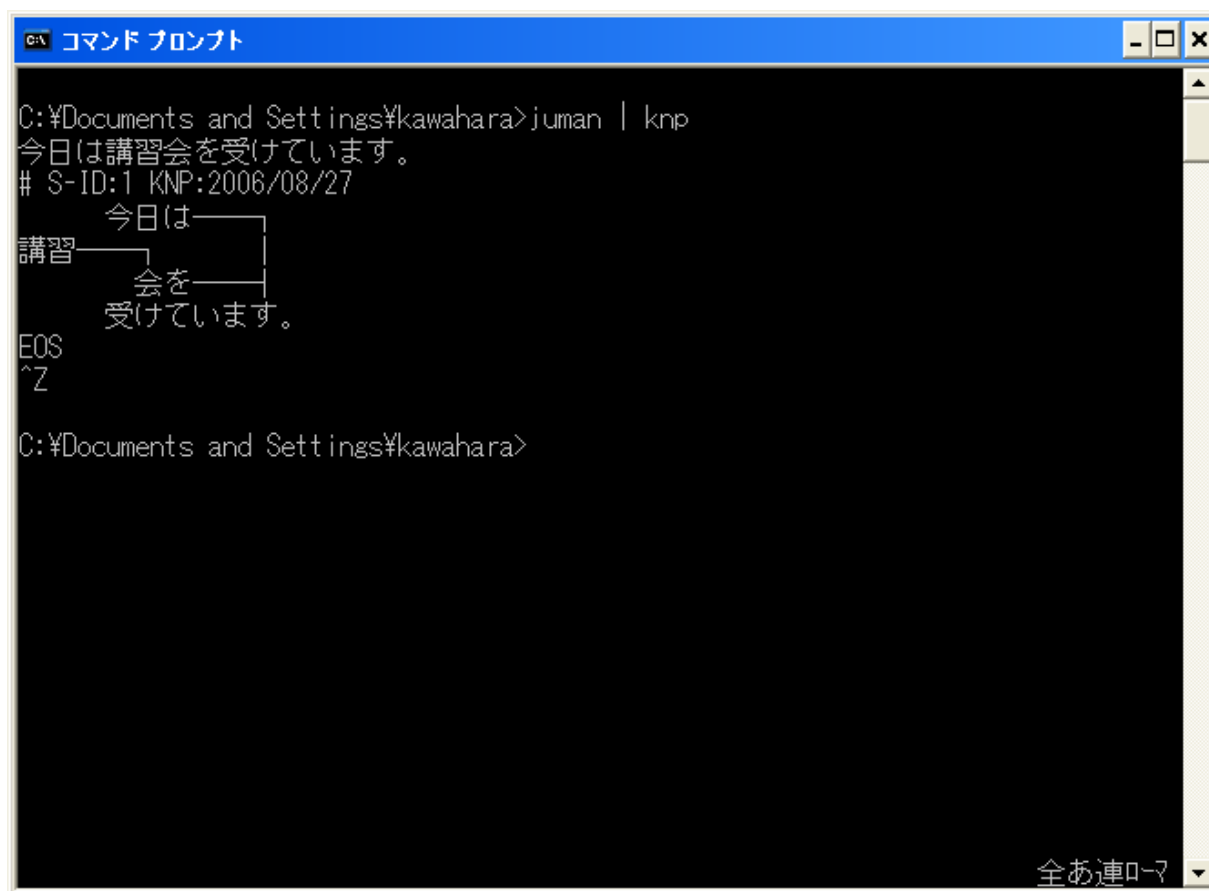
格フレームID	格	頻度
する.動1	ヲ格	13820
できる.動1	ガ格	5775
なる.動1	ニ格	1591
出来る.動1	ガ格	1264
なる.動2	ト格	1234
語る.動1	ヲ格	1216
ある.動1	ガ格	1098
行う.動1	ヲ格	1009
届け.動1	ヲ格	691
話す.動1	ヲ格	490
生かす.動1	ヲ格	443
書く.動1	ヲ格	441
実施.動32	ヲ格	429
積む.動1	ヲ格	414
楽しむ.動12	ヲ格	361

