

テキストデータの統計科学入門

3章 形態素解析と構文解析

12NM712L 小幡智裕

日本語テキストの計量分析

- 日本語のテキストは英語のように単語や句に分かれていない
 - 語、文節などを計量するためにテキストを語や文節などを単位として分割し、関連情報を付与しておくことが必要

形態素解析

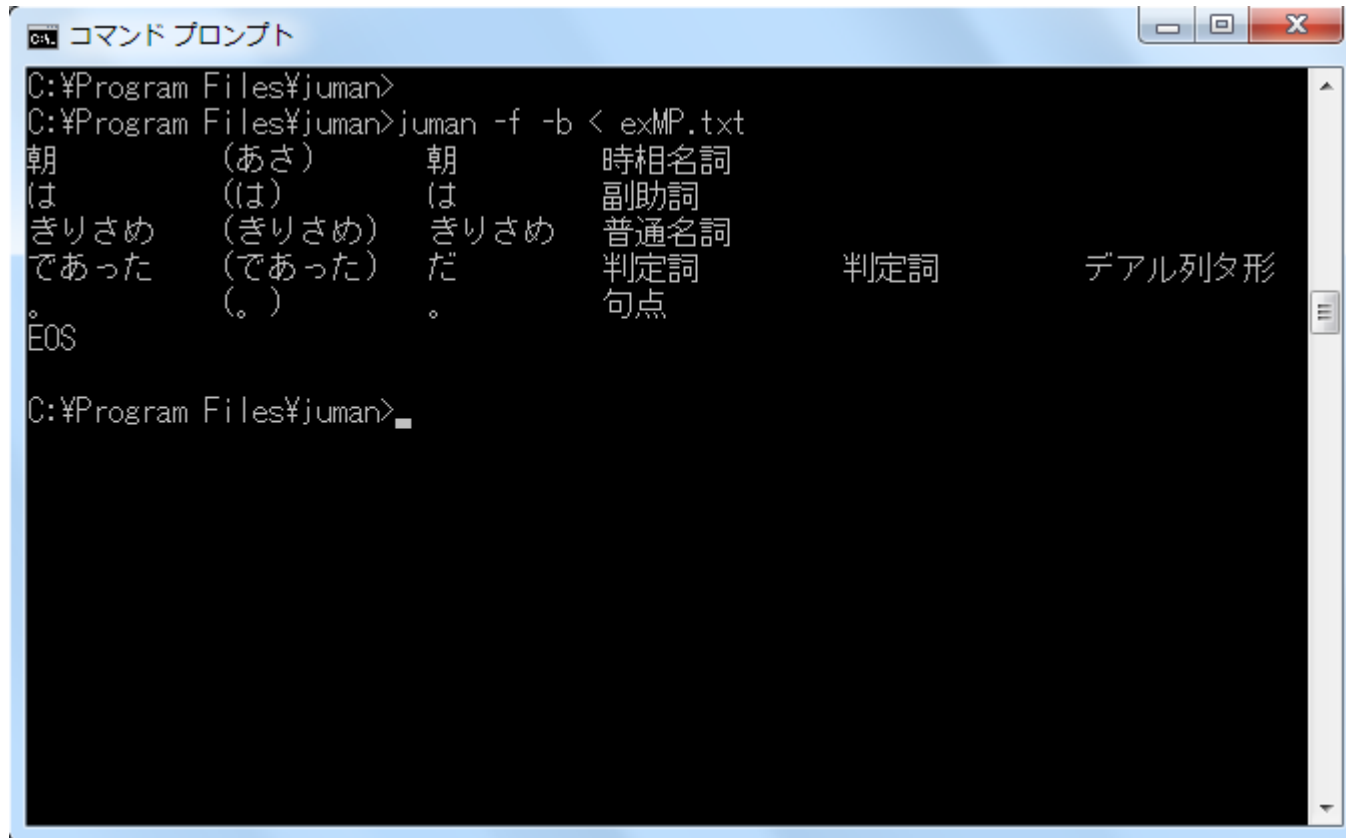
- 形態素
 - 意味を持つ最小の文字列の単位
- 形態素解析
 - 文を単語ごとに分割し、品詞情報などを付け加える作業

例：花が咲く

「花」→名詞、「が」→助詞、「咲く」→動詞

JUMAN

- 長尾真氏の研究室を中心に開発され、1992年に公開された



```
コマンドプロンプト
C:\Program Files\juman>
C:\Program Files\juman>juman -f -b < exMP.txt
朝          (あさ)      朝          時相名詞
は          (は)        は          副助詞
きりさめ   (きりさめ)  きりさめ   普通名詞
であった   (であった)  だ         判定詞      判定詞      デアル列々形
。          (。)        。         句点
EOS
C:\Program Files\juman>.
```

茶筌 (ChaSen)

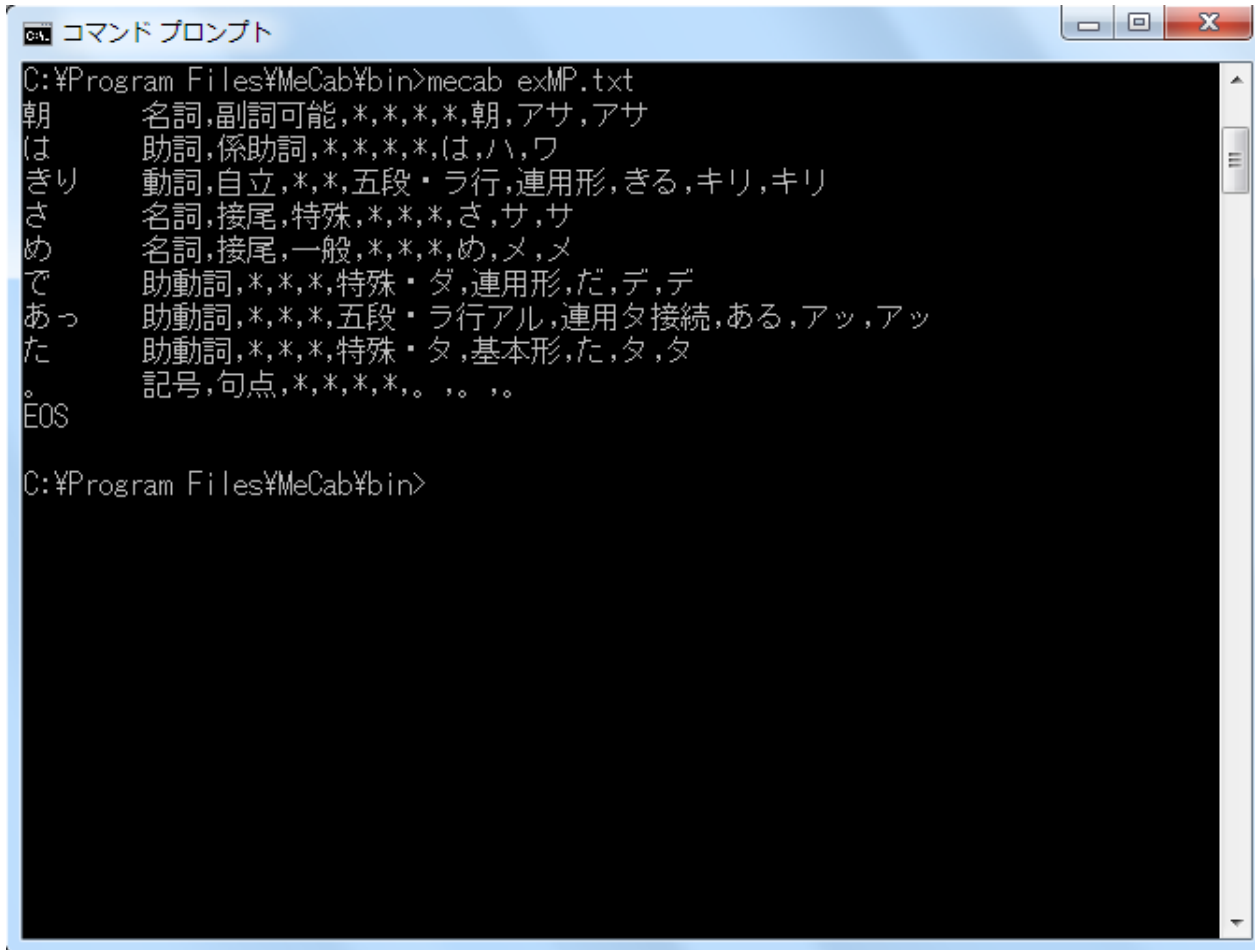
- 奈良先端科学技術大学院大学の松本裕治氏の研究室で開発され、1997年に公開

```
コマンドプロンプト
C:\Program Files\chasen20>chasen exMP.txt
朝      アサ   朝      名詞-副詞可能
は      ハ     は      助詞-係助詞
きり    キリ   きる    動詞-自立      五段・ラ行      連用形
さめ    サメ   さめる  動詞-自立      一段          連用形
で      デ     だ      助動詞 特殊・ダ 連用形
あっ    アッ   ある    助動詞 五段・ラ行アル 連用形接続
た      タ     た      助動詞 特殊・タ 基本形
。      。     。      記号-句点
EOS

C:\Program Files\chasen20>
```