ページが表示されました

インターネット

4.2 KNP を使ってみよう

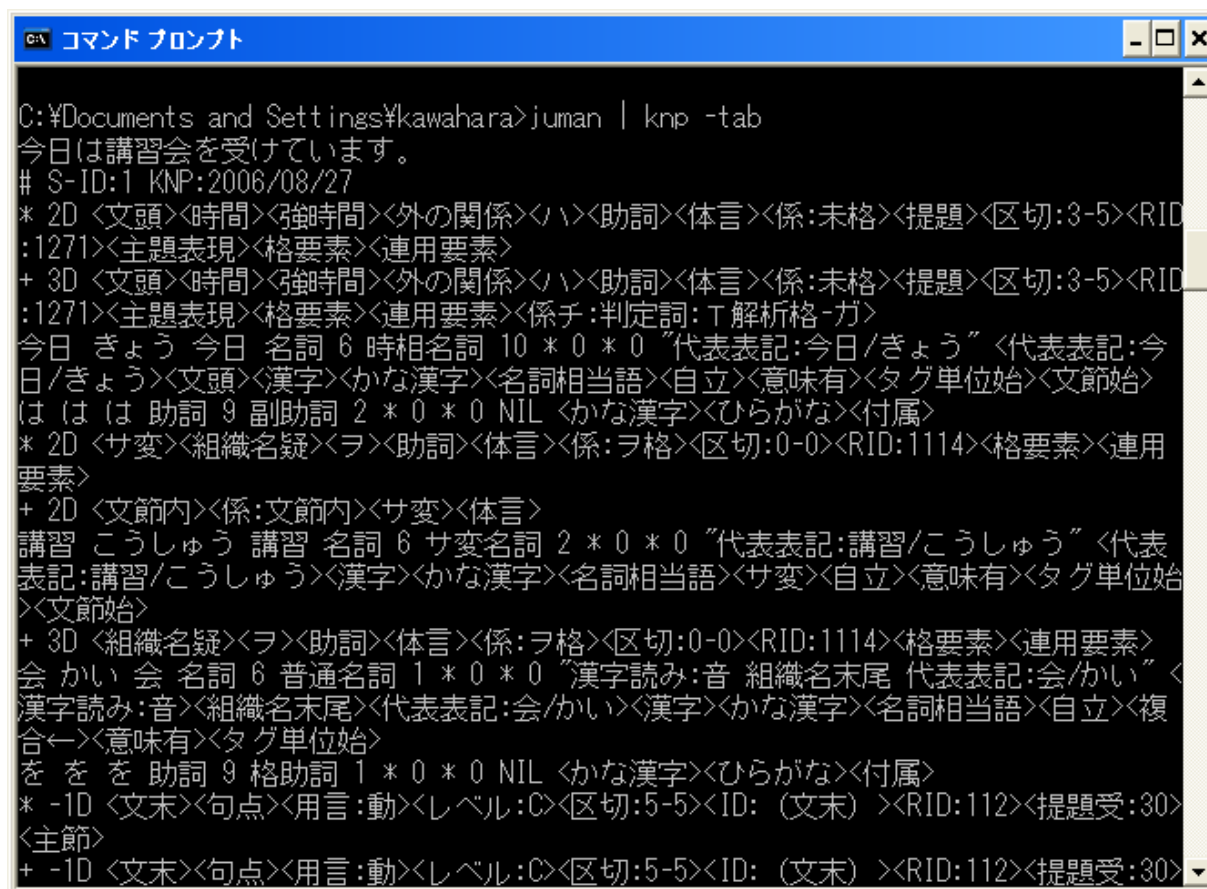
- juman | knp と打つ



```
C:\Documents and Settings\kawahara>juman | knp
今日は講習会を受けています。
# S-ID:1 KNP:2006/08/27
    今日は――
講習――|
    会を――|
    受けています。
EOS
^Z
C:\Documents and Settings\kawahara>
```