MeCab

- 工藤拓氏が開発し、2002年に公開



```
コマンド プロンプト
C:\Program Files\MeCab\bin>mecab exMP.txt
朝      名詞,副詞可能,*,*,*,*,朝,アサ,アサ
は      助詞,係助詞,*,*,*,*,は,ハ,ワ
ぎり    動詞,自立,*,*,五段・ラ行,連用形,きる,キリ,キリ
さ      名詞,接尾,特殊,*,*,*,さ,サ,サ
め      名詞,接尾,一般,*,*,*,め,メ,メ
で      助動詞,*,*,*,特殊・ダ,連用形,だ,デ,デ
あっ    助動詞,*,*,*,五段・ラ行アル,連用タ接続,ある,アッ,アッ
た      助動詞,*,*,*,特殊・タ,基本形,た,タ,タ
。      記号,句点,*,*,*,*,。,,。,,。
EOS

C:\Program Files\MeCab\bin>
```

構文解析

- 構文解析

- 文法規則にもとづいて、文の構造を句・文節を単位として解析すること
- 日本語では、文節を単位に係り受け関係を用いて構文を解析するのが一般的であり、文における1つの文節はその文節の後の少なくとも1つの文節と係り受け関係を持つ

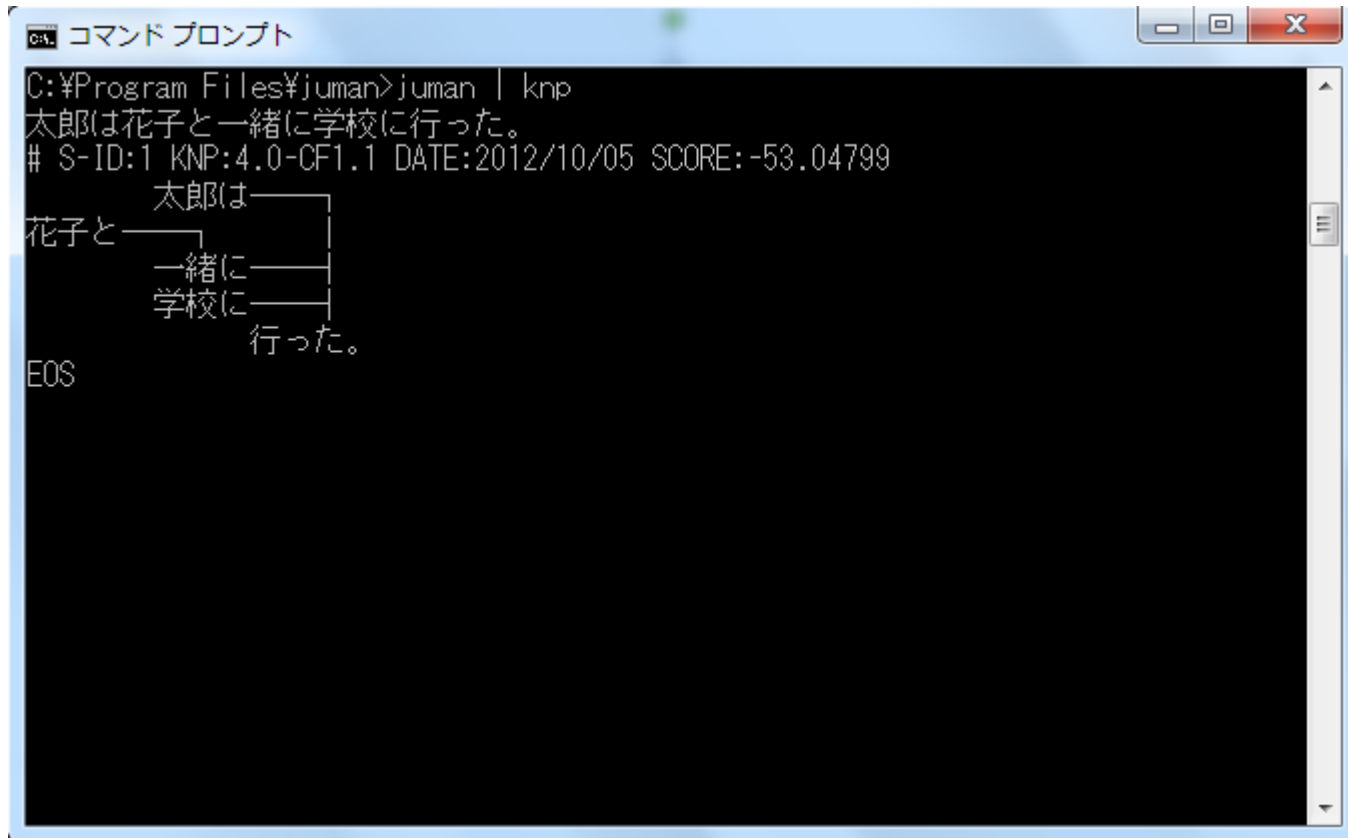
例：太郎は花子と一緒に学校に行った。



太郎は/ 花子と/ 一緒に/ 学校に/ 行った。

KNP

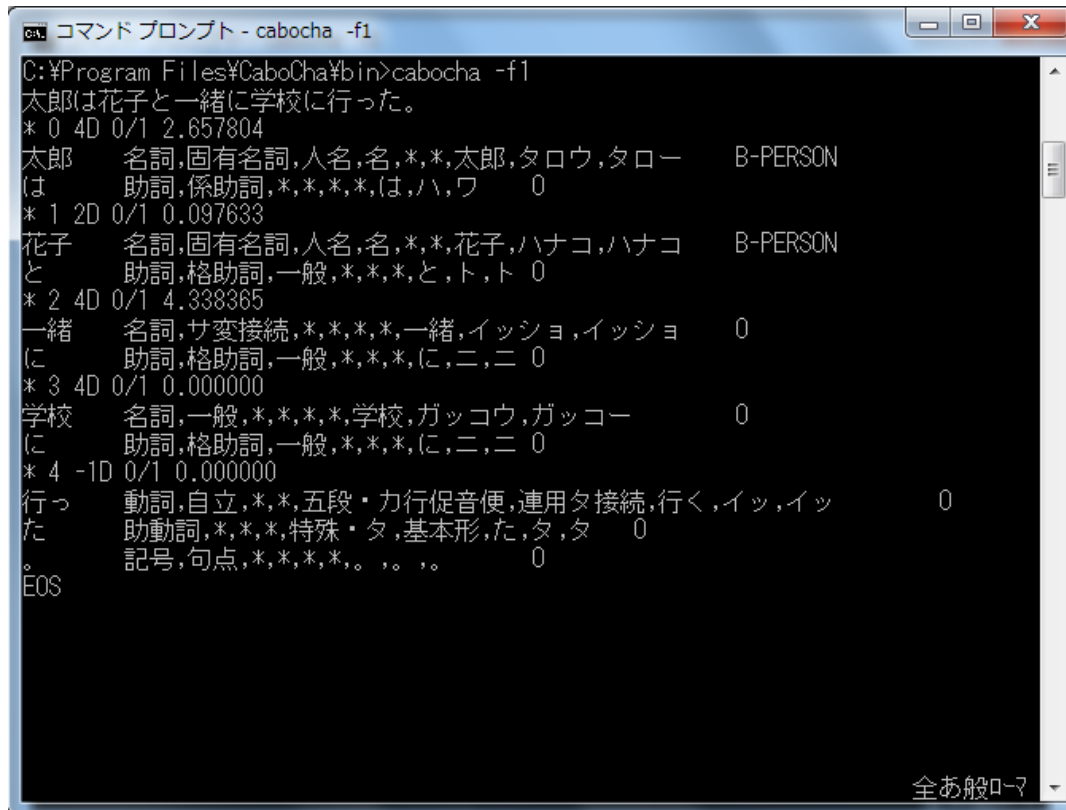
- JUMANをベースとした係り受け関係による構文解析器。1993年に公開



```
コマンドプロンプト
C:\Program Files\juman>juman | knp