KNP を使ってみよう

- juman | knp -tab と打つ



```
C:\Documents and Settings\kawahara>juman | knp -tab
今日は講習会を受けています。
# S-ID:1 KNP:2006/08/27
* 2D <文頭><時間><強時間><外の関係><い><助詞><体言><係:未格><提題><区切:3-5><RID:1271><主題表現><格要素><連用要素>
+ 3D <文頭><時間><強時間><外の関係><い><助詞><体言><係:未格><提題><区切:3-5><RID:1271><主題表現><格要素><連用要素><係子:判定詞:T解析格-ガ>
今日 きょう 今日 名詞 6 時相名詞 10 * 0 * 0 "代表表記:今日/きょう" <代表表記:今日/きょう><文頭><漢字><かな漢字><名詞相当語><自立><意味有><タグ単位始><文節始>
は は は 助詞 9 副助詞 2 * 0 * 0 NIL <かな漢字><ひらがな><付属>
* 2D <サ変><組織名疑><ヲ><助詞><体言><係:ヲ格><区切:0-0><RID:1114><格要素><連用要素>
+ 2D <文節内><係:文節内><サ変><体言>
講習 こうしゅう 講習 名詞 6 サ変名詞 2 * 0 * 0 "代表表記:講習/こうしゅう" <代表表記:講習/こうしゅう><漢字><かな漢字><名詞相当語><サ変><自立><意味有><タグ単位始><文節始>
+ 3D <組織名疑><ヲ><助詞><体言><係:ヲ格><区切:0-0><RID:1114><格要素><連用要素>
会 かい 会 名詞 6 普通名詞 1 * 0 * 0 "漢字読み:音 組織名末尾 代表表記:会/かい" <漢字読み:音><組織名末尾><代表表記:会/かい><漢字><かな漢字><名詞相当語><自立><複合><意味有><タグ単位始>
を を を 助詞 9 格助詞 1 * 0 * 0 NIL <かな漢字><ひらがな><付属>
* -1D <文末><句点><用言:動><レベル:C><区切:5-5><ID:(文末)><RID:112><提題受:30><主節>
+ -1D <文末><句点><用言:動><レベル:C><区切:5-5><ID:(文末)><RID:112><提題受:30>
```

knp -tab フォーマット

0 番目の基本句

0 番目の文節

係り先の文節番号

S-ID:1 KNP:3.0 DATE:2009/09/30 SCORE:-20.56067

* 2D < 文頭 > < 時間 > < 強時間 > < 外の関係 > < ハ > < 助詞 > < 体言 > < 係 : 未格 > < 提題

+ 3D < 文頭 > < 時間 > < 強時間 > < 外の関係 > < ハ > < 助詞 > < 体言 > < 係 : 未格 > < 提題

今日 きょう 今日 名詞 6 時相名詞 10 * 0 * 0 " 代表表記 : 今日 / きょう " < 代表表記 : 今

は は は 助詞 9 副助詞 2 * 0 * 0 NIL < かな漢字 > < ひらがな > < 付属 >

* 2D < サ変 > < 組織名疑 > < ヲ > < 助詞 > < 体言 > < 係 : ヲ格 > < 区切 : 0-0 > < RID:1114 > <

+ 2D < 文節内 > < 係 : 文節内 > < サ変 > < 体言 >

講習 こうしゅう 講習 名詞 6 サ変名詞 2 * 0 * 0 " 代表表記 : 講習 / こうしゅう " < 代表表記 :

+ 3D < 組織名疑 > < ヲ > < 助詞 > < 体言 > < 係 : ヲ格 > < 区切 : 0-0 > < RID:1114 > < 格要素

会 かい 会 名詞 6 普通名詞 1 * 0 * 0 " 漢字読み : 音 組織名末尾 代表表記 : 会 / かい "

味有 > < タグ単位始 >

を を を 助詞 9 格助詞 1 * 0 * 0 NIL < かな漢字 > < ひらがな > < 付属 >

1 番目の文節

1 番目の基本

句

2 番目の基本

句

ファイルから入力する場合

- `cd c:\juman-knp-20090930\text`
- `juman < cook_small.txt > cook_small.jmn`
- `knp -tab < cook_small.jmn > cook_small.knp`

Tips

ファイル・ディレクトリ名は
Tab で補完できます

目次

1. インストール確認
2. 環境設定
3. JUMAN の仕組み、使い方
4. KNP の仕組み、使い方
5. Perl 超入門
6. JUMAN/KNP と Perl を用いたいろいろな頻度統計の取り方
7. 自動構築した大規模格フレームとそれに基づく格解析

5. Perl 超入門

- Perl は言語処理に適したプログラミング言語

スチームコンベクションオーブン100℃で4分間加熱する。
イラストやツクール素材など、自作ゲームも制作しています。
海洋生物資源の有効利用
含芒硝石膏泉:無色透明、源泉約50℃
オレンジのビタミンCが風邪を予防してくれます。
活用の部屋戻る
くっきもさん、お粗末様でございましたこの場合はおかしいですね
:

テキスト



633 、
607 の
578 に
:
43 料理
43 塩
42 かける
41 や
38 など
36 梅
:

頻度つき単語リスト

Perl 超入門

- まずはじめに以下のようなファイルをエディタ（メモ帳など）で作成

```
use encoding "shiftjis";
```

おまじない

```
print "こんにちは。\\n";
```

文字列を表示する関数

改行を表す記号

- C:\juman-knp-20090930\src\test.pl に保存
- cd C:\juman-knp-20090930\src
- コマンド プロンプトで、perl test.pl を実行

Perl 超入門

- パターンにマッチする行を表示するプログラム
(src\grep.pl)

```
use encoding 'shiftjis';  
$ARGV[0] = Encode::decode('shiftjis', $ARGV[0]);  
  
while (<STDIN>) {  
    print if (/$ARGV[0]/);  
}
```

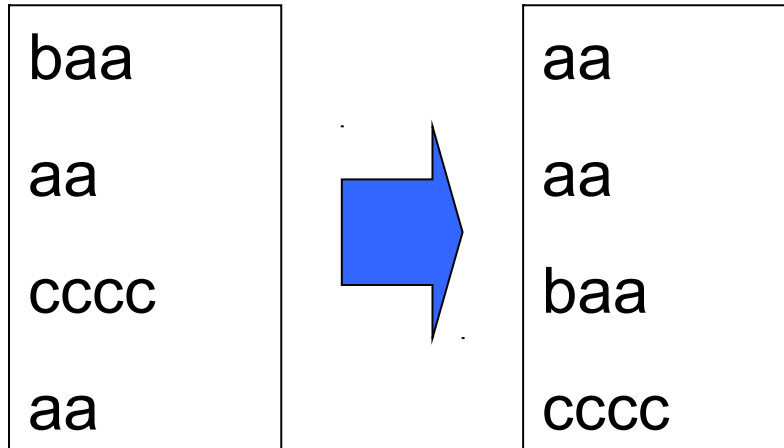
条件

1 つ目の引数

「 /.../ 」はパターンマッチを行う

Perl 超入門

- ソートするプログラム (src\sort.pl)



Perl 超入門

- ソートするプログラム (src\sort.pl)

...

```
while (<STDIN>) {  
    push(@buffer, $_);  
}
```

@... は配列

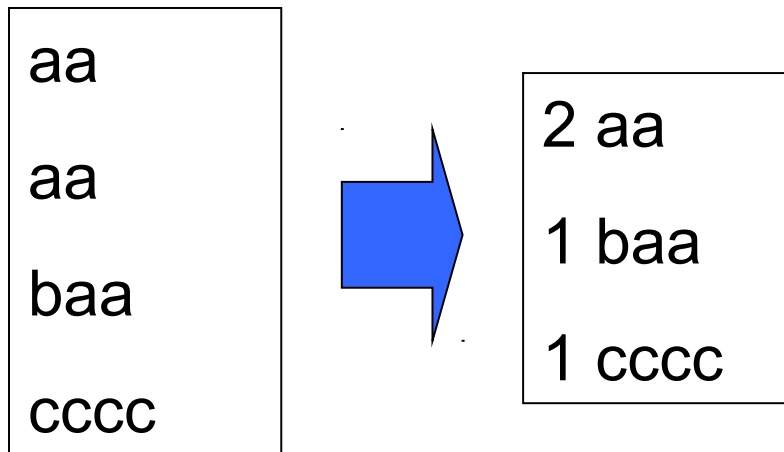
```
if ($rflag) {  
    print reverse sort @buffer;  
} else {  
    print sort @buffer;  
}
```

配列を逆順にする関数

ソートする関数

Perl 超入門

- ソートされたファイルから重複行の行数を数えるプログラム (src\uniq.pl)



Perl 超入門

- ソートされたファイルから重複行の行数を数えるプログラム (src\uniq.pl)

```
use encoding 'shiftjis';

$pre = <STDIN>;
$count = 1;

while (<STDIN>) {
    if ($pre eq $_) {
        $count++;
    } else {
        printf "%6d $pre", $count;
        $pre = $_;
        $count = 1;
    }
}
printf "%6d $pre", $count;
```

Perl 超入門

- 各行から指定したカラムを表示するプログラム
(src\cut.pl)

```
今日 きょう 今日 名詞 6 時相名詞 10 * 0 * 0 " 代表表記 : 今日 / きょう "  
は は は 助詞 9 副助詞 2 * 0 * 0 NIL  
講習 こうしゅう 講習 名詞 6 サ変名詞 2 * 0 * 0 " 代表表記 : 講習 / こうしゅう "  
会 かい 会 名詞 6 普通名詞 1 * 0 * 0 " 漢字読み : 音 組織名末尾 代表表記 : 会 / か  
い "  
を を を 助詞 9 格助詞 1 * 0 * 0 NIL  
受けて うけて 受ける 動詞 2 * 0 母音動詞 1 タ系連用テ形 14 " 代表表記 : 受ける / う  
ける "  
いい いる 接尾辞 14 動詞性接尾辞 7 母音動詞 1 基本連用形 8 " 代表表記 : いる /  
いる "  
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 ...  
。。。特殊 1 句点 1 * 0 * 0 NIL  
EOS
```