太郎は花子と一緒に学校に行った。
# S-ID:1 KNP:4.0-CF1.1 DATE:2012/10/05 SCORE:-53.04799
      太郎は———|
花子と———|
      一緒に———|
      学校に———|
                行った。
EOS
```


CaboCha

- 工藤拓氏と松本裕治氏により開発されたSVMにもとづいた係り受け解析器



```
コマンド プロンプト - cabochoa -f1
C:\Program Files\CaboCha\bin>cabochoa -f1
太郎は花子と一緒に学校に行った。
* 0 4D 0/1 2.657804
太郎 名詞,固有名詞,人名,名,*,*,太郎,タロウ,タロー B-PERSON
は 助詞,係助詞,*,*,*,は,ハ,ワ 0
* 1 2D 0/1 0.097633
花子 名詞,固有名詞,人名,名,*,*,花子,ハナコ,ハナコ B-PERSON
と 助詞,格助詞,一般,*,*,*,と,ト,ト 0
* 2 4D 0/1 4.338365
一緒 名詞,サ変接続,*,*,*,一緒,イッショ,イッショ 0
に 助詞,格助詞,一般,*,*,*,に,ニ,ニ 0
* 3 4D 0/1 0.000000
学校 名詞,一般,*,*,*,学校,ガッコウ,ガッコウ 0
に 助詞,格助詞,一般,*,*,*,に,ニ,ニ 0
* 4 -1D 0/1 0.000000
行っ 動詞,自立,*,*,五段・力行促音便,連用タ接続,行く,イッ,イッ 0
た 助動詞,*,*,*,特殊・タ,基本形,た,タ,タ 0
。 記号,句点,*,*,*,.,.,. 0
EOS
```