Perl 超入門

- 各行から指定したカラムを表示するプログラム
(src\cut.pl -3)

今日 きょう 今日 名詞 6 時相名詞 10 * 0 * 0 " 代表表記 : 今日 / きょう "
は は は 助詞 9 副助詞 2 * 0 * 0 NIL
講習 こうしゅう 講習 名詞 6 サ変名詞 2 * 0 * 0 " 代表表記 : 講習 / こうしゅう "
会 かい 会 名詞 6 普通名詞 1 * 0 * 0 " 漢字読み : 音 組織名末尾 代表表記 : 会 / か
い "
を を を 助詞 9 格助詞 1 * 0 * 0 NIL
受けて うけて 受ける 動詞 2 * 0 母音動詞 1 タ系連用テ形 14 " 代表表記 : 受ける / う
ける "
いい いる 接尾辞 14 動詞性接尾辞 7 母音動詞 1 基本連用形 8 " 代表表記 : いる /
いる "
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 ...
。。。特殊 1 句点 1 * 0 * 0 NIL
EOS

目次

1. インストール確認
2. 環境設定
3. JUMAN の仕組み、使い方
4. KNP の仕組み、使い方
5. Perl 超入門
6. JUMAN/KNP と Perl を用いたいろいろな頻度統計の取り方
7. 自動構築した大規模格フレームとそれに基づく格解析

6.1 単語の頻度を数える

- `cd c:\juman-knp-20090930\text`
- テキストの例として `cook_small.txt` を用いる

スチームコンベクションオーブン100℃で4分間加熱する。
イラストやツクール素材など、自作ゲームも制作しています。

海洋生物資源の有効利用

含芒硝石膏泉：無色透明、源泉約50℃

オレンジのビタミンCが風邪を予防してくれます。

活用の部屋戻る

くっきもさん、お粗末様でございましたこの場合はおかしいですね、調理されたのはあなたですものね

甘さ辛さを味わいつくし、苦さも知った人生、その人の舌が、それを知るはずだ。

:

単語の頻度を数える

- 形態素解析

juman < cook_small.txt > cook_small.jmn

- 単語の原形を抽出する

perl ..\src\cut.pl -3 < cook_small.jmn | more

3 目 = 原形を抽出

スチームコンベクションオーブン
100
℃
で
4
分間
加熱
する
。
:

単語の頻度を数える

- 抽出した単語の頻度を数える

```
perl ..\src\cut.pl -3 < cook_small.jmn | perl ..\src\sort.pl |  
perl ..\src\uniq.pl | perl ..\src\sort.pl -r | more
```

:
ある
ある
ある
ある
ある
ある
ある
ある
ある
:

:
3 ありがとう
52 ある
1 あわせた
1 あわせて
2 あわせる
1 あわび
6 あん
1 あんしん
1 あんず
:

633 、
607 の
578 に
:
43 料理
43 塩
42 かける
41 や
38 など
36 梅
:

src\phrase.pl

- 文節や係り受けを抽出するプログラム
 - 文節の抽出: `phrase.pl -1`
 - 係り受けの抽出: `phrase.pl -2`
- `cook_large.knp` でやってみる
 - `perl ../src/phrase.pl -1 knp/cook_large.knp > cook_large.dat1`
 - `perl ../src/phrase.pl -2 knp/cook_large.knp > cook_large.dat2`

src\phrase.pl

- `#!/usr/bin/env perl`

`# 文節または係り受けを抽出するスクリプト`

`# UNIX 系 OS の環境で、標準入力からテキストを読みながら解析する場合は以下のようにする`

`# use KNP;`

`# $KNP = new KNP;`

`# while (<STDIN>) {`

`# $result = $KNP->parse($_);`

`# for my $bnst ($result->bnst) {`

`# ...`

`use KNP::File;`

`use encoding 'shiftjis';`

`if ($ARGV[0] =~ /\-(1|2)$/ && -f $ARGV[1]) {`

`$type = $1;`

`# 解析済みファイルを読み込む`

`$KNP = new KNP::File($ARGV[1]) || die;`

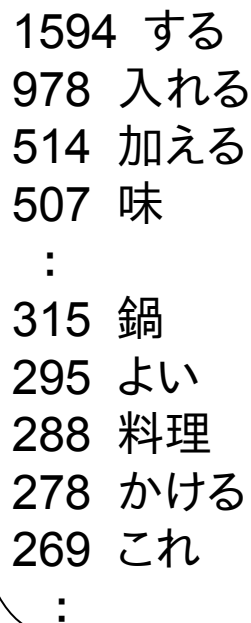
`} else {`

`;`

6.2 文節の頻度を数える

- 抽出した文節の頻度を数える

```
perl ..\src\sort.pl < cook_large.dat1 | perl ..\src\uniq.pl |  
perl ..\src\sort.pl -r | more
```

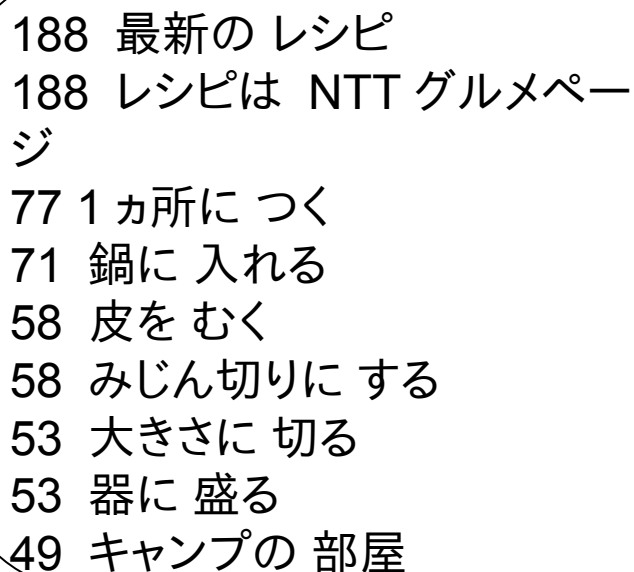


1594 する
978 入れる
514 加える
507 味
:
315 鍋
295 よい
288 料理
278 かける
269 これ
:

6.3 係り受けの頻度を数える

- 抽出した係り受けの頻度を数える

```
perl ../src\sort.pl < cook_large.dat2 | perl ../src\uniq.pl |  
perl ../src\sort.pl -r | more
```



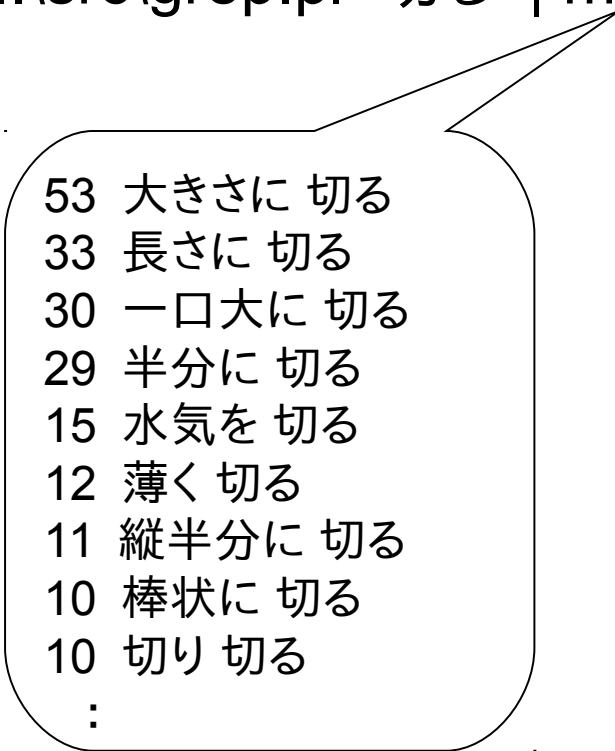
188 最新のレシピ
188 レシピは NTT グルメペー
ジ
77 1カ所につく
71 鍋に入れる
58 皮をむく
58 みじん切りにする
53 大きさに切る
53 器に盛る
49 キャンプの部屋

:

表現の検索

- 「切る」を検索

```
perl ..\src\sort.pl < cook_large.dat2 | perl ..\src\uniq.pl |  
perl ..\src\sort.pl -r | perl ..\src\grep.pl 切る | more
```



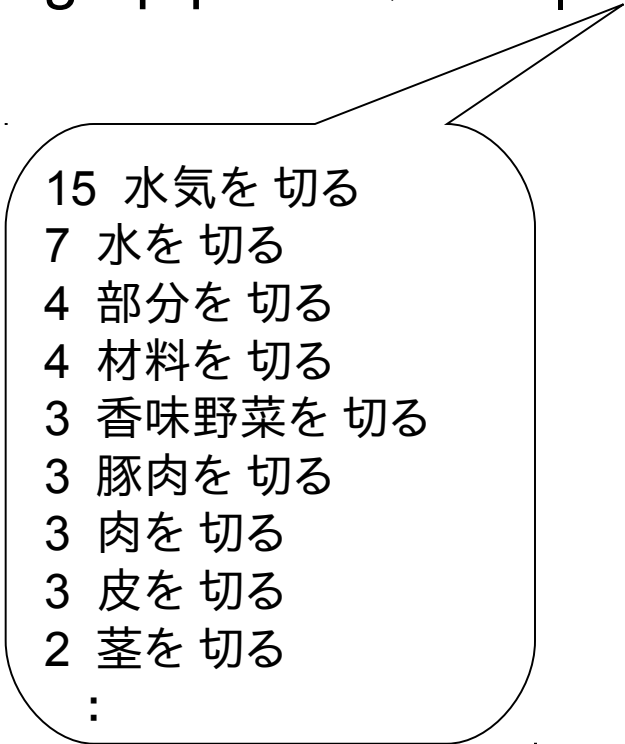
53 大きさに 切る
33 長さに 切る
30 一口大に 切る
29 半分に 切る
15 水気を 切る
12 薄く 切る
11 縦半分に 切る
10 棒状に 切る
10 切り 切る

:

表現の検索

- 「を 切る」を検索

```
perl ..\src\sort.pl < cook_large.dat2 | perl ..\src\uniq.pl |  
perl ..\src\sort.pl -r | perl ..\src\grep.pl "を 切る" | more
```



15 水気を切る
7 水を切る
4 部分を切る
4 材料を切る
3 香味野菜を切る
3 豚肉を切る
3 肉を切る
3 皮を切る
2 茎を切る

:

試してみよう

- 別のテキストでやってみる
- 自分の知りたい表現を検索してみる

目次

1. インストール確認
2. 環境設定
3. JUMAN の仕組み、使い方
4. KNP の仕組み、使い方
5. Perl 超入門
6. JUMAN/KNP と Perl を用いたいろいろな頻度統計の取り方
7. 自動構築した大規模格フレームとそれに基づく格解析

格解析

ドイツ語も話す先生

ガ格

ヲ格

話す (1)

ガ

人, 私, ...

ヲ

英語, 言語, ...

話す (2)

ガ

患者, ...

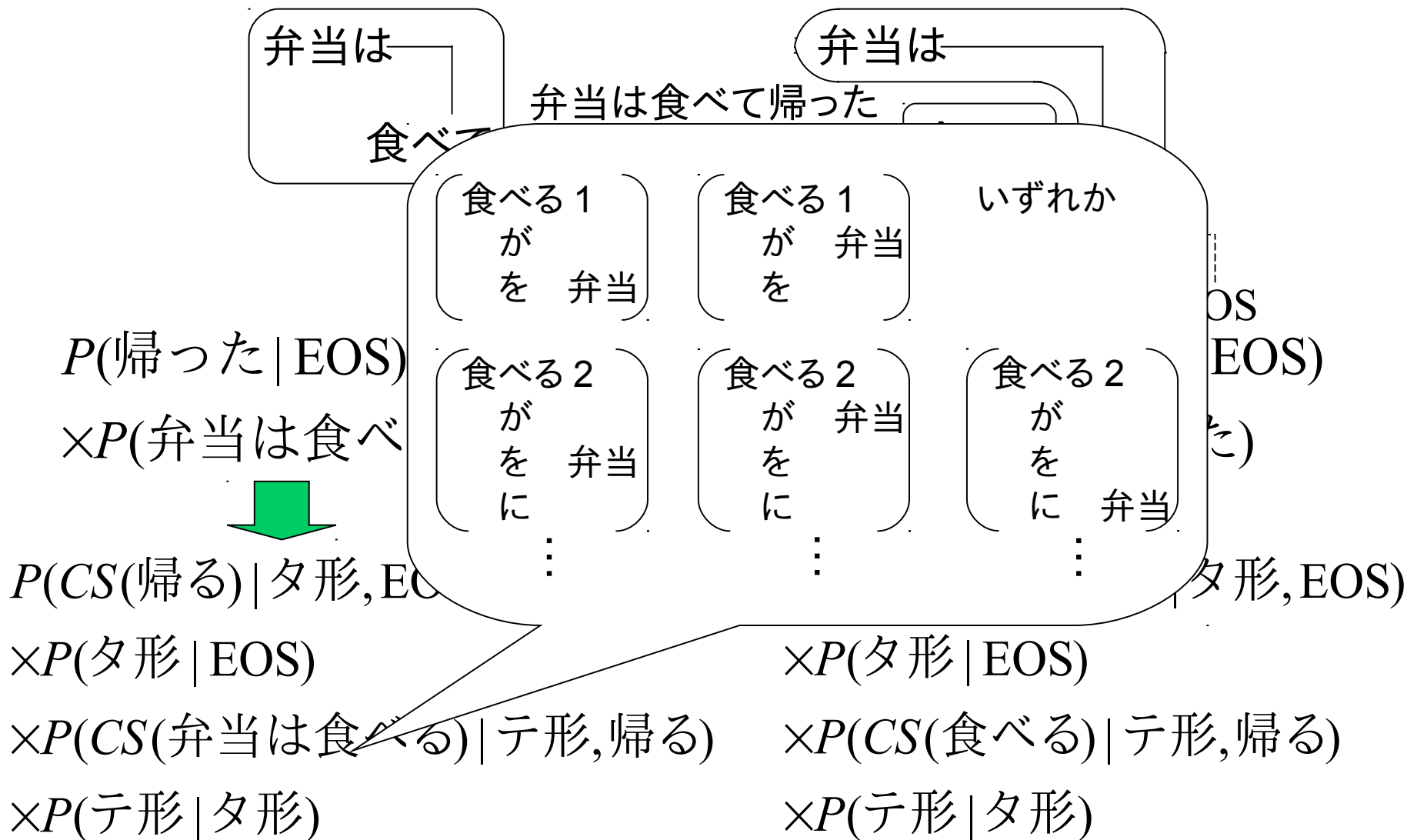
ヲ

症状, 状態, ...

二

医師, 医者, ...

構文・格解析の統合的確率モデル



KNP における格解析の使い方

- 格解析の実行
 - KNP3.0 ではデフォルトでオンになっている
 - tab 出力 : `knp -tab`
 - 詳細出力 : `knp -detail`
- 格解析結果(述語項構造)の表示

格解析結果の表示

- 述語項構造を木構造とともに表示

```
perl ../src/print_pa.pl -m knp/cook_small.knp |  
more
```

S-ID:1 KNP:3.0 DATE:2009/09/30

鱈 n は p ——— 時期 : ガ
 寒い j ———
 時期 n に p ———
 旬 n を p ———
 迎える v ——— 魚 : ガ 旬 : ヲ 時期 : ニ
魚 n です c . * 鱈 : ガ

参考情報

ツールの公開:

<http://nlp.kuee.kyoto-u.ac.jp/nl-resource/>

参考文献

黒橋禎夫, 長尾 眞: 長い日本語文における並列構造の推定, 情報処理学会論文誌, Vol.33, No.8, pp.1022-1031 (1992.8).

黒橋禎夫, 長尾 眞: 並列構造の検出に基づく長い日本語文の構文解析, 自然言語処理, Vol.1, No.1, pp.35-57 (1994.10).

黒橋禎夫, 長尾 眞: 京都大学テキストコーパス・プロジェクト, 言語処理学会 第3回年次大会, pp.115-118 (1997.3).

黒橋禎夫: 開発されるべきシステムとしての言語, 月刊「言語」, Vol.27, No.6, pp.66-73 (1998.6).

黒橋禎夫: コーパスが先か, パーサーが先か, 情報処理 Vol.41, No.7, pp.769-773 (2000.7).

黒橋禎夫: 結構やるな, KNP, 情報処理 Vol.41, No.11, pp.1215-1220 (2000.11).

河原大輔, 黒橋禎夫: 用言と直前の格要素の組を単位とする格フレームの自動構築, 自然言語処理, Vol.9, No.1, pp.3-19 (2002.1).

黒橋禎夫: 自然言語処理を支える文法, 月刊「言語」, Vol.31, No.4, pp.52-57 (2002.4).

河原大輔, 黒橋禎夫: 格フレーム辞書の漸次的自動構築, 自然言語処理, Vol.12, No.2, pp.109-131 (2005.3).

黒橋禎夫: 言語のセマンティックス, 人工知能学会誌, Vol.21, No.6, pp.718-723 (2006.11).

河原大輔, 黒橋禎夫: 自動構築した大規模格フレームに基づく構文・格解析の統合的確率モデル, 自然言語処理, Vol.14, No.4, pp.67-81, (2007.7).

黒橋禎夫: 言語コンピューティング, 人工知能学会誌, Vol.22, No.5 (2007.9